frontiers | Research Topics

AI innovations in neuroimaging: transforming brain analysis

Edited by
S. B. Goyal, Deepti Deshwal and Pardeep Sangwan

Published in
Frontiers in Medicine
Frontiers in Neuroscience

FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-7410-2
DOI 10.3389/978-2-8325-7410-2

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

AI innovations in neuroimaging: transforming brain analysis

Topic editors

S. B. Goyal — City University, Malaysia

Deepti Deshwal — Maharaja Surajmal Institute of Technology, India

Pardeep Sangwan — Maharaja Surajmal Institute of Technology, India

Citation

Goyal, S. B., Deshwal, D., Sangwan, P., eds. (2026). *AI innovations in neuroimaging: transforming brain analysis*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-7410-2

Table of contents

frontiers | Frontiers in Medicine

TYPE Editorial
PUBLISHED 13 January 2026
DOI 10.3389/fmed.2025.1755373

Check for updates

OPEN ACCESS

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

*CORRESPONDENCE
Deepti Deshwal
✉ deshwaldeepti@msit.in

RECEIVED 27 November 2025
ACCEPTED 29 December 2025
PUBLISHED 13 January 2026

CITATION
Goyal SB, Deshwal D and Sangwan P (2026) Editorial: AI innovations in neuroimaging: transforming brain analysis.
Front. Med. 12:1755373.
doi: 10.3389/fmed.2025.1755373

COPYRIGHT
© 2026 Goyal, Deshwal and Sangwan. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Editorial: AI innovations in neuroimaging: transforming brain analysis

Shyam Bihari Goyal[1], Deepti Deshwal[2]* and Pardeep Sangwan[2]

[1]School of Computer Science and Engineering, Chitkara University, Rajpura, India, [2]Department of Electronics and Communication Engineering (ECE), Maharaja Surajmal Institute of Technology, New Delhi, India

KEYWORDS

brain-inspired computing, EEG signal classification, fMRI pattern analysis, neuroscience, recurrent neural networks (memory + sequences)

Editorial on the Research Topic
AI innovations in neuroimaging: transforming brain analysis

Over the last decade, artificial intelligence (AI) has transformed nearly every branch of medical imaging, but its impact on neuroimaging has been particularly revolutionary (1–5). From automated segmentation of magnetic resonance imaging (MRI) data to deep learning–assisted disease prediction, AI techniques—especially machine learning (ML), deep learning (DL), and emerging quantum computing paradigms are reshaping how clinicians interpret the human brain. These computational advances are accelerating the diagnosis of neurological disorders, optimizing patient management, and opening new frontiers in personalized medicine (6–10).

The Research Topic "*AI Innovations in Neuroimaging: Transforming Brain Analysis*" brings together a diverse collection of studies that harness advanced algorithms and hybrid models to address key clinical challenges in brain analysis, ranging from tumor classification and stroke detection to autism spectrum disorder (ASD) assessment and schizophrenia identification. Each contribution underscores how AI, when aligned with clinical neuroimaging, can enable faster, non-invasive, and highly interpretable diagnostics.

This Research Topic presents 11 articles that collectively highlight the breadth of AI-driven neuroimaging research. The contributions span a wide range of applications from brain tumor detection and stroke prediction to epilepsy monitoring and autism diagnosis demonstrating how interdisciplinary advances are transforming precision medicine and neuroscience.

Among the notable contributions, Priyadharshini et al. introduce *QBrainNet*, a hybrid quantum-classical neural network that leverages quantum superposition and entanglement to improve stroke prediction accuracy to 96%, outperforming traditional CNN-based approaches. By combining quantum feature extraction with variational quantum circuits, this model demonstrates the transformative role of quantum-assisted intelligence in medical imaging. In another important development, Cüce et al. propose a hybrid deep learning radiomics framework that analyzes cerebrospinal fluid (CSF) signals in central nervous system infections (CNSIs). Their approach accurately identifies infection-related CSF alterations on MRI scans, offering a promising non-invasive alternative to lumbar puncture, traditionally the gold standard in CNS infection diagnosis.

Broadening the perspective beyond imaging, Farhah et al. present a Double Deep Q-Network (DDQN) model to identify ASD traits from social media text, demonstrating

how digital footprint analysis can complement neuroimaging by capturing behavioral and emotional cues indicative of neurodevelopmental disorders. Similarly, Yuan et al. apply a robust multi-task feature selection strategy with counterfactual explanations to identify schizophrenia-related functional brain networks from resting-state fMRI data, enhancing both classification accuracy and clinical interpretability. These studies illustrate how AI-driven behavioral and cognitive analysis extends neuroimaging beyond the scanner to the digital and functional realms of brain health.

Advancing the field of brain tumor detection, Han et al. modify the YOLOv11 architecture by integrating novel attention mechanisms and a hybrid loss function (HKCIoU), achieving improved accuracy and reduced computational cost—an essential step toward real-time tumor detection in clinical environments. Naeem et al. complement this effort with a lightweight CNN tailored for small MRI datasets, achieving 99% accuracy and proving that data-efficient deep learning can yield high reliability even with limited samples. Alsubai et al. further expand diagnostic scope by combining transfer learning and explainable AI (XAI) for multi-disease MRI classification, accurately identifying both brain tumors and Alzheimer's disease across datasets. The integration of SHapley Additive exPlanations (SHAP) ensures transparency, allowing clinicians to visualize model reasoning. Meanwhile, Chen et al. introduce a Mixed Local and Global (MLG) model that fuses CNN and Transformer architectures through a gated attention mechanism. By integrating fine-grained and contextual features, their model achieves near-perfect accuracies (99.02% and 97.24%) and sets a new benchmark for hybrid architectures in neuroimaging.

Moving from structural MRI to electrophysiological data, Al-Adhaileh et al. employ EEG-based ML and DL frameworks for epileptic seizure detection, achieving an exceptional 99.9% accuracy using Random Forests. This demonstrates the capability of non-invasive EEG-based AI systems for reliable real-time seizure monitoring. Complementarily, Yuan et al. enhance feature interpretability in schizophrenia detection by applying counterfactual modeling to identify functional connectivity abnormalities, providing a neurobiological rationale behind model predictions.

In the domain of multimodal neuroimaging, Chandrasekaran et al. propose a powerful ensemble model combining VGG19 and Bidirectional LSTM with LightGBM for MRI-based brain simulations, achieving 97% accuracy and an AUC of 0.997. This hybrid design demonstrates how spatial and temporal feature fusion can improve diagnostic performance while supporting sustainable healthcare AI, a crucial step toward scalable clinical deployment. Collectively, these contributions highlight the evolution of AI in neuroimaging from task-specific models toward integrated, interpretable, and efficient systems capable of supporting real-world clinical decision-making. Ciftci et al. present a dual-model AI framework that synergistically combines clinical analytics and neuroimaging to improve Alzheimer's disease diagnosis. An Artificial Neural Network (ANN) trained on demographic and behavioral data from 1,200 patients provides risk prediction with 87.08% accuracy, while a Convolutional Neural Network (CNN) analyzes 4,876 MRI scans to stage disease progression with 97% accuracy using explainable Grad-CAM visualizations. By integrating structured clinical features with imaging-based assessment, the hybrid system enhances both diagnostic precision and clinical interpretability, aligning with the growing trend toward multimodal, scalable, and AI-assisted neuroimaging solutions for neurodegenerative disorders.

Emerging themes across the Research Topic

Across the 11 studies in this Research Topic, several unifying themes emerge. First, hybrid intelligence, the integration of quantum computing, CNNs, Transformers, and ensemble learning, is redefining neuroimaging accuracy and adaptability. Second, explainability has become a cornerstone of modern neuro-AI research. Through SHAP, counterfactual reasoning, and attention visualization, the models presented here strive not only for accuracy but also for interpretability, fostering clinical trust in AI-driven diagnostics. Third, the move toward data-efficient models such as lightweight CNNs and transfer learning underscores a shift toward accessibility, enabling AI adoption even in data-constrained healthcare systems.

Additionally, multimodal integration combining MRI, fMRI, EEG, and behavioral data reflects a growing recognition that brain disorders are inherently multifactorial and cannot be captured through a single data source. These multimodal approaches bridge the gap between structure and function, allowing for more holistic assessments of neurological conditions. Finally, the emphasis on sustainability and scalability ensures that emerging AI technologies can transition from research prototypes to clinical practice, empowering healthcare systems globally.

Author contributions

DD: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. SG: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. PS: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Acknowledgments

The Guest Editors gratefully acknowledge all contributing authors and reviewers whose rigorous efforts made this Research Topic possible. Their work exemplifies global collaboration at the intersection of AI and neuroimaging.

Conflict of interest

The author(s) declared that this work 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) declared that generative AI was not 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

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. Tushita, Srivastava V, Singh RK. Transforming brain research: neuroimaging breakthroughs driven by AI. In: *AIP Conference Proceedings*, Vol. 3254. Melville, NY: AIP Publishing LLC (2025). p. 020021. doi: 10.1063/5.0248504

2. Bacon EJ, He D, Achi NBADA, Wang L, Li H, Yao-Digba PDZ, et al. Neuroimage analysis using artificial intelligence approaches: a systematic review. *Med Biol Eng Comput.* (2024) 62:2599–627. doi: 10.1007/s11517-024-03097-w

3. Gadgil AA, Selvakumar P, Gnanaselvi GS, Malathi G. AI in neuroimaging and brain analysis. In: *Transforming Neuropsychology and Cognitive Psychology with AI and Machine Learning*. Hershey, PA: IGI Global Scientific Publishing. (2025). p. 185–21. doi: 10.4018/979-8-3693-9341-3.ch008

4. Hassan MM, Yasmin F, Hasan M, Sharma C. Neuroimaging techniques: innovations and applications. In: *Brain Networks in Neuroscience: Personalization Unveiled Via Artificial Intelligence*. Gistrup: River Publishers (2025). p. 191–210. doi: 10.1201/9788770047371-9

5. Bhatt S, Sharma S, Bhadula S. August. NeuroAI: emerging artificial intelligence and image processing techniques in neuroscience for enhanced medical diagnosis of brain tumor. In: *2024 First International Conference on Pioneering Developments in Computer Science & Digital Technologies (IC2SDT)*. Piscataway, NJ: IEEE (2024). p. 204–9. doi: 10.1109/IC2SDT62152.2024.10696499

6. Du Y, Niu J, Xing Y, Li B, Calhoun VD. Neuroimage analysis methods and artificial intelligence techniques for reliable biomarkers and accurate diagnosis of schizophrenia: achievements made by Chinese scholars around the past decade. *Schizophrenia Bull.* (2025) 51:325–42. doi: 10.1093/schbul/sbae110

7. Szmyd B, Podstawka M, Wiśniewski K, Zaczkowski K, Puzio T, Tomczyk A, et al. AI-driven innovations in neuroradiology and neurosurgery: scoping review of current evidence and future directions. *Cancers*. (2025) 17:2625. doi: 10.3390/cancers17162625

8. Wright SN, Anticevic A. Generative AI for precision neuroimaging biomarker development in psychiatry. *Psychiatry Res.* (2024) 339:115955. doi: 10.1016/j.psychres.2024.115955

9. Zhang Y, Yu L, Lv Y, Yang T, Guo Q. Artificial intelligence in neurodegenerative diseases research: a bibliometric analysis since 2000. *Front Neurol.* (2025) 16:1607924. doi: 10.3389/fneur.2025.1607924

10. Li Y, Zhong Z. Decoding the application of deep learning in neuroscience: a bibliometric analysis. *Front Comput Neurosci.* (2024) 18:1402689. doi: 10.3389/fncom.2024.1402689

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 10 April 2025
DOI 10.3389/fmed.2025.1574428

Check for updates

OPEN ACCESS

EDITED BY
Deepti Deshwal,
Maharaja Surajmal Institute of Technology, India

REVIEWED BY
Annie Sujith,
Visvesvaraya Technological University, India
Tanveer Baig Z.,
Amity University Tashkent, Uzbekistan

*CORRESPONDENCE
Abdulmajeed Alqhatani
✉ aaalqhatni@nu.edu.sa

RECEIVED 10 February 2025
ACCEPTED 04 March 2025
PUBLISHED 10 April 2025

CITATION
Chandrasekaran S, Aarathi S, Alqhatani A, Khan SB, Quasim MT and Basheer S (2025) Improving healthcare sustainability using advanced brain simulations using a multi-modal deep learning strategy with VGG19 and bidirectional LSTM.
Front. Med. 12:1574428.
doi: 10.3389/fmed.2025.1574428

COPYRIGHT
© 2025 Chandrasekaran, Aarathi, Alqhatani, Khan, Quasim and Basheer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Improving healthcare sustainability using advanced brain simulations using a multi-modal deep learning strategy with VGG19 and bidirectional LSTM

Saravanan Chandrasekaran[1], S. Aarathi[2], Abdulmajeed Alqhatani[3]*, Surbhi Bhatia Khan[4,5,6], Mohammad Tabrez Quasim[7] and Shakila Basheer[8]

[1]Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, India, [2]Department of Computer Science and Engineering (Data Science), Dayananda Sagar College of Engineering, Bangalore, India, [3]Department of Information Systems, College of Computer Science and Information Systems, Najran University, Najran, Saudi Arabia, [4]School of Science, Engineering, and Environment, University of Salford, Salford, United Kingdom, [5]University Centre for Research and Development, Chandigarh University, Mohali, India, [6]Centre for Research Impact and Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, India, [7]Department of Computer Science and Artificial Intelligence, College of Computing and Information Technology, University of Bisha, Bisha, Saudi Arabia, [8]Department of Information Systems, College of Computer and Information Science, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

Background: Brain tumor categorization on MRI is a challenging but crucial task in medical imaging, requiring high resilience and accuracy for effective diagnostic applications. This study describe a unique multimodal scheme combining the capabilities of deep learning with ensemble learning approaches to overcome these issues.

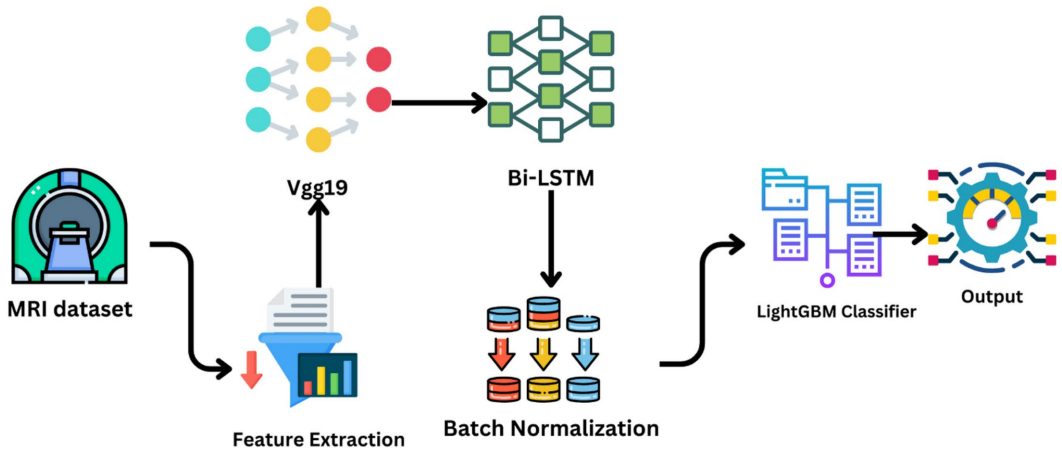
Methods: The system integrates three new modalities, spatial feature extraction using a pre-trained VGG19 network, sequential dependency learning using a Bidirectional LSTM, and classification efficiency through a LightGBM classifier.

Results: The combination of both methods leverages the complementary strengths of convolutional neural networks and recurrent neural networks, thus enabling the model to achieve state-of-the-art performance scores. The outcomes confirm the efficacy of this multimodal approach, which achieves a total accuracy of 97%, an F1-score of 0.97, and a ROC AUC score of 0.997.

Conclusion: With synergistic harnessing of spatial and sequential features, the model enhances classification rates and effectively deals with high-dimensional data, compared to traditional single-modal methods. The scalable methodology has the possibility of greatly augmenting brain tumor diagnosis and planning of treatment in medical imaging studies.

KEYWORDS

brain tumor classification, multi-modal learning, VGG19, bidirectional LSTM, LightGBM, MRI imaging, deep learning, ensemble learning



GRAPHICAL ABSTRACT

1 Introduction

Brain tumor segmentation from MRI images is an important component of medical imaging, and serious consequences follow for the diagnosis, treatment, and prognosis of the patient. The heterogeneity and complexity of brain tumors and the high-dimensionality of MRI data pose significant challenges to traditional diagnostic approaches. These include problems like tumor variability in appearance due to size, shape, and location, which can complicate detection and classification. Diagnosis with a human expert is generally cumbersome, subjective, and prone to error, and traditional machine learning approaches rely on manually designed features, which are prone to missing out on the complexities of MRI data. It uses advanced preprocessing techniques like image normalization and data augmentation to enhance training and model stability. Improvements in machine learning and deep learning enabled the automation and accurate classification of brain cancers. In this work, a new multi-modal approach is introduced that uses deep learning and ensemble learning methods to tackle these challenges, thus providing a scalable and effective approach to classifying brain tumors (1). Employing bidirectional long-term memory networks to represent sequential dependencies in MRI slices, deep convolutional neural networks to enhance spatial feature extraction, and LightGBM for high-dimensional data classification in an efficient way, the proposed VGG19-BiLSTM-LightGBM model. This multimodal approach synergistically improves brain tumor categorization by combining the strengths of each model component, thereby enhancing the model's ability to handle the intricacies of MRI data and improving diagnostic accuracy. Figure 1 shows the brain tumor images from the dataset.

The motivation for this work is the limitation imposed by existing techniques due to their inability to transcend such limitations. Because traditional diagnostic techniques, though effective within their confines, suffer from the heterogeneity of tumor size, shape, and location (2), and single-modal techniques account only for spatial or sequential characteristics and cannot harness the full richness of MRI image information, therefore, a method has to be developed those accounts for the interplay between spatial and sequential factors. This is capable of building more robust and precise classification by including these techniques as a multi-modal technique. Ensemble learning algorithms like LightGBM provides stable classification, effectively handling the high-dimensional data and aggregating the strengths of individual models (3).

This work centralizes to the creation of a multi-modal deep learning architecture for brain tumor classification that synergistically integrates the spatial and sequential features of MRI images. Spatial feature extraction was carried out through a pre-trained VGG19 model, thereby making it feasible and accurate for representing MRI images. To improve the model's capacity to learn the underlying patterns, a bidirectional LSTM layer is used to monitor temporal relationships among the extracted features (4).

This work is on the integration of multiple modalities, such as sequential modeling using Bidirectional LSTM and spatial feature learning using VGG19. The drawbacks of the traditional methods are alleviated through this work by giving an end-to-end solution to brain tumor classification. MRI image description becomes more realistic with the use of an integration of multiple modalities. The classification performance is further augmented by LightGBM being utilized as a final classifier to enable effective processing of high-dimensional data (5). Large and high-dimensional data can be handled using the proposed framework, which renders it easy to implement on actual healthcare challenges. High validation accuracy with minimal amounts of loss indicates its generalization capabilities to unseen data. The following sections of this paper are classified as given below. Section 2 gives an overview of the major research on brain tumor classification including deep learning and ensemble learning techniques. Section 3 provides a thorough explanation of the suggested methodology, i.e., data preparation, feature extraction, and classification. Section 4 discusses the experimental results, including performance metrics and comparisons with baseline models.

2 Literature review

The field of brain tumor classification from MRI scans has experienced tremendous expansion in the recent past with momentum building for the application of deep learning and machine learning techniques. Traditional methods in brain tumor diagnosis have employed close to all visual inspection by

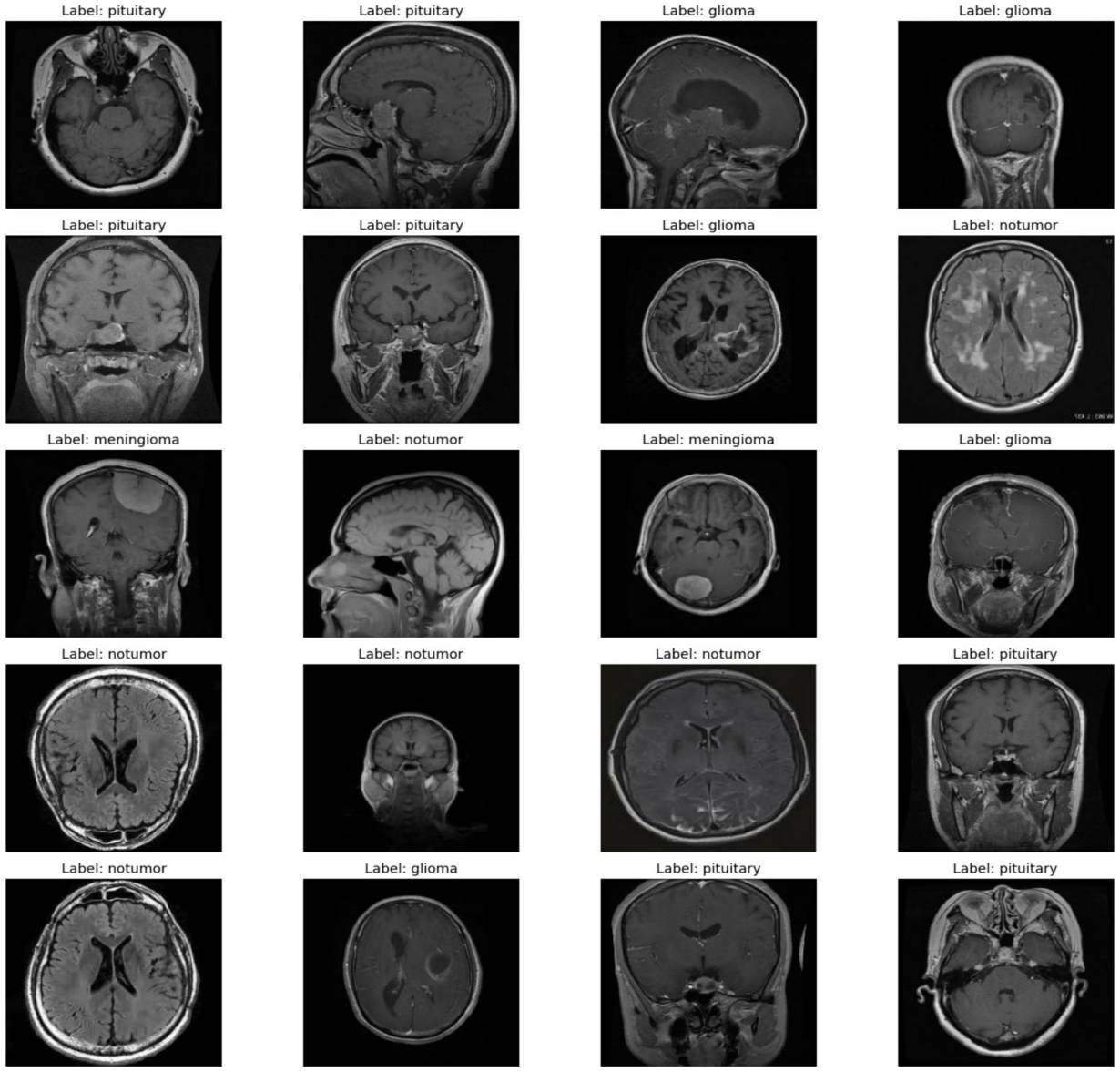


FIGURE 1
A sample of images from the dataset.

radiologists, not just time-consuming but also prone to human error (6). Such methods tend to employ extraction of inherent features like texture, shape, and intensity that might not reflect the complex patterns present in medical images. Therefore, there has been a move toward automated methods that take advantage of the strengths of deep learning to achieve improved accuracy and efficiency.

Convolutional Neural Networks have become a backbone of modern medical image analysis. Their ability to learn spatial features automatically from images has made them particularly suitable to applications such as tumor detection and classification (7). From pre-trained CNN models, VGG19, ResNet, and Inception can be broadly applied in medical imaging because they can generalize toward a large range of datasets. The early layers typically freeze, and the final layers are fine-tuned on the target dataset toward a specific application, such as the classification of brain tumors. This reduces not only the computational cost but also enhances performance with knowledge gained from large-scale datasets, such as ImageNet.

While CNNs excel at capturing spatial features, they may not fully exploit the sequential or temporal dependencies present in medical images. LSTMs are designed to represent sequential data, making them optimal for identifying temporal trends in medical images (8). Bidirectional LSTMs, which process data in both forward and backward directions, have been shown to further enhance performance by capturing more comprehensive dependencies. The combination of CNNs and LSTMs has been explored in various medical imaging tasks, including brain tumor classification, where it has demonstrated superior performance compared to standalone models. Table 1 shows the exiting studies through multiple techniques.

Ensemble learning methods have also found relevance in medical image analysis because they can enhance classification accuracy and robustness. Techniques such as Random Forests, Gradient Boosting,

TABLE 1 Existing studies from different techniques.

Study	Objective	Remark
Maqsood et al. (13)	To present an automated technique for precise brain tumor identification and classification by using deep learning and MRI.	The method achieved high accuracy (97.47 and 98.92%) and outperformed prior methods.
Jiang et al. (14)	To develop SwinBTS, a 3D medical image segmentation approach combining transformers and CNNs for brain tumor classification.	SwinBTS beat state-of-the-art algorithms on BraTS 2019, 2020, and 2021 datasets.
Zhu et al. (15)	Present a brain tumor segmentation approach that integrates deep semantics and edge information in multimodal MRI.	The method outperformed state-of-the-art methods on BraTS benchmarks.
Zhang et al. (16)	Introducing mmFormer: A Transformer-based approach to strong multimodal brain tumor segmentation with incomplete modalities.	mmFormer outperformed state-of-the-art approaches, particularly with missing modalities.
Razzaghi et al. (17)	A multimodal deep transfer learning system that can be used with MRI brain image processing should have domain flexibility.	The strategy outperformed equivalent algorithms on IBSR and Figshare datasets.
Ali et al. (18)	Analyze the progresses in brain tumor segmentation, feature extraction, and classification using MRI along with deep learning.	Highlights the move from traditional approaches to deep learning and hybrid methodologies.
Peng and Sun (19)	To propose AD-Net, an autonomous weighted dilated convolutional network for multimodal brain tumor feature extraction.	Achieved high Dice scores (0.90, 0.80, 0.76) on BraTS20 dataset.
Fang and Wang (20)	To propose MFF-DNet, a dual-path network for multi-modal feature fusion in brain tumor segmentation.	Achieved high precision (0.92 and 0.90) for whole tumor and core tumor segmentation.
Hossain et al. (21)	To propose a strategy for brain tumor segmentation using 3D U-Net and ResNet50 with image fusion.	Achieved high accuracy (98.96% for ResNet50, 97.99% for 3D U-Net).
Liu et al. (22)	To present SF-Net, a multi-task model for brain tumor segmentation leveraging segmentation-fusion.	Achieved higher segmentation accuracy than VAE-based approaches on BraTS 2020.
Prasad et al. (23)	To enhance medical imaging capabilities using a CNN-based approach for detecting and classifying brain tumors.	The proposed model achieves superior accuracy, recall, F1-score, and precision compared to traditional methods, contributing to more effective brain tumor analysis.
Kargar Nigjeh et al. (24)	To optimize brain tumor classification using deep learning models and advanced image enhancement techniques.	The study demonstrates high classification accuracy (95%) and provides insights into the strengths and limitations of various deep learning architectures for medical imaging.
Sharma et al. (25)	To improve efficiency in brain tumor categorization through a hybrid model approach.	The model achieves 97% classification accuracy by integrating multiple learning techniques, enhancing robustness in tumor classification.
Bibi et al. (26)	To address computational inefficiencies and improve classification accuracy through a transfer learning approach.	The InceptionV4 model achieves 98.7% accuracy, significantly improving diagnostic precision and reducing computation time.
Albalawi et al. (27)	To develop an advanced CNN architecture for more accurate and efficient brain tumor diagnosis.	The CNN model achieves an exceptional 99% accuracy, marking a major advancement in automated MRI analysis and early tumor detection.

and LightGBM combine the predictions of many models to produce more accurate and reliable results. LightGBM is specifically widely used because of its ability to work on enormous datasets and high-dimensional data (9). By combining deep learning models with ensemble techniques, scientists have been able to develop hybrid frameworks that leverage the strengths of both methods.

While great advances have been made, brain tumor categorization still presents some challenges. One of the most significant is that the tumors are very variable in how they look, which could vary greatly by size, shape, and even placement. All this variability makes it difficult to build a model that generalizes well over all datasets. Because the dimension of the MRI data is high, their computation presents serious challenges in particular when a lot of them is involved. Methods such as flipping, rotating and adjusting the brightness randomly, used to enlarge training data variety while preventing overfitting have commonly been employed for overcoming this difficulty (10). The third is interpretability in medical imaging models.

Multi-modal combination is a key component in improving categorization. Multi-modal techniques provide a more comprehensive explanation of the underlying issue by combining multiple data modalities, such as MRI images, clinical data, and genomic data. Multi-modal techniques have been shown to perform better than single-modal approaches in the categorization of brain tumors by complementarily gathering information from diverse data sources. It is presently known that the fusion of MRI images with clinical data, like the patient's age and medical history, improves classification performance and provides more individualized predictions (11). Brain tumor classification has greatly improved in the past few years due to advances in deep learning, ensemble learning, and multi-modal methods.

3 Methodology

The multi-modal nature of the proposed method for MRI-based brain cancer diagnosis is becoming increasingly popular. For sequence

modeling and feature extraction, it uses deep learning models like VGG19 and Bidirectional LSTM, for classification, it uses LightGBM. Figure 2 illustrates a step-by-step overview of the preferred model's approach.

3.1 Dataset description

The 7,023 MRI images of the human brain that make up the Brain Tumor MRI dataset are split into four categories: pituitary, meningioma, glioma, and no tumor. Glioma tumors are made up of glial cells, while Meningioma malignancies arise from the meninges, protective coverings of the brain and spinal cord. The Pituitary class contains cancers that originate in the pituitary gland, a small gland at the base of the brain that is responsible for the production of hormones. The "No Tumor" class contains normal brain scans to act as a control set for comparative analysis. The data were intentionally divided into training, validation, and testing sets in 70, 15, and 15% ratios, respectively. The ratio of splitting was aimed at achieving a trade-off between enough training data to learn the model parameters well and adequate validation and test data to analyse the performance and generalizability of the model comprehensively. The significant portion dedicated to training ensures deep learning models, which demand huge amounts of data, get well-trained. Equal partitioning of the rest of the data for validation and testing helps refine model parameters and test the model on data not seen by it, reducing the risk of overfitting. The method also ensures that the evaluation measures capture the model's ability to function under varying conditions, thus offering a truer measure of its potential effectiveness in actual use. The wide scope of categorization ensures total research over a wide range of common situations of the brain, thus enhancing representativeness when the model is used in practical applications. The dataset, though, has its limitations in the shape of potential class imbalance and heterogeneity in tumor locations and sizes, which could hinder learning as well as predictive capacity. Three primary sources make up this dataset: the SARTAJ dataset, which initially consisted of glioma images but contained inconsistencies that led to their replacement with images sourced from figshare; the Br35H dataset, which provides images for the "No Tumor" class; and figshare, which offers images for glioma, meningioma, and pituitary tumors. It is thought that this data would make it possible to design automated systems for the classification of brain cancers with proper early detection and a proper diagnosis. It has been divided into training and test sets, with images resized to 224 × 224 pixels for deep learning models such as VGG19. The dataset's size and heterogeneity render it a valuable source of information upon which researchers and medical imaging professionals can formulate generalizable and robust brain tumor classification algorithms.

3.2 Data preprocessing

The first preprocessing operation is scaling of images. The MRI images in the dataset are resized to a uniform size of 224 × 224 pixels. Standardization is necessary because deep learning models like VGG19 need to have fixed input sizes. Resizing enables all images to be compatible with the model architecture, thus enabling effective batch processing during training. Resizing also reduces the computational complexity by downsampling high-resolution images without significantly reducing their quality. Equation 1 shows the resizing of images.

$$I' = \text{resize}(I, h, w) \tag{1}$$

The resized images are then normalized, which is the process of scaling pixel values to a particular range. In this case, pixel values are normalized to the range [0, 1] by dividing the pixel intensity by 255. Normalization is necessary since it ensures the input data have a fixed scale, which improves the convergence of the model while training. If not normalized, the model will fail to learn since the magnitudes of pixel values vary from image to image. Equation 2 illustrates the formula to normalize the images.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{2}$$

In order to improve the strength and variety of the dataset, data augmentation techniques are applied. Data augmentation is artificially

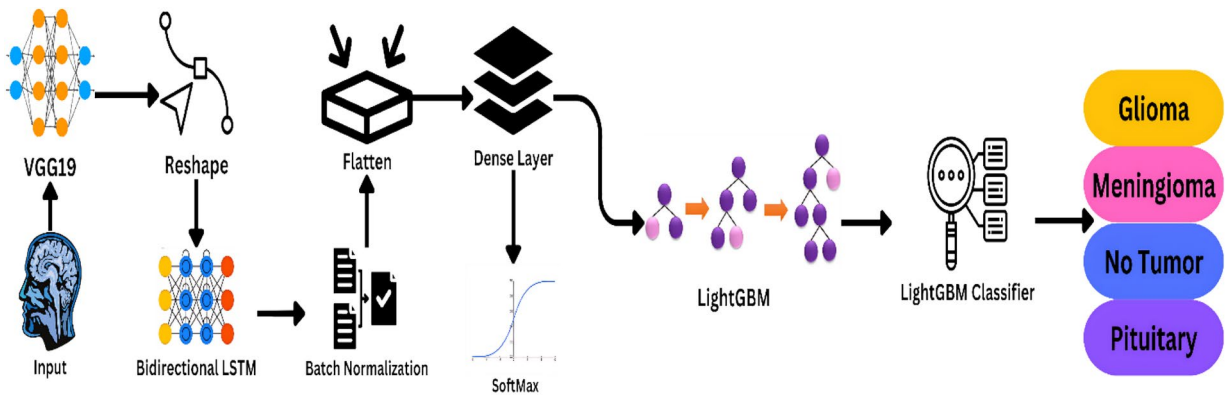


FIGURE 2
Workflow of the proposed model the framework is ideal for real clinical applications as it strives for high accuracy and generality.

conducted to enlarge the size of the training dataset by creating multiple copies of the original images. This process not only addresses the issue of limited data in medical imaging but also simulates varying imaging conditions, which helps in building a robust model. The data augmentation techniques applied in this system not just random horizontal and vertical flips, but also random horizontal and vertical flip, which mimic different brain orientations; random change in brightness, which introduces lighting variability; random change in contrast, which changes the difference in intensity of pixels; random change in saturation, which changes the colour intensity; and random change in hue, which changes the tonal quality of images. These transformations are essential for training the model to recognize tumors under different imaging conditions and enhance its ability to generalize across new, unseen datasets. Equation 3 represent the mean and standard deviation of the pixel values in the image. Equation 4 applies a flip transformation along a specified axis (horizontal or vertical) to the image. Equation 5 brightens the image by adding a constant β, being possibly positive (to brighten) or negative (to darken). Equation 6 adjusts the pixel values of I''' to change the contrast.

$$I'''' = I''' \cdot R(\theta) \tag{3}$$

$$I''''' = \text{flip}(I''', \text{axis}) \tag{4}$$

$$I'''''' = I''' + \beta \tag{5}$$

$$I''''''' = \alpha \cdot (I''' - \mu) + \mu \tag{6}$$

Data preprocessing pipeline is built to transform raw MRI images to an appropriate form for deep learning models. By resizing, normalizing, augmenting, and organizing the data, the pipeline enables the model to learn and generalize effectively to unseen new data. These preprocessing steps are important to achieve high accuracy and robustness in brain tumor classification and are therefore an integral part of the proposed approach.

3.3 Model architecture

The suggested classification system of brain tumors uses a combination of deep learning models to achieve great accuracy and robustness. The construction of automated approaches for the classification of brain cancers with sufficient early detection and precise diagnosis is anticipated to be enabled by dataset. For deep learning models such as VGG19, it has been split into training and test sets, and the photographs have been resized to 224 × 224 pixels. Due to the volume and diversity of the dataset, researchers and medical image professionals can utilize it to construct valid and generalisable analysis. Figure 3 shows the Model Architecture of VGG19-BiLSTM-LightGBM Framework.

The technique of converting raw MRI scans into an applicable set of features suitable for classification is referred to as feature extraction, and it is the initial step of the deep learning pipeline. A pre-trained VGG19 model is used to do this. VGG19 is a very deep convolutional neural network (CNN) architecture that has been widely applied in computer vision tasks due to its capability to extract hierarchical features from images. The architecture of VGG19 comprises 19 layers, including 16 layers of convolutional layers, 3 of fully connected layers, and 5 max-pooling layers. On this model, they apply pre-training from ImageNet dataset, incorporating over 1 million images in 1,000 categories. The pre-trained model gives the feature of identifying general features, such as edges, textures, and shapes, which can be further fine-tuned for any other task. In this case, it is for medical image analysis. Equation 7 gives the output size of a convolutional layer.

$$O = \frac{W - K + 2P}{S} + 1 \tag{7}$$

Transfer learning is employed in the suggested framework to fine-tune the VGG19 model for brain tumor classification. Transfer learning is the reuse of a pre-trained model with fine-tuning for a task. The model is set up to receive input images of size 224 × 224 pixels. The pre-trained weights are imported, to focus on extracting the most relevant features for brain tumor classification, only the early convolutional layers of the model are frozen, allowing the deeper layers, which are more specific to the task at hand, to adjust during the training process. This keeps the model to retain the common features learned from ImageNet while learning task-specific features in the later layers. The VGG19 model processes the input MRI images and extracts high-level spatial features from its final convolutional layer. These features represent the most discriminative aspects of the images, such as tumor boundaries, texture, and intensity variations. The output of the VGG19 model is a feature map with dimensions 7 × 7 × 512, which is then passed to the next stage of the pipeline for further processing. To effectively use both sequential and spatial information, a Bidirectional LSTM layer has been added within the pipeline. LSTMs are a family of RNNs, the architecture of which is well-suited to the modeling of sequence data. Adding a Bidirectional LSTM allows the model to not only extract forward temporal dynamics but also backward dynamics, giving complete insight into sequence data. To enhance the ability of the model to learn the inherent patterns, a bidirectional LSTM layer is employed to track temporal relationships between the extracted features (4). This project is on the fusion of various modalities, like sequential modeling by Bidirectional LSTM and spatial feature learning by VGG19. The shortcomings of the conventional methods are overcome through this project with the provision of an end-to-end solution to brain tumor classification. MRI image description is made more realistic with the provision of an integration of various modalities. The classification efficiency is also enhanced through the use of LightGBM as a final classifier for efficient handling of high-dimensional data (5) using Equation 8. High-dimensional and large data are handled using the proposed framework, making it simple to deploy on real-life healthcare problems.

$$\widehat{x_i} = \gamma \left(\frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \right) + \beta \tag{8}$$

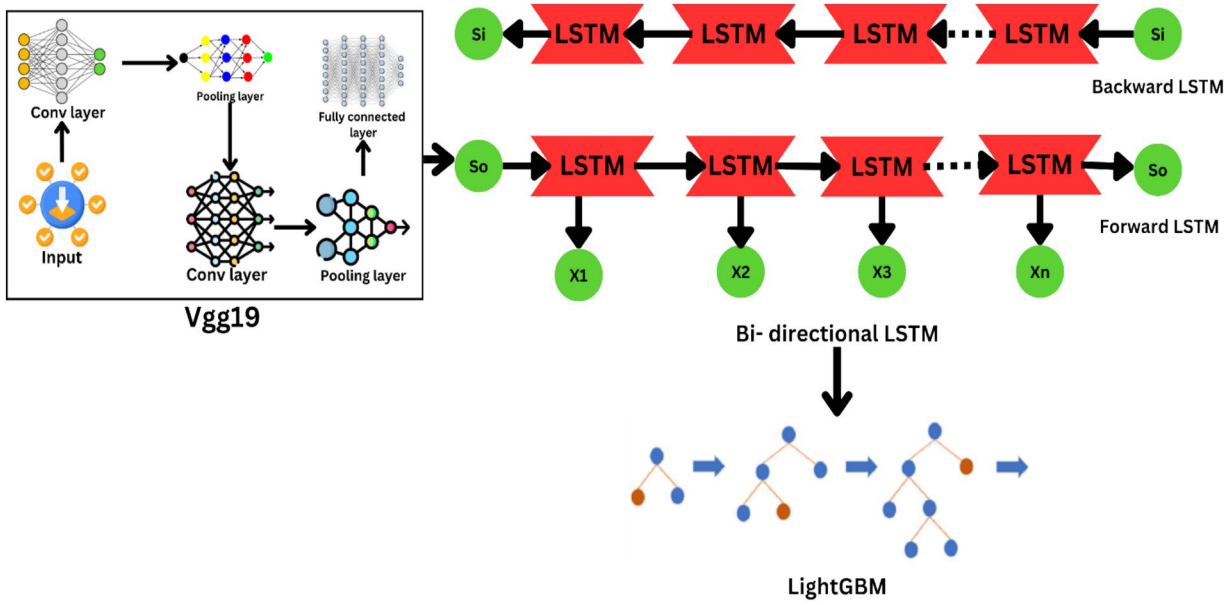


FIGURE 3
Architecture of the VGG19-BiLSTM-LightGBM framework.

The last layer of classification takes the flattened output of the LSTM layer in the form of a 1D vector. It does this so that the features are brought in a form that allows easy classification. It is also designed for progressive learning. It allows the model to incrementally update its knowledge base whenever there is new information without rigid retraining needs. The inclusion of the LightGBM classifier within the model is highly significant in this case, as this classifier supports online learning environments. This aspect allows the model to update continuously with new data, hence enhancing its prediction with the passage of time. This is a highly significant feature in medical imaging, where shifting patterns of data require flexible models that can update with minimal downtime and computational costs.

To find the brain tumors consistently, the features that are extracted are used to train a LightGBM classifier, which is the final step in the deep learning process. A very good gradient boosting library capable of handling large high-dimensional data is known as LightGBM. The trained VGG19 and LSTM layers are used for building another feature extraction model. The gradient descent update rule is found in Equation 9. The logistic loss function for binary classification is found in Equation 10.

$$\theta := \theta - \eta \cdot \nabla_{\theta} J(\theta) \tag{9}$$

$$l(\hat{y}, y) = \sum_{i=1}^{n} \left[y_i \log\left(1 + e^{-\hat{y}_i}\right) + (1 - y_i) \log\left(1 + e^{\hat{y}_i}\right) \right] \tag{10}$$

The features extracted are standardized with StandardScaler, thus obtaining a zero mean and unit variance for all the variables. This step is essential for maximizing the LightGBM classifier's performance since it ensures that each feature contributes evenly to the classification process. With default hyperparameters, i.e., 200 estimators and a learning rate of 0.05, the LightGBM classifier is trained on scaled features. The retrieved features are used to train the algorithm to categorize different types of tumors. LightGBM is employed because it can generate precise and reliable predictions and is effective at managing big datasets. The operational flow and interdependencies between the various components of this multi-modal deep learning technique for MRI-based brain tumor classification are outlined sequentially in Algorithm 1.

The training process of the proposed VGG19-BiLSTM-LightGBM framework involves a multi-stage pipeline designed to optimize the model's performance and generalization capabilities. This uses pre-trained VGG19 as the spatial feature extractor from the MRI images with all layers frozen so that weights learned during ImageNet can be preserved. Features from these layers are passed to the Bidirectional LSTM layer, which then encodes the temporal dependencies, followed by repeated processes of Batch Normalization and Flattening so that the data is made ready for classification. Equations 11, 12 can be used to compute the accuracy and precision of the model, respectively, which are two key parameters that can establish the efficiency of the model for real-world implementation.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{12}$$

The whole pipeline is trained over the Brain Tumor MRI Dataset. To enhance training data variations, the entire dataset has methods applied that consist of random flips in any two planes and various combinations of changing brightness and contrast. Equations 13, 14 compute recall and F1-score thus yielding more criteria that are

Input:
- A set of MRI images

Output:
- Classifications of brain tumors

1. Preprocess the MRI_images:
   1.1. Resize images to 224x224 pixels
   1.2. Normalize pixel values to [0, 1]
   1.3. Apply data augmentation (e.g., random rotations, flipping)

2. Extract features using pre-trained VGG19:
   2.1. Load VGG19 model, discard final layers
   2.2. Pass each image through VGG19 to get feature maps

3. Sequential modelling with Bidirectional LSTM:
   3.1. Reshape feature maps to sequences
   3.2. Feed sequences into a Bidirectional LSTM to obtain feature vectors

4. Classify with LightGBM:
   4.1. Flatten LSTM output to prepare feature vectors
   4.2. Normalize feature vectors
   4.3. Train LightGBM classifier on feature vectors
   4.4. Predict tumor type using the trained LightGBM model

ALGORITHM 1
Multi-modal deep learning method for classifying brain tumors based on MRI.

essential in judging performance concerning the positive values correctly discovered but at some expense in recall/precision ratio (12).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{13}$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

The model is trained on parameters like accuracy, precision, recall, F1-score, and ROC AUC so that it is able to classify the brain tumors robustly and accurately. The long training process makes sure that the model learns not only to be precise but also to be generalizable in nature and hence usable in real-world clinical practice.

4 Results

The proposed VGG19-BiLSTM-LightGBM model for brain cancer classification was outstanding in classifying the Brain Tumor MRI Dataset, subjecting it to being able to handle the uncertainty and complexity of the MRI images. The model achieved a training accuracy of 98.69%, validation accuracy of 96.64%, and total test accuracy of 97%, evidence of its ability to generalize to unseen data. Precision, recall, and F1-score metrics also testified to the stability of the model, with its performance being more than 0.92 across all classes. Interestingly, the "No Tumor" and "Pituitary" classes achieved 100% accuracy and recall, while the Glioma and Meningioma classes achieved comparatively lower but still outstanding performance because of their visual similarity. Figure 4 illustrates the categorization report of the suggested model according to all four classes.

The model's discriminative ability was confirmed by an ROC AUC score of 0.997, indicating its strong capability to distinguish between different tumor types. Figure 5 shows the ROC AUC score of all four classes.

Error metrics, including Mean Squared Error (MSE = 0.01), Root Mean Squared Error (RMSE = 0.10), and Mean Absolute Error (MAE = 0.10), further underscored the model's accuracy and reliability. These results demonstrate that the integration of spatial feature extraction (VGG19), sequential modeling (Bidirectional LSTM), and robust classification (LightGBM) provides a powerful framework for brain tumor classification, outperforming traditional single-modal approaches. Figure 6 shows the error metrices of the proposed model.

The confusion matrix indicated that the majority of the misclassifications were between the Glioma and Meningioma classes, consistent with the difficulty caused by their visual similarity. The overall misclassification rate was low, and the model performed high accuracy in all classes. The superior performance of the proposed framework compared to baseline procedures, including isolated VGG19 and Random Forest classifiers, supports the advantage of the combination of deep learning and ensemble learning methods. Figure 7 displays the confusion matrix of the utilized dataset.

These findings are important to clinical use as the model has the potential to assist radiologists in more precise and effective diagnosis of brain tumors. But the task can be expanded with other modalities

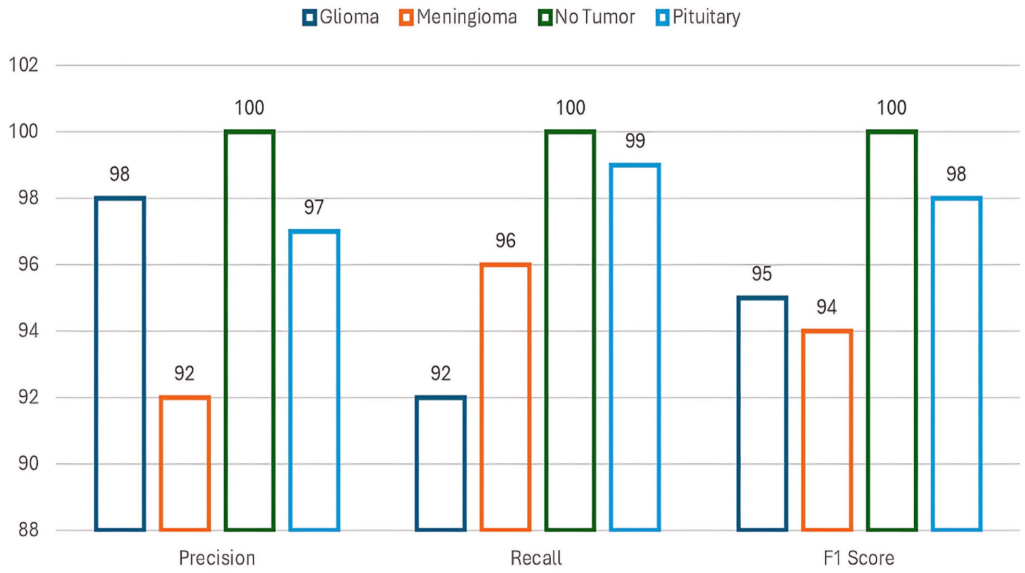


FIGURE 4
Classification report of the proposed model.

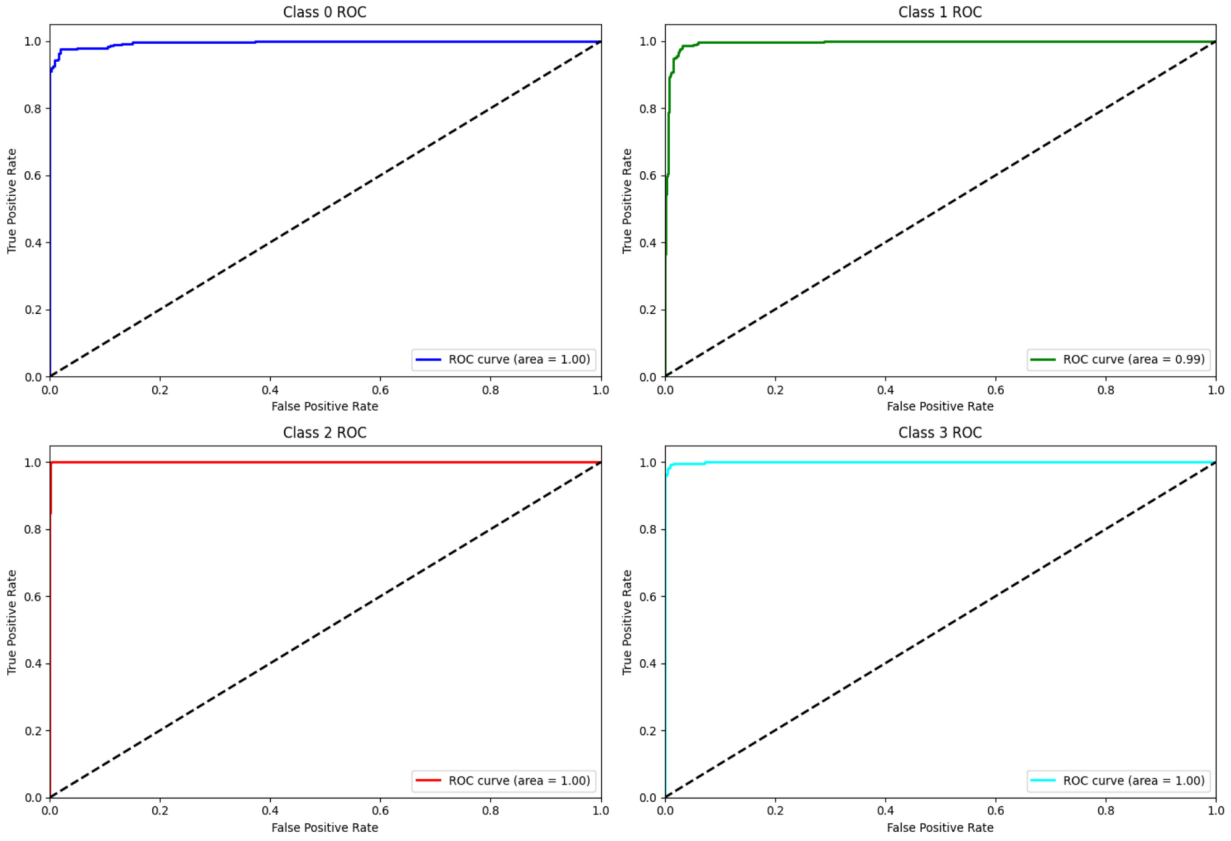


FIGURE 5
ROC and AUC curve.

being added, e.g., clinical data or genomic data, to further improve the performance of the model. Table 2 shows the comparison study of many Techniques.

An important development in brain tumor classification is the VGG19-BiLSTM-LightGBM framework, which provides a reliable and expandable solution for medical imaging applications. The VGG19-

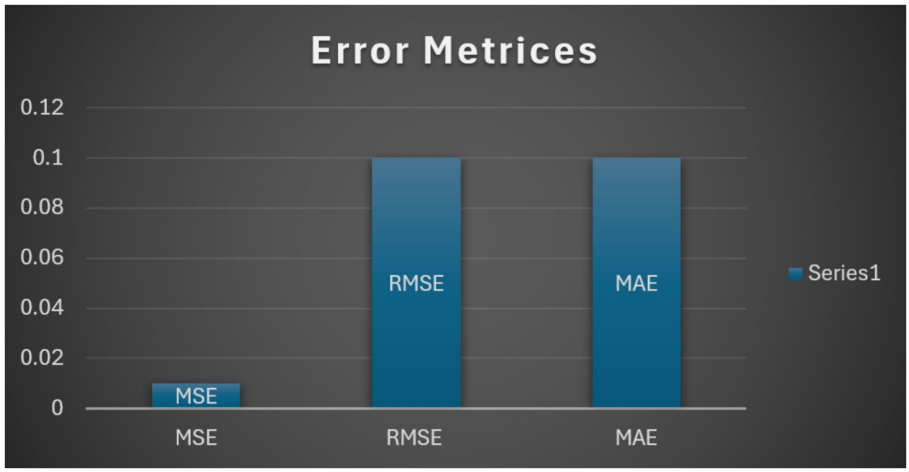


FIGURE 6
Error metrices.

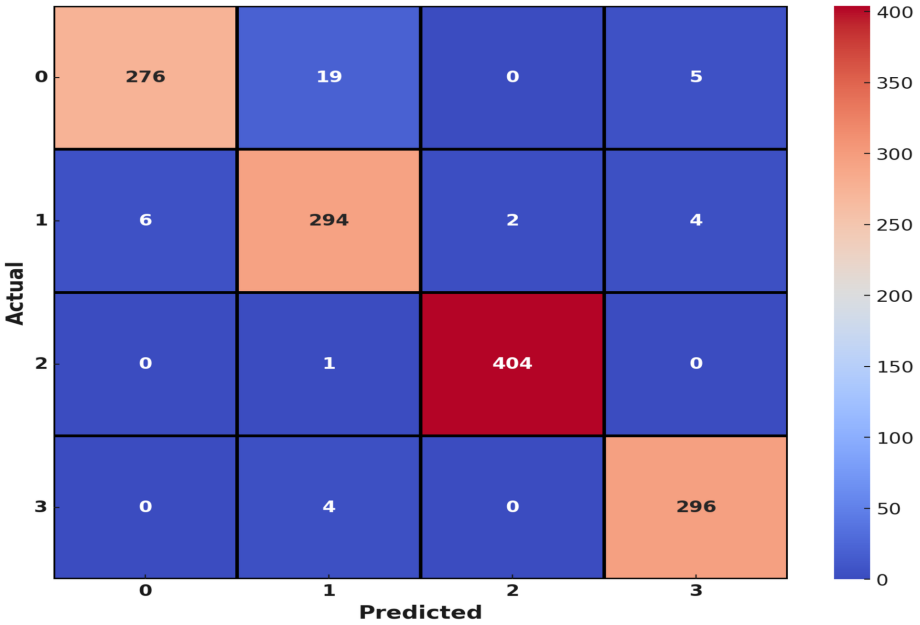


FIGURE 7
Confusion matrix.

BiLSTM-LightGBM model achieves excellent accuracy but requires extensive processing resources due to its complex construction. This can result in longer training times and higher costs, which might not be desirable for most clinical scenarios, particularly real-time scenarios. To address this, techniques such as pruning and quantization could be used to reduce model size and speed up inference times without sacrificing accuracy.

5 Discussion

In balancing for potential class imbalances in the MRI data sets, a common problem in medical images due to different rates of occurrence of different types of tumors, application of data augmentation techniques and weighted loss function assists in achieving balanced model training and prevents class bias toward majority classes. Scalability of the VGG19-BiLSTM-LightGBM architecture is beyond brain tumor classification. The model's structure is inherently flexible enough so that it may be utilized to process a wide variety of sickness classes over a large number of imaging modalities. The same structural concepts could reasonably be applied with the goal of classifying chest X-ray abnormalities or skin imaging lesions. This adaptability is primarily attributed to the VGG19 component of the model, which is widely renowned for its capacity to extract informative features from the majority of images, and the very flexible nature of the LSTM and LightGBM components that can be fine-tuned to detect and classify various pathological

TABLE 2 Comparison study from different techniques.

Study	Techniques	Accuracy
Pan et al. (28)	Convolutional neural networks (CNNs)	96%
Filatov and Yar (29)	EfficientNetB1	89.55%
Ma et al. (30)	CNN	80%
Shilaskar et al. (31)	Extreme gradient boosting (XG Boost)	92.02%
Binish et al. (32)	CBAM	96.70%
Upadhyay et al. (33)	CNN	91%
Ullah et al. (34)	SVM	95.73%
Pandiyaraju et al. (35)	LinkNet architecture	95.84%
Zhu (36)	FT-CNN	96%
Asiri et al. (37) (ML Model)	SVM	95.3%
Stadlbauer et al. (38) (ML Model)	Random forest	0.87%
Proposed model	VGG19-BiLSTM-LightGBM framework	97%

features with high efficiency. This approach should be applied in low-resource environments. Techniques such as model simplification, quantization, and the use of light-weight neural networks can sufficiently reduce the computational requirements. These parameters are important in maintaining the diagnostic integrity of the model for all categories of tumors. In addition to the computational efficiency and model complexity, there is an inherent trade-off between accuracy and computational requirement. The VGG19, Bidirectional LSTM, and LightGBM together, although computationally expensive, are warranted by the size of accuracy gain and medical diagnostics stability needed. The architecture's complexity makes it challenging to use in the clinic with real-time requirements.

Existing model implementation into clinical environments may be compromised by latency in processing and loading demands. Future development will center on refining these components to enable real-time analysis, possibly by model reduction or employing more effective processing methods like model quantization and pruning. Future studies will also continue to explore scalability, namely how this system can be adapted or scaled to support different types of tumors or medical imaging tests. This can involve training on larger, more heterogeneous sets of data or modifying the architecture to more effectively encode unique features of individual medical diseases, increasing model flexibility and utility across a broad array of clinical applications.

6 Conclusion

The paper offers an important contribution to the brain tumor identification from MRI images using a VGG19-BiLSTM-LightGBM model. The multi-modal approach overcomes complexity and heterogeneity, which are inherently linked to medical imaging data, by using space feature extraction, sequential modeling, and high-performance classification algorithms. Deploying a pre-trained VGG19 model for spatial feature extraction, a Bidirectional LSTM to process sequential information, and LightGBM for efficient and accurate classification, the model improves on diagnostic capability.

With a strong output of 98.69% training accuracy, 96.64% validation accuracy, and 97% test accuracy, it excels over currently available methods such as the VGG19 when isolated and the Random Forest classifier. Such a paradigm, in addition to lowering the chances of error in diagnosis, also aids radiologists in successfully diagnosing brain cancers efficiently and in a timely manner, enhancing patient care. Future upgrades can involve the integration of new data types, e.g., clinical or genetic data, to improve the accuracy as well as the robustness of the model. Additionally, employing explainable AI techniques can enhance the interpretability of the model as a more practical tool for application in clinical contexts. VGG19-BiLSTM-LightGBM is a cost-effective and effective approach to classifying brain tumors and can potentially transform computer-aided diagnosis in radiology.

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

SC: Conceptualization, Data curation, Writing – original draft. SA: Data curation, Investigation, Writing – original draft. AA: Formal analysis, Resources, Writing – review & editing. SK: Formal analysis, Methodology, Supervision, Writing – review & editing. MQ: Formal analysis, Investigation, Software, Writing – review & editing. SB: Formal analysis, Validation, Visualization, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research is supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R195), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors are thankful to the Deanship of Graduate Studies and Scientific Research at Najran University for funding this work under the Growth Funding Program grant code (NU/GP/SERC/13/575). The authors are thankful to the Deanship of Graduate Studies and Scientific Research at University of Bisha for supporting this work through the Fast-Track Research Support Program.

Acknowledgments

The authors acknowledge this research is supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R195), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors are thankful to the

Deanship of Graduate Studies and Scientific Research at Najran University for funding this work under the Growth Funding Program grant code (NU/GP/SERC/13/575). The authors are thankful to the Deanship of Graduate Studies and Scientific Research at University of Bisha for supporting this work through the Fast-Track Research Support Program.

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. Hussain D, al-masni MA, Aslam M, Sadeghi-Niaraki A, Hussain J, Gu YH, et al. Revolutionizing tumor detection and classification in multimodality imaging based on deep learning approaches: methods, applications and limitations. *J Xray Sci Technol*. (2024) 32:857–911. doi: 10.3233/xst-230429

2. Xie Y, Zaccagna F, Rundo L, Testa C, Agati R, Lodi R, et al. Convolutional neural network techniques for brain tumor classification (from 2015 to 2022): review, challenges, and future perspectives. *Diagnostics*. (2022) 12:1850. doi: 10.3390/diagnostics12081850

3. Mienye ID, Sun Y. A survey of ensemble learning: concepts, algorithms, applications, and prospects. *IEEE Access*. (2022) 10:99129–49. doi: 10.1109/access.2022.3207287

4. Liu X, Mu J, Pang M, Fan X, Zhou Z, Guo F, et al. A male patient with hydrocephalus via multimodality diagnostic approaches: a case report. *Cyborg Bionic Syst*. (2024) 5:0135. doi: 10.34133/cbsystems.0135

5. Li Q, You T, Chen J, Zhang Y, Du C. LI-EMRSQL: linking information enhanced Text2SQL parsing on complex electronic medical records. *IEEE Trans Reliab*. (2024) 73:1280–90. doi: 10.1109/TR.2023.3336330

6. Arabahmadi M, Farahbakhsh R, Rezazadeh J. Deep learning for smart healthcare—A survey on brain tumor detection from medical imaging. *Sensors*. (2022) 22:1960. doi: 10.3390/s22051960

7. Soomro TA, Zheng L, Afifi AJ, Ali A, Soomro S, Yin M, et al. Image segmentation for MR brain tumor detection using machine learning: A review. *IEEE Rev Biomed Eng*. (2023) 16:70–90. doi: 10.1109/rbme.2022.3185292

8. Montaha S, Azam S, Rafid AKMRH, Hasan MZ, Karim A, Islam A. Time Distributed-CNN-LSTM: A hybrid approach combining CNN and LSTM to classify brain tumor on 3D MRI scans performing ablation study. *IEEE Access*. (2022) 10:60039–59. doi: 10.1109/access.2022.3179577

9. Alomar K, Aysel HI, Cai X. Data augmentation in classification and segmentation: a survey and new strategies. *J Imaging*. (2023) 9:46. doi: 10.3390/jimaging9020046

10. Yao Z, Wang H, Yan W, Wang Z, Zhang W, Wang Z, et al. Artificial intelligence-based diagnosis of Alzheimer's disease with brain MRI images. *Eur J Radiol*. (2023) 165:110934. doi: 10.1016/j.ejrad.2023.110934

11. Liu Y, Mu F, Shi Y, Chen X. SF-net: A multi-task model for brain tumor segmentation in multimodal MRI via image fusion. *IEEE Signal Processing Letters*. (2022) 29:1799–803. doi: 10.1109/lsp.2022.3198594

12. Pan H, Wang Y, Li Z, Chu X, Teng B, Gao H. A complete scheme for multi-character classification using EEG signals from speech imagery. *IEEE Trans Biomed Eng*. (2024) 71:2454–62. doi: 10.1109/TBME.2024.3376603

13. Maqsood S, Damaševičius R, Maskeliūnas R. Multi-modal brain tumor detection using deep neural network and multiclass SVM. *Medicina*. (2022) 58:1090. doi: 10.3390/medicina58081090

14. Jiang Y, Zhang Y, Lin X, Dong J, Cheng T, Liang J. SwinBTS: A method for 3D multimodal brain tumor segmentation using Swin transformer. *Brain Sci*. (2022) 12:797. doi: 10.3390/brainsci12060797

15. Zhu Z, He X, Qi G, Li Y, Cong B, Liu Y. Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI. *Inf Fusion*. (2023) 91:376–87. doi: 10.1016/j.inffus.2022.10.022

16. Zhang Y, He N, Yang J, Li Y, Wei D, Huang Y, et al. mmFormer: multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. *MICCAI*. (2022) 13435:107–17. doi: 10.1007/978-3-031-16443-9_11

17. Razzaghi P, Abbasi K, Shirazi M, Rashidi S. Multimodal brain tumor detection using multimodal deep transfer learning. *Appl Soft Comput*. (2022) 129:109631:109631. doi: 10.1016/j.asoc.2022.109631

18. Ali S, Li J, Pei Y, Khurram R, Rehman KU, Mahmood T. A comprehensive survey on brain tumor diagnosis using deep learning and emerging hybrid techniques with multi-modal MR image. *Arch Comput Methods Eng*. (2022) 29:4871–96. doi: 10.1007/s11831-022-09758-z

19. Peng Y, Sun J. The multimodal MRI brain tumor segmentation based on AD-net. *Biomed Signal Proc Control*. (2023) 80:104336. doi: 10.1016/j.bspc.2022.104336

20. Fang L, Wang X. Brain tumor segmentation based on the dual-path network of multi-modal MRI images. *Pattern Recogn*. (2022) 124:108434. doi: 10.1016/j.patcog.2021.108434

21. Hossain E, Shazzad Hossain M, Selim Hossain M, al Jannat S, Huda M, Alsharif S, et al. Brain tumor auto-segmentation on multimodal imaging modalities using deep neural network. *Comput Mat Continua*. (2022) 72:4509–23. doi: 10.32604/cmc.2022.025977

22. Singh RB, Datta A. Dimensionality reduction and gradient boosting for in-vivo hyperspectral brain image classification. In *IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)* (2024), pp. 1–6.

23. Prasad CR, Srividya K, Jahnavi K, Srivarsha T, Kollem S, Yelabaka S. Comprehensive CNN model for brain tumour identification and classification using MRI images. In *2024 IEEE International Conference for Women in Innovation, Technology & Entrepreneurship (ICWITE)* (2024), pp. 524–8.

24. Kargar Nigjeh M, Ajami H, Mahmud A, Hoque MSU, Umbaugh SE. Comparative analysis of deep learning models for brain tumor classification in MRI images using enhanced preprocessing techniques. *Appl Digital Image Proc*. (2024) XLVII:318. doi: 10.1117/12.3028318

25. Sharma A, Mittal S. Hybrid deep learning model for enhanced brain tumor classification using VGG19, LSTM, and SVM algorithms. In *Global Conference on Communications and Information Technologies (GCCIT)* (2024), pp. 1–5.

26. Bibi N, Wahid F, Ma Y, Ali S, Abbasi IA, Alkhayyat A, et al. A transfer learning-based approach for brain tumor classification. *IEEE Access*. (2024) 12:111218–38. doi: 10.1109/access.2024.3425469

27. Albalawi E, Thakur A, Dorai DR, Bhatia Khan S, Mahesh TR, Almusharraf A, et al. Enhancing brain tumor classification in MRI scans with a multi-layer customized convolutional neural network approach. *Front Comput Neurosci*. (2024) 18:1418546. doi: 10.3389/fncom.2024.1418546

28. 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

29. Filatov D, Yar GNAH. Brain tumor diagnosis and classification via pre-trained convolutional neural networks. CSH (2022).

30. Ma Q, Zhang Y, Hu F, Zhou H, Hu H. Nip it in the bud: the impact of China's large-scale free physical examination program on health care expenditures for elderly people. *Hum Soc Sci Commun*. (2025) 12:27. doi: 10.1057/s41599-024-04295-5

31. Shilaskar S, Mahajan T, Bhatlawande S, Chaudhari S, Mahajan R, Junnare K. Machine learning based brain tumor detection and classification using HOG feature descriptor. In *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)* (2023), pp. 67–75.

32. Binish MC, Raj RS, Thomas V. Brain Tumor Classification using Multi-Resolution Averaged Spatial Attention Features with CBAM and Convolutional Neural Networks. In *2024 1st International Conference on Trends in Engineering Systems and Technologies (ICTEST)* (2024), pp. 1–7.

33. Upadhyay P, Saifi S, Koul J, Rani R, Bansal P, Sharma A. Classification of Brain Tumors Using Augmented MRI Images and Deep Learning. In *2024 2nd International Conference on Computer, Communication and Control (IC4)* (2024), pp. 1–7.

34. Ullah S, Ahmad M, Anwar S, Khattak MI. An intelligent hybrid approach for brain tumor detection. *Pakistan J Eng Technol*. (2023) 6:42–50. doi: 10.51846/vol6iss1pp34-42

35. Pandiyaraju V, Ganapathy S, Senthil Kumar AM, Jesher Joshua M, Ragav V, Sree Dananjay S, et al. A new clinical diagnosis system for detecting brain tumor using integrated ResNet_Stacking with XGBoost. *Biomed Signal Proc Control*. (2024) 96:106436. doi: 10.1016/j.bspc.2024.106436

36. 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

37. Asiri AA, Khan B, Muhammad F, Alshamrani HA, Alshamrani KA, Irfan M, et al. Machine learning-based models for magnetic resonance imaging (MRI)-based brain tumor classification. *Int Autom Soft Comput*. (2023) 36:299–312. doi: 10.32604/iasc.2023.032426

38. Stadlbauer A, Marhold F, Oberndorfer S, Heinz G, Buchfelder M, Kinfe TM, et al. Radiophysiomics: brain tumors classification by machine learning and physiological MRI data. *Cancers*. (2022) 14:2363. doi: 10.3390/cancers14102363

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 20 May 2025
DOI 10.3389/fmed.2025.1577474

Check for updates

OPEN ACCESS

EDITED BY
Deepti Deshwal,
Maharaja Surajmal Institute of Technology, India

REVIEWED BY
Mahmood Safaei,
University of Akron, United States
Eid Rehman,
University of Mianwali, Pakistan

*CORRESPONDENCE
Mosleh Hmoud Al-Adhaileh
madaileh@kfu.edu.sa
Theyazn H. H. Aldhyani
taldhyani@kfu.edu.sa

RECEIVED 15 February 2025
ACCEPTED 14 April 2025
PUBLISHED 20 May 2025

CITATION
Al-Adhaileh MH, Ahmad S, Alharbi AA, Alarfaj M, Dhopeshwarkar M and Aldhyani THH (2025) Diagnosis of epileptic seizure neurological condition using EEG signal: a multi-model algorithm. *Front. Med.* 12:1577474. doi: 10.3389/fmed.2025.1577474

COPYRIGHT
© 2025 Al-Adhaileh, Ahmad, Alharbi, Alarfaj, Dhopeshwarkar and Aldhyani. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Diagnosis of epileptic seizure neurological condition using EEG signal: a multi-model algorithm

Mosleh Hmoud Al-Adhaileh[1,2]*, Sultan Ahmad[3], Alhasan A. Alharbi[4], Mohammed Alarfaj[1,5], Mukta Dhopeshwarkar[4] and Theyazn H. H. Aldhyani[6]*

[1]King Salman Center for Disability Research, Riyadh, Saudi Arabia, [2]Deanship of E-Learning and Information Technology, King Faisal University, Al-Ahsa, Saudi Arabia, [3]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia, [4]Department of Computer Science, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India, [5]Department of Electrical Engineering, College of Engineering, King Faisal University, Al-Ahsa, Saudi Arabia, [6]Applied College in Abqaiq, King Faisal University, Al-Ahsa, Saudi Arabia

Introduction: Affecting millions of individuals worldwide, epilepsy is a neurological condition marked by repeated convulsions. Monitoring brain activity and identifying seizures depends much on electroencephalography (EEG). An essential step that may help clinicians identify and treat epileptic seizures is the differentiation between epileptic and non-epileptic signals by use of epileptic seizure detection categorization.

Methods: In this work, we investigated Machine learning algorithms including Random Forest, Gradient Boosting, and K-Nearest Neighbors, alongside advanced DL architectures such as Long Short-Term Memory networks and Long-term Recurrent Convolutional Networks for detecting epileptic seizures in terms of difficulties and procedures evolved depending on EEG data. The EEG data classification by applying ML and DL framework to improve the accuracy of seizure detection. The EEG dataset consisted of 102 patients (55 seizure and 47 non-seizure cases), and the data underwent comprehensive preprocessing, including noise removal, frequency band extraction, and data balancing using SMOTE to address class imbalance. Key features, including delta, theta, alpha, beta, and gamma bands, as well as spectral entropy, were extracted to aid in the classification process.

Results: A comparative analysis was conducted, resulting in high classification accuracy, with the Random Forest model achieving the best results at 99.9% accuracy.

Discussion: The study demonstrates the potential of EEG data for reliable seizure detection while emphasizing the need for further development of more practical and non-invasive monitoring systems for real-world applications.

KEYWORDS

electroencephalography, EEG data classification, seizure detection, epilepsy, SMOTE

1 Introduction

Epilepsy is a neurological condition that affects neurons in the brain. In many instances, epilepsy may not be curable, but it can be managed and controlled with proper care. This involves taking essential steps to ensure patients' safety, especially in situations in which they might be driving, cooking, or simply being at home. With effective monitoring, patients can feel more confident in their daily activities, knowing that help is available when needed. This

can minimize potential harm and reduce their dependence on others. This highlights the significance of proper management in epilepsy.

Epilepsy, a neurological condition, is recognized as a widespread issue that poses a significant risk to human life. Global statistics from the World Health Organization (WHO) indicate that around 50 million people worldwide are affected by epilepsy, establishing it as one of the most prevalent neurological diseases globally. Epilepsy affects individuals of all genders, including males and females, and it is also observed in children (1). Epilepsy refers to a neurological condition in which there are irregular disruptions in the usual functioning of the brain. These disruptions lead to seizures, which can differ in duration and effect from one individual to another. Seizures may be brief and go unnoticed or affect specific body parts or the entire body, occasionally resulting in unconsciousness.

Epilepsy can arise from acquired neurological insults (2) (e.g., oxygen deprivation, head trauma, and strokes) that damage brain tissue and disrupt normal electrical functioning. Genetic mutations affecting ion channels, neurotransmitters, and neural transmission can also predispose individuals to chronic seizures. Elucidating these precipitating factors enables better prevention and treatment of epilepsy. EEG is a non-invasive diagnostic tool that captures the electrical activity generated by brain neurons. Given the multi-channel signals from scalp electrodes and the necessity for long-term recordings, advanced signal processing methods have become indispensable for EEG-based detection (3).

A critical component of managing epilepsy is seizure detection, which involves categorizing EEG signals into seizure or non-seizure classes. This process is facilitated by identifying prominent features within the EEG signals. An important step in reducing the human and monetary costs of uncontrolled epilepsy is the development of methods for more precise seizure detection (4). According to Van de Vel et al. (5), beyond the pursuit of epilepsy treatment options, there is an increasing recognition of the need for effective epilepsy management strategies to enhance patient and caregiver quality of life. Non-EEG-based seizure detection technologies are receiving growing research attention due to their potential to improve care quality, peace of mind, and independence. A comprehensive literature review was carried out, and discussions were held with manufacturers of commercially available devices to gain further insights. The reported performance of non-EEG-based seizure detection devices showed a wide range of sensitivity, from as low as 2.2%–100%. In terms of false detections per hour, the range was 0–3.23 when compared with the gold standard of video-EEG. This underscores the varying reliability of these devices and the need for further research and development in this field.

EEG signals are prone to human error and are impractical for continuous monitoring. While automated systems leveraging machine learning and deep learning have shown promise, significant challenges hinder their widespread adoption in the real world.

Data Limitations EEG datasets often suffer from class imbalance, with far fewer seizure events than non-seizure data, leading models to overlook critical seizure patterns. Signal Complexity: EEG signals are inherently noisy, contaminated by artifacts from muscle movements, eye blinks, or environmental interference, complicating feature extraction. Computational Trade-offs: Deep Learning (DL) models (e.g., CNNs, LSTMs, transfer learning in DL, GRU, and transformers) excel at automatic feature learning but require substantial computational resources, making them unsuitable for low-power wearable devices (5). Conversely, traditional ML models, while efficient, rely on manual feature engineering, which risks missing subtle seizure signatures. Generalizability: Many algorithms perform well on controlled datasets but falter with patient-specific variability or ambulatory recordings.

This study aims to explore the potential of EEG data classification using machine learning techniques to enhance seizure detection. We conducted extensive preprocessing of the EEG data, including noise filtering, frequency band extraction, and data balancing, to ensure robust feature extraction and to improve model performance. By evaluating the effectiveness of different machine learning models, this work contributes to the growing body of research aimed at developing more accurate and efficient tools for epilepsy management. Furthermore, we emphasize the need for non-invasive, user-friendly monitoring systems that can complement EEG-based detection in real-world clinical applications. The main contributions of the article include a robust preprocessing pipeline combining noise filtering, frequency band extraction, and SMOTE-based class balancing, coupled with a comparative analysis of five models: Random Forest (RF), Gradient Boosting, KNN, LSTM, and LRCN. The RF classifier achieves state-of-the-art accuracy (99.9%). The paper is structured as follows: Section 2 reviews existing methodologies, Section 3 details the proposed framework, Section 4 presents empirical results and comparisons, and Section 5 concludes with clinical implications and future directions.

2 Literature review

Over 50 million individuals throughout the world are afflicted with epilepsy, a neurological disorder. Seizures that cannot be controlled occur repeatedly. To improve medical results and quality of life for epileptic patients, it is essential to monitor and diagnose seizures in a timely manner. Seizures may be quickly and accurately diagnosed using EEG data, which records the brain's electrical activity. On the other hand, patients may find it obtrusive and complicated gear is usually required.

Recent years have seen tremendous growth in the area of epileptic seizure identification using EEG data, merit to the use of several ML and DL approaches. This literature review examines 23 studies that have contributed to this domain, categorizing them based on their methodological approaches, datasets used, and the specific aspects of seizure detection they address. The studies are grouped into four main categories: Traditional ML Approaches, DL Methods, Hybrid and Novel Approaches, and Comparative Studies and Reviews.

2.1 Traditional machine learning approach

Several studies have employed traditional ML techniques for seizure detection and classification, often focusing on feature extraction and selection methods. Fergus et al. (6) proposed a supervised ML method using the real dataset, achieving a sensitivity and specificity of 88%. This study demonstrated the potential of traditional ML methods in creating generalizable seizure detection models. Raghu et al. (7) presented a model that is computationally efficient by using a new feature known as a successive decomposition index. The system was evaluated using three different databases. Authors proposed support vector machine (SVM) classifiers, they achieved high sensitivity (95.80–97.53%) and low false detection rates

(0.4–0.57/h) across all datasets. The use of multiple datasets in this study provided robust validation of their approach, highlighting the importance of diverse data in developing reliable seizure detection methods. Rani et al. (8) developed SVM approach for classifying a peak signal EEG signal. The system was used dataset that collected from Bonn University dataset. The SVM model achieved a remarkable 99.60% accuracy rate and a low error rate of 0.039. Almustafa (9) conducted a comprehensive comparison of various ML. These studies have demonstrated the continued relevance and effectiveness of traditional ML approaches in seizure detection, particularly when combined with innovative feature extraction methods. The high accuracies achieved by these methods suggest that they remain competitive with more complex DL approaches in certain scenarios.

2.2 Deep learning method

Due to automatically learn essential characteristics from raw EEG data, DL approaches have improved seizure detection accuracy and resilience. Liu et al. (10) created a hybrid bilinear DL network using CNNs and RNNs, model was scored 97.4% on the Temple University Hospital Seizure Corpus and 97.2% on EPILEPSIAE, demonstrating the power of neural network architectural composition. This research showed that CNNs, which excel in spatial feature extraction, and RNNs, which capture temporal relationships in EEG data, work well together.

The linear graph convolution network (LGCN) introduced by Zhao et al. (11) uses spatial interactions in EEG data using a Pearson correlation matrix to identify seizures. This novel method showed graph-based neural networks could capture intricate spatial correlations between EEG channels. Gabeff et al. (12) used the REPO2MSE cohort of scalp-EEG recordings from 568 epilepsy patients to construct a CNN-based model for online seizure identification. For clinical applications, online detection is key. This work addressed it. Chou et al. (13) tested four CNN architectures for video-EEG data analysis and found that their best model had 97.7% ictal stage accuracy. This work showed that CNNs can interpret multimodal data for seizure detection, indicating that adding visual information to EEG signals may improve detection. A 3D CNN-based automated epilepsy detection method by Sun and Chen (14) was very accurate. Their method used CNNs' three-dimensionality to collect EEG signals' spatial and temporal properties. This research proved the generalizability of their 3D-CNN-based technique by performing well across numerous datasets. Kunekar et al. (15) employed LSTM networks to identify seizures with 97% validation accuracy on the UCI-Epileptic Seizure Recognition dataset. It is observed that LSTM outperformed traditional algorithms in accuracy and precision. This work showed that RNNs can identify seizures by recording EEG data temporal dynamics. These DL methods demonstrate automated feature learning and complicated, high-dimensional EEG data processing. High accuracies across datasets show DL seizure detection technologies are getting more dependable.

2.3 Hybrid and novel approaches

Several studies have proposed innovative methods that combine different techniques or introduce novel concepts to improve seizure detection, often addressing specific challenges in the field or exploring unconventional approaches.

Bandarabadi et al. (16) presented a statistical methodology for selecting the preictal period, which serves as an indicator of seizure predictability. This study was used EGG recordings from 18 patients, provided insights into optimizing preictal periods for more precise classification models. This study contributed to the important area of seizure prediction, which has implications for early intervention and improved patient care.

Mert and Akan (3) introduced novel EEG analysis methodologies that achieved accuracy rates as high as 97.89%, demonstrating the potential of innovative signal-processing techniques in seizure detection. While the specific details of their approach were not provided in the summary, the high accuracy achieved suggests that there is still room for improvement in EEG signal analysis techniques.

Brari and Belghith (17) developed a machine learning framework leveraging chaos and fractal theories. Their approach, which included reconstructing EEG signals and extracting the Hurst fractal dimensions, achieved 100% accuracy on the Bonn EEG database using a small number of features and a linear classifier. This study highlighted the potential of applying concepts from complex systems theory to EEG analysis, offering a novel perspective on seizure detection.

Shah et al. (18) combined RNNs with a discrete wavelet transform for seizure detection. This hybrid approach demonstrated the benefits of combining wavelet-based feature extraction with the modeling capabilities of random neural networks.

Kantipudi et al. (19) presented an advanced complex Neural Network. This complex approach achieved an overall detection performance of 99.6% with a high F-measure (99%) and G-mean (98.9%). The study showed the potential of combining multiple advanced techniques, including bio-inspired optimization and specialized neural network architectures.

Ein Shoka et al. (20) introduced CNN model to classify EEG data using chaotic maps for addressing the crucial aspect of data privacy in medical applications while maintaining high classification performance. This study addressed the important issue of privacy preservation in medical data analysis, which is becoming increasingly relevant in the era of big data and interconnected healthcare systems.

Zeng et al. (21) applied a method that integrates deep and shallow learning techniques. The combined approach used a deep neural network for feature extraction, followed by PCA for dimensionality reduction and shallow classifiers for final classification, achieving nearly 100% accuracy on the Bonn dataset. This hybrid approach leveraged the strengths of both deep and traditional machine learning methods, demonstrating the potential benefits of such integrations.

These hybrid and novel approaches demonstrate the potential for significant improvements in seizure detection by combining different techniques or introducing innovative concepts. They often address specific challenges in the field, such as privacy preservation, computational efficiency, or the need for more interpretable models.

2.4 Comparative studies and reviews

Several studies have focused on comparing different methods or providing comprehensive reviews of the field, offering valuable insights into the relative performance of various approaches and highlighting areas for future research.

Bhandari et al. (22) introduced a comparative study in which seven raters reviewed EEG sharp. Their results showed that certain

criteria in sensor space and source space analysis could achieve accuracy rates comparable to expert scoring, providing insights into the effectiveness of different EEG analysis methods. Singh and Kaur (23) designed a neural network classifiers and nonlinear EEG features, demonstrating high accuracy and AUC. Their study provided a comparison point for the effectiveness of nonlinear feature extraction in seizure detection and highlighted the importance of feature engineering in machine learning approaches.

Polat and Nour (24) proposed a hybrid method for seizure detection and classification and compared different SVM kernels and normalization techniques. Their study, which achieved accuracies of 76.70%–82.50%, showed the effects of preprocessing and classifier selection on detection performance. This study underscored the importance of careful parameter tuning and preprocessing in achieving optimal performance with traditional machine learning methods.

Farooq et al. (25) conducted a systematic literature review of ML techniques for seizure detection. Their review identified common feature extraction methods and classifiers, created a taxonomy of state-of-the-art solutions, and highlighted research gaps and challenges. This comprehensive review provided a valuable overview of the field, insights into trends, and directions for future research.

Hamlin et al. (26) explored the use of non-cerebral sensor data for seizure detection and compared the effectiveness of different sensor types and features. Their study, which achieved a mean ROC value of 0.9682, suggested the potential of multimodal approaches in improving seizure detection accuracy. This study opened up new possibilities for seizure detection by incorporating data from sensors beyond traditional EEG, potentially leading to more robust and versatile detection systems.

These comparative studies and reviews provide valuable insights into the relative performance of different methods and highlight areas for future research. They offer a broader perspective on the field and help researchers and practitioners understand the strengths and limitations of various approaches.

2.5 EEG datasets review

Epilepsy research and seizure detection have greatly benefited from the availability of diverse and comprehensive EEG datasets. This section provides review all type of datasets utilized in recent studies on epilepsy classification and seizure detection. These datasets vary in size, patient population, and recording methods.

2.5.1 CHB-MIT dataset

The CHB-MIT dataset has been widely used in several studies for seizure detection and classification. Fergus et al. (6) employed this dataset in their supervised machine learning approach, achieving 88% sensitivity and specificity. Raghu et al. (7) utilized SVM classifiers on this dataset, resulting in 97.28% sensitivity and a false detection rate of 0.57/h. Zhao et al. (11) implemented a Linear Graph Convolution Network (LGCN) on the CHB-MIT data, achieving impressive results with 99.30% accuracy, 98.82% specificity, and 99.43% sensitivity. Shah et al. (18) combined Random Neural Networks (RNN) with Discrete Wavelet Transform (DWT) on this dataset, achieving 93.27% accuracy. Sun and Chen (14) also used this dataset in their 3D-CNN approach, reporting high accuracy, although the specific value was not provided in the summary.

2.5.2 Bonn University dataset

The Bonn University dataset has been the foundation for several innovative approaches in seizure detection. Rani and Chellam (8) achieved a remarkable 99.60% accuracy using their Peak Signal Features (PSF) method combined with an SVM classifier on this dataset. Brari and Belghith (17) applied concepts from chaos and fractal theories to the Bonn dataset, achieving 100% accuracy. (18), in addition to their work on the CHB-MIT dataset, also used the Bonn dataset, achieving an even higher accuracy of 99.84% with their RNN and DWT combination. Zeng et al. (21) employed a hybrid approach combining deep and shallow learning techniques on this dataset, reporting nearly 100% accuracy.

2.5.3 Temple University Hospital (TUH) dataset

The TUH dataset has been utilized in studies employing various ML and DL techniques. Liu et al. (10) achieved a 97.4% F1-score on this dataset using their hybrid bilinear DL network. Raghu et al. (7), as part of their multi-dataset study, applied SVM classifiers to the TUH data, achieving 95.80% sensitivity and a false detection rate of 0.49/h. Sun and Chen (14) included the TUH dataset in their 3D-CNN study, reporting high accuracy, although the specific value for this dataset was not provided in the summary.

2.5.4 EPILEPSIAE dataset

The EPILEPSIAE dataset was used by Liu et al. (10) in their comprehensive study employing a hybrid bilinear deep learning network. On this dataset, their approach achieved a 97.2% F1-score, demonstrating the effectiveness of their method across different datasets.

2.5.5 UCI-epileptic seizure recognition dataset

Kunekar et al. (15) utilized the UCI-Epileptic Seizure Recognition dataset in their study focusing on LSTM networks for seizure detection. Their approach achieved a validation accuracy of 97% on this dataset, highlighting the potential of recurrent neural networks in capturing the temporal dynamics of EEG signals for seizure detection.

2.5.6 REPO2MSE dataset

Gabeff et al. (12) used the REPO2MSE dataset, which consists of scalp-EEG recordings from 568 epilepsy patients, to develop their CNN-based model for online epileptic seizure detection. Table 1 given highlight the importance of standardized, publicly available datasets in advancing seizure detection research.

2.6 Conclusion of the EEG section review

The reviewed studies demonstrate significant progress in seizure classification and detection based on EEG signals. Traditional machine learning approaches continue to show effectiveness, particularly when combined with innovative feature extraction methods. The studies of Fergus et al. (6), Raghu et al. (7), and Rani and Chellam (8) show the potential of these methods when applied with careful feature engineering and selection.

TABLE 1 Summary of EEG datasets.

Studies	Dataset	Description
Fergus et al. (6), Raghu et al. (7), Zhao et al. (11), Sun and Chen (14), and Shah et al. (18)	CHB-MIT	Scalp EEG data from 23 pediatric subjects with intractable seizures, recorded at the Children's Hospital Boston. Contains 686 h of EEG recordings.
Rani et al. (8), Brari and Belghith (17), Shah et al. (18), and Zeng et al. (21)	Bonn University	Consists of 5 subsets (Z, O, N, F, S) each containing 100 single-channel EEG segments of 23.6-s duration. Sets Z and O are from healthy subjects, N and F from seizure-free intervals, and S contains seizure activity.
Raghu et al. (7), Liu et al. (10), and Sun and Chen (14)	Temple University Hospital (TUH)	Large-scale dataset of clinical EEG recordings from Temple University Hospital. Contains over 30,000 EEG records from more than 16,000 patients.
Liu et al. (10)	EPILEPSIAE	European database of long-term EEG data from epilepsy patients. Contains both scalp and intracranial EEG recordings.
Kunekar et al. (15)	UCI-Epileptic Seizure Recognition	Dataset from UCI Machine Learning Repository, containing 11,500 EEG recordings, each 1 s long, classified into 5 categories.
Gabeff et al. (12)	REPO2MSE	Cohort of scalp-EEG recordings from 568 epilepsy patients. Specific details not provided in the summary.

Deep learning techniques, especially CNNs and LSTMs, have demonstrated remarkable performance in automatically learning relevant features from raw EEG data. Liu et al. (10), Zhao et al. (11), and Sun and Chen (14) revealed the power of these approaches in capturing complex spatial and temporal patterns in EEG signals. The high accuracies achieved by these methods across various datasets suggest that they are becoming increasingly reliable for seizure detection tasks.

Hybrid and novel approaches, such as those leveraging Brari and Belghith's chaos theory (17), fractal dimensions, and Zhao et al. (11) graph neural networks have shown promise in improving detection accuracy and addressing specific challenges in the field. These innovative methods often combine the strengths of different approaches or introduce new concepts from other domains, pushing the boundaries of what is possible in seizure detection.

The integration of multiple data sources and sensor types, as seen in Hamlin et al.'s study (26), suggests promising directions for more robust seizure detection systems. This multimodal approach could lead to detection systems that are less prone to false positives and more adaptable to different patient populations.

Comparative studies and reviews, such as those by Kural et al. (22) and Farooq et al. (25), provide valuable insights into the relative performance of different methods and highlight areas for future research. These studies help contextualize individual research efforts within the broader landscape of seizure detection techniques.

However, challenges remain in terms of generalizability across different datasets and patient populations, as well as in reducing false-positive rates and detection delays. The need for larger, more diverse datasets and standardized evaluation metrics is evident from the literature. Many studies use different datasets and evaluation metrics, making direct comparisons challenging. Table 2 reviews studies on EEG-based seizure detection by summarizing the methodologies, technologies, and results of various research efforts and focusing on the effectiveness and accuracy of EEG applications in detecting seizures.

Figure 1 illustrates a summary of the EEG classification results. It provides a visual representation of how different EEG signals have been classified and shows the accuracy and performance of the classification model. It presents the various metrics and comparisons, helping to understand the effectiveness of the approach used to distinguish between different brain wave patterns.

3 Methodology

The proposed system is being investigated using a real EGG dataset. Various algorithms were employed to enhance the existing methods for modeling and detecting seizure diseases. This research presents a detailed overview of the training and validation methodologies employed for the RF, GB, LSTM, and LRCN models. The outlined method structures the approach employed to identify seizures through EEG data, as illustrated in Figure 2.

3.1 EEG dataset acquisition

EEG data were collected from a group of patients who had continuous video-EEG monitoring for an extended duration at two medical institutions in Denmark: Aarhus University Hospital and the Danish Epilepsy Center in Dianalund (22). The data collection period was from January 2012 to September 2017. During the diagnostic evaluation phase, sharp transients were initially identified and marked. Subsequently, two authors conducted a comprehensive review of these marked transients. Through collaborative analysis, a consensus was established among the experts, confirming the initial marking as a sharp transient, regardless of its manifestation of epileptiform characteristics. This selected sharp transient was then subjected to further evaluation to ensure compliance with the predetermined selection criterion. In the dataset, there were 100 files in the European Data Format (EDF), comprising data from 55 epileptic patients and 47 non-epileptic patients of different ages and genders. On December 18, 2017, the dataset that was used for this research was recorded. A sample rate of 500 Hz was used to get the EEG data, since this is the industry standard for collecting the important frequency content in EEG signals. The raw data was further processed using a 250 Hz low-pass filter. The EEG recording system employed in this study comprised 26 channels, enabling the simultaneous measurement of brain activity from multiple scalp locations. Table 3 outlines the EEG dataset content and features, such as the number of patients and class.

TABLE 2 A review of studies of EEG-based seizure detection.

Study	Data	Preprocessing	Models/Algorithms	Results
Liu et al. (10)	Temple University, EPILEPSIAE dataset	exploit the frequency (STFT), analysis data	Hybrid bilinear deep learning network (CNNs + RNNs)	F1-score: 97.4%
Fergus et al. (6)	CHB-MIT dataset	Simple, filter, features extraction	k-NN, SVM, NN, DT	Sensitivity 88%, AUC: 93%
Mert and Akan (3)	Various EEG recordings	Digitalize, filter, Normalize frequency	Novel EEG analysis methods	Accuracy: 97.89%
Raghu et al. (7)	Ramaiah Medical College, CHB-MIT	Feature extraction (SDI)	SVM	Sensitivity: 97.53%
Bhandari et al. (22)	1,001 patients (video-EEG) EMG Data	Record, sample and filter the data	Analysis of EEG sharp transients	92% Accuracy
Zhao et al. (11)	CHB-MIT dataset	Pearson correlation matrix	Linear Graph Convolution Network (LGCN)	Accuracy: 99.30%, Sensitivity: 99.43%
Rani et al. (8)	Bonn University dataset	Peak Signal Features (PSF)	SVM, DT, KNN	Accuracy up to 99.60% with SVM
Aayesha et al. (28)	Bonn and CHB-MIT datasets	Feature extraction	KNN, FRNN	Accuracy: up to 99.81%
Gabeff et al. (12)	REPO2MSE cohort	Simple, segment and split the data	CNN	F1-score: 0.873, 90% seizure detection
Brari and Belghith(17)	Bonn EEG database	EEG signal reconstruction	Chaos and fractal theories	Accuracy: 100%
Chou et al. (13)	Video-EEG data	Not specified	Four CNN architectures	97.7% accuracy for ictal stage
Shah et al. (18)	CHB-MIT, BONN datasets	DWT	RNN, ANN, SVM	CHB-MIT: 93.27%, BONN: 99.84%
Polat and Nour (24)	Not specified	Z-score, Minimum-Maximum, MAD normalizations	SVM (Linear, Cubic, Medium Gaussian)	76.70–82.50%
Kantipudi et al. (19)	Not specified	FLHF	GBSO, TAENN	99.6%, F-measure: 99%, G-mean: 98.9%
Almustafa (9)	Not specified	Not specified	Random Forest, K-NN, Naïve Bayes, Logistic Regression, DT, Random Tree, J48, SGD	97% accuracy,
Kunekar et al. (15)	UCI-Epileptic Seizure Recognition dataset	Not specified	LSTM, Logistic Regression, SVM, KNN, ANN	97% Accuracy
Hamlin et al. (26)	Data from 15 patients	LDA	Not specified	Mean ROC: %96.8
Zeng et al. (21)	Bonn dataset	PCA	CNN, shallow classifiers	~100% Accuracy
George et al. (29)	KITS, TUH databases	TQWT, entropies	PSO, ANN	KITS: 100%, TUH: 88.8–97.4% Accuracy

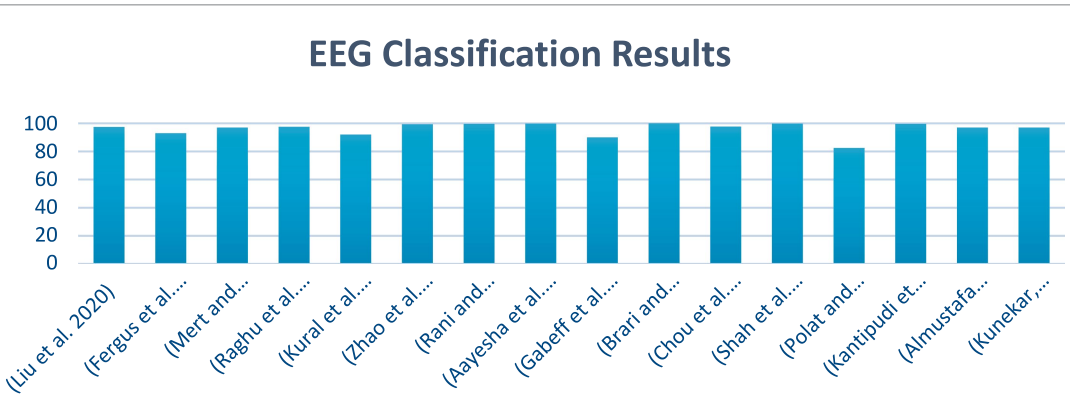


FIGURE 1
EEG classification result.

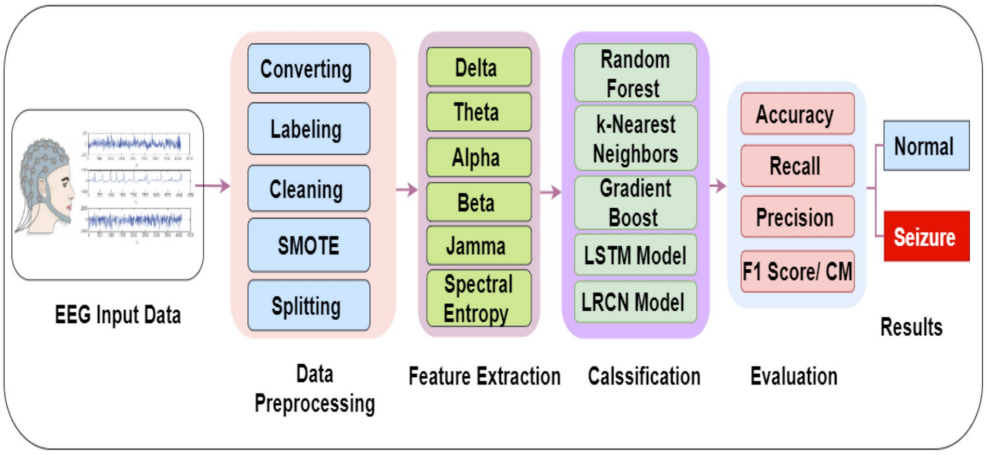


FIGURE 2
Proposed methodology for EEG data classification and seizure detection.

3.2 Preprocessing

In the data preprocessing phase, the raw EEG data undergo filtering to extract the relevant frequency bands of interest. Specifically, the following frequency bands are extracted: alpha (8–12 Hz), beta (13–30 Hz), theta (4–7 Hz), and gamma (above 30 Hz). These frequency bands are commonly analyzed in EEG studies because of their associations with various cognitive and physiological processes. It is crucial to preprocess the EEG data appropriately to ensure the reliability and validity of subsequent analyses (27). The filtering step is essential for isolating the frequency bands of interest and minimizing the influence of irrelevant signal components or noise. The extraction of these specific frequency bands facilitates the investigation of their potential correlations with the cognitive or physiological processes under study, as shown in Figure 3.

3.2.1 Data labeling

In this process, we labelled all of the EEG recordings in the dataset according to the patient's status. We used the numbers "1" to denote normal EEG data and the number "0" to denote seizures. While training, the classification algorithm benefits from this labeling as it allows it to differentiate between the two groups.

3.2.2 Data normalization

The EEG characteristics were on the same scale, we normalized the data. If you want to make sure that the learning process is not overloaded with features with out-of-range values, normalization is a must. Z-score normalization method was used for scaling the rows of EGG dataset.

3.2.3 Data cleaning

Initial data cleaning was performed to address any missing values within the features. The mean imputation technique was utilized, where missing values in any given feature were replaced with the mean value of that feature. This method was implemented using the SimpleImputer class from the sklearn.impute module, configured with strategy = 'mean'. The transformation was applied to all feature columns, excluding the 'label' column, which represents the target variable.

TABLE 3 EEG dataset content.

Class	Number of patients
Normal	55
Seizure	47

3.2.4 Data balancing using SMOTE

SMOTE technique used to address class imbalances in datasets. One step in processing SMOTE data is to use synthetic samples for the minority class. This ensures that the distribution of classes is balanced. The algorithm works by identifying the KNN for each minority class sample and creating new synthetic samples along the line segments that join the minority class sample and its neighbors. The synthetic samples are generated by randomly selecting one of the KNN and introducing a perturbation along the line segment joining the two samples. This approach was implemented using the SMOTE class from the imblearn. over_sampling library with a random_state set for the reproducibility of results. The resampling process adjusted the dataset to ensure an equal representation of both classes, mitigating the potential effect of class imbalance on the subsequent analysis and modeling steps. Figure 4 illustrates the distribution of EEG data before and after applying SMOTE.

3.2.5 Data splitting

Two subsets, training and testing, were taken from the dataset. A data allocation of 80% for training and 20% for testing the machine learning model is known as an 80/20 split. By splitting the data in this way, we can train the model on one set of data and then evaluate it on another set, which stops overfitting and lets the model generalize.

3.2.6 Heatmap of amplitude differences

The profound complexities underlying epileptic seizures necessitate a multifaceted approach to elucidate their intricate mechanisms. The study presents a comprehensive spatiotemporal analysis of EEG data, leveraging the visual potency of heat maps to

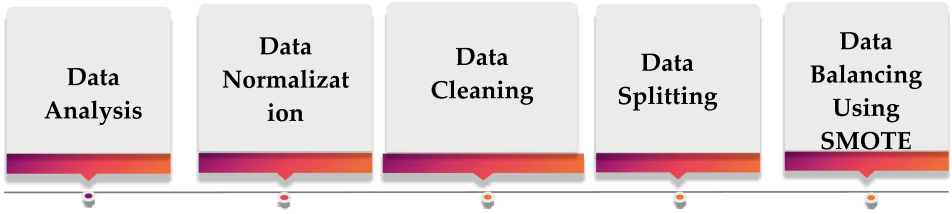


FIGURE 3
EEG data preprocessing steps.

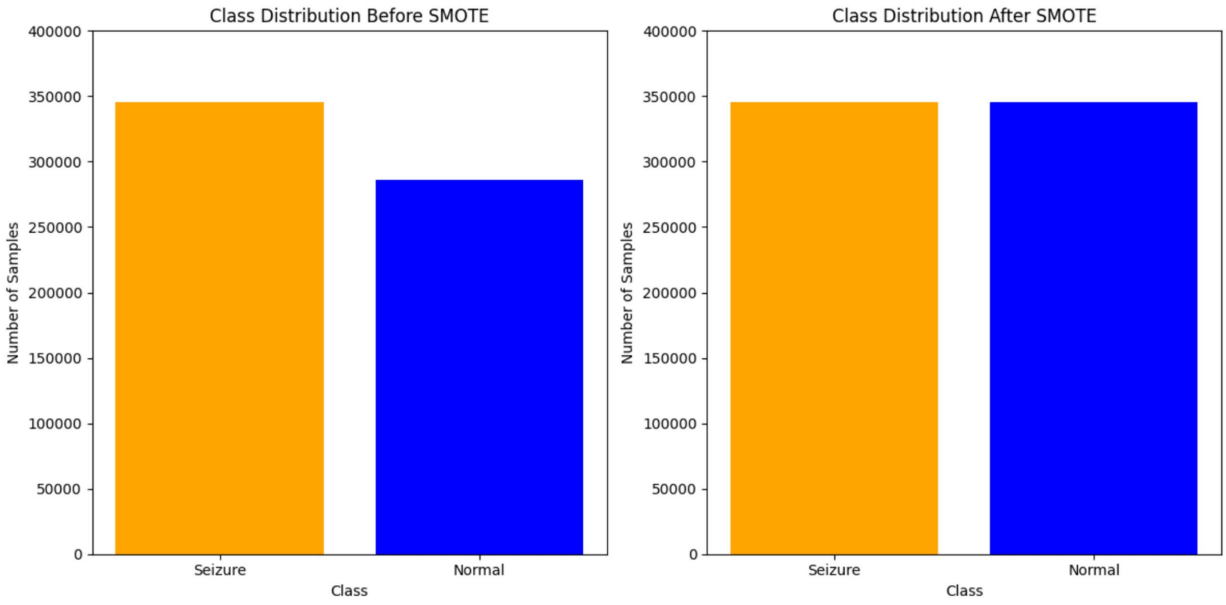


FIGURE 4
EEG data distribution before and after applying SMOTE.

delineate amplitude variations across cortical regions. By comparing seizure and non-seizure conditions, the proposed methodology quantifies the dynamic shifts in neural activity, transitioning seamlessly from negative to positive amplitude deviations through a "coolwarm" color palette. This graphical representation not only facilitates the localization of epileptogenic foci but also elucidates the propagation patterns of seizure activity, thereby contributing to a holistic understanding of the pathophysiological processes underlying this neurological disorder. As shown in Figure 5, the knowledge acquired from this study has great consequences for the formulation of focused treatment strategies and the progress of our understanding of the complex neural dynamics controlling seizure events.

3.2.7 Spectral analysis

This study used Fourier spectral analysis of EEG data to elucidate the frequency domain signatures that differentiate seizure and non-seizure neural dynamics in epilepsy. The spectral power distributions derived from these analyses revealed pronounced amplitudes across specific frequency bands during seizure activity, which is indicative of heightened neuronal synchronization. By contrast, the non-seizure condition exhibited reduced spectral power, reflecting normal neural oscillations. By characterizing these distinct frequency profiles, this work sheds light on the neurophysiological underpinnings of epileptic seizures and pathological hypersynchrony and paves the way for improved therapeutic interventions, as shown in Figure 6.

3.3 Feature extraction

This study analyzed the power spectral density (PSD) levels across different frequency bands to investigate the differences in neural activity between epileptic and non-epileptic patients. The epileptic patient exhibited distinct PSD levels compared with the non-epileptic patient, suggesting variations in their underlying neural activity patterns. The frequencies at which the difference in PSD between the two patients was statistically significant ($p < 0.05$) were identified, indicating that the observed differences in brain activity were unlikely due to chance. Significant differences at certain frequencies, such as increased power in the theta and gamma bands, could reveal specific brain activity patterns associated with epilepsy, including the presence of epileptic networks outside of seizure events. These findings contribute to a better

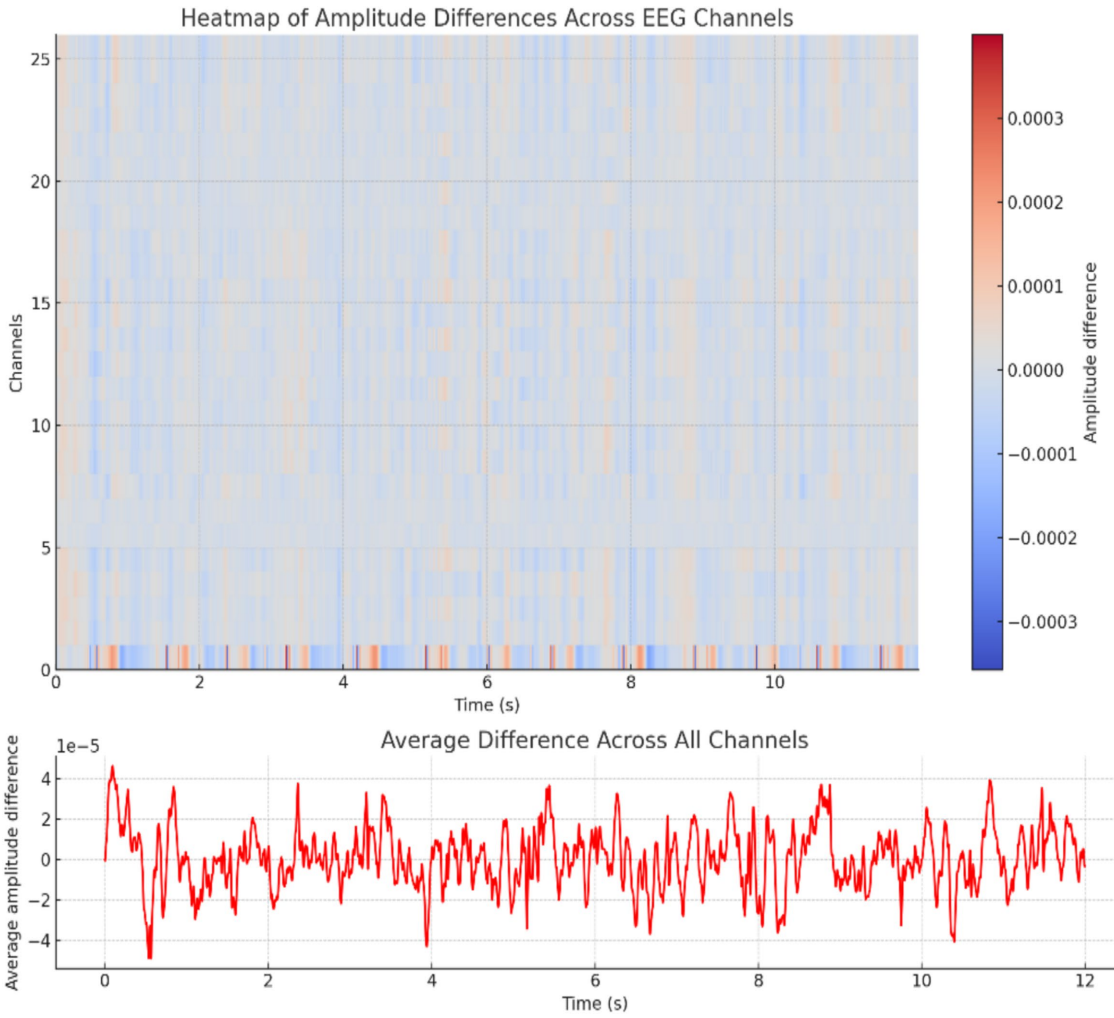


FIGURE 5
Heat map of amplitude differences.

understanding of the neurophysiological underpinnings of epilepsy and hold promise for improving diagnostic and monitoring techniques and for guiding more targeted interventions for the management of epilepsy, as shown in Figure 7.

In this section, several features were extracted from the EEG signals to enable the classification of epileptic and non-epileptic patients. These features capture different aspects of neural activity and provide valuable information for distinguishing between the two groups. The extracted features are as follows:

- *Delta*

Usually covering 0.5 to 4 Hz, this function shows the PSD in the delta frequency region. Deep sleep phases are linked to delta waves, which are also well-known to be involved in many cognitive functions like memory and attention.

- *Theta*

The theta feature corresponds to the PSD in the theta frequency band, which ranges from 4 to 8 Hz. Theta oscillations are linked to cognitive processes such as memory formation, spatial navigation, and emotional regulation.

- *Alpha*

The alpha feature is derived from the PSD in the alpha frequency band, typically between 8 and 12 Hz. Alpha waves are prominent during relaxed wakefulness and are believed to play a role in attention and information processing.

- *Beta*

This feature represents the PSD in the beta frequency band, ranging from 13 to 30 Hz.

- *Gamma*

The gamma feature corresponds to the PSD in the gamma frequency band, which encompasses frequencies above 30 Hz. Gamma oscillations are involved in various cognitive functions, including perception, attention, and memory.

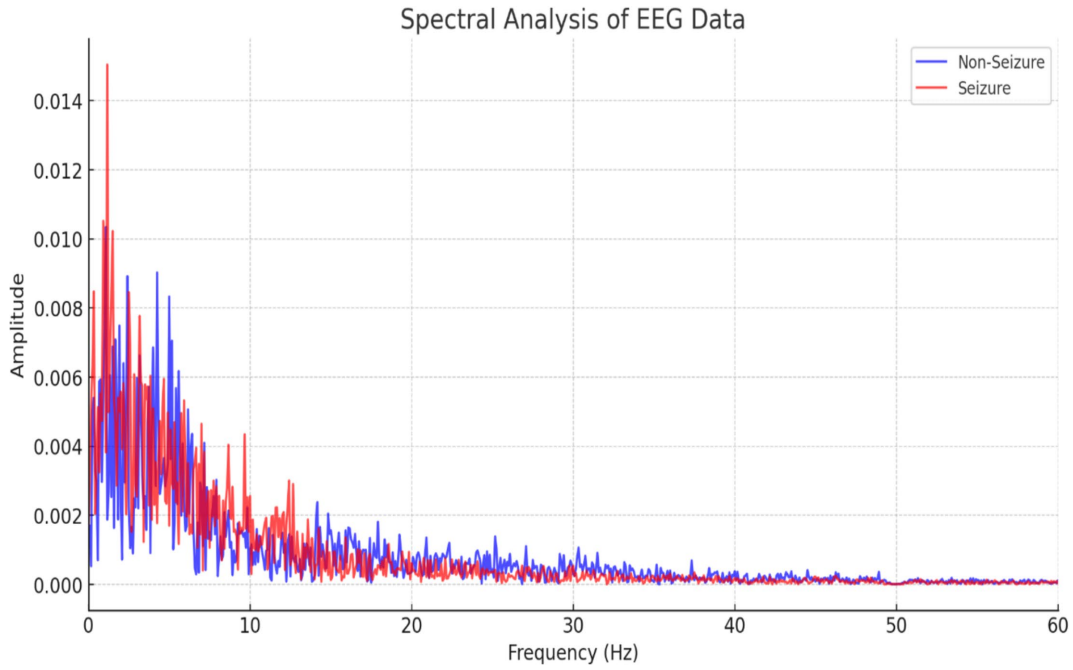


FIGURE 6
Spectral analysis of EEG signals of seizure and non-seizure cases.

- *Spectral entropy*

A estimate of the complexity or irregularity of the EEG signal, spectral entropy It may help to identify aberrant patterns of brain activity by providing details on the distribution of power across many frequency ranges.

The power spectral density (PSD) of gamma band (30 + Hz) emerged as the most discriminative feature, showing statistically significant amplitude increases during seizures ($p < 0.05$, Figure 7). This aligns with neurophysiological evidence linking high-frequency oscillations to epileptic hyperexcitability. The theta band (4–8 Hz) also demonstrated utility, though with marginally lower significance. Other bands (delta, alpha, and beta) contributed minimally, as their PSD distributions overlapped between seizure and non-seizure states.

Spectral entropy, quantifying signal irregularity, effectively captured abrupt changes in EEG complexity during seizures. It achieved a feature importance score of 0.180.18 in the Random Forest (RF) model, complementing gamma band analysis to reduce false positives caused by non-stationary noise.

Commonly utilized in EEG analysis, these characteristics have been shown to be useful in distinguishing and defining many brain states and disorders, including epilepsy. Table 4 summarizes the obtained characteristics; they will be input for categorization techniques.

3.4 Modeling

In the classification stage, the EEG data was analyzed using four models: RF, GB, KNN, LSTM, and LRCN. The RF constructs multiple decision trees and uses majority voting for classification, well-suited for high-dimensional, nonlinear data like EEG signals. Gradient Boosting iteratively combines weak models to capture complex patterns. LSTM, a recurrent neural network variant, can learn long-term dependencies in sequential data such as EEG for identifying seizure patterns. LRCN combines convolutional layers for spatial feature extraction with LSTM for temporal modeling, making it effective for seizure detection and classification from EEG recordings. The specific architectures of these diverse machine learning and deep learning models were previously detailed, Table 5 lists EEG classification models. Justifications for each model in the context of EMG data classification between normal and seizure cases:

3.4.1 Random Forest model

Random Forest Classifier excels in handling complex EMG data due to its ensemble nature. Combining many decision trees, each tuned on random selections of data and attributes, helps to detect complex trends in muscle activity signals. This approach is particularly effective for seizure detection, as it can identify subtle differences in EMG characteristics. The model's feature importance ranking also provides insights into which aspects of the EMG signal are most predictive of seizures, aiding in both classification and physiological understanding.

3.4.2 Gradient boost model

Gradient Boosting is well-suited for EMG classification due to its sequential learning process. Approaches the building of a series of weak learners, generally decision trees, in a stage-by-stage manner, with the main aim of fixing errors generated by previous models. This approach allows it to capture fine-grained differences in EMG patterns between normal and seizure states. Gradient Boosting's ability to handle non-linear relationships and its robustness to outliers make it

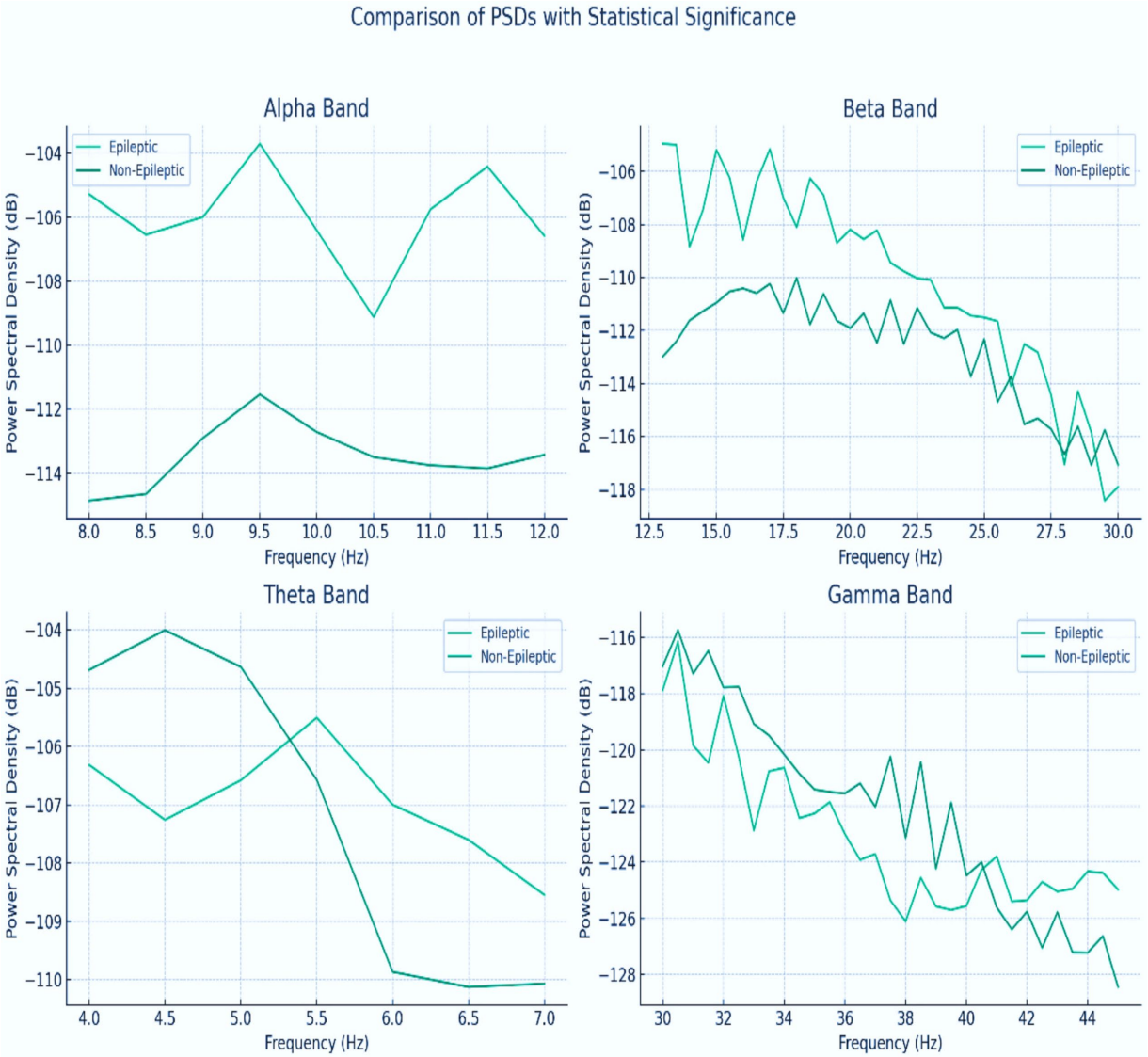


FIGURE 7
PSD for the alpha, beta, theta, and gamma bands between epileptic and non-epileptic patients.

TABLE 4 EEG extracted features summary.

Feature	Description
Delta	PSD in the delta frequency band (0.5–4 Hz)
Theta	PSD in the theta frequency band (4–8 Hz)
Alpha	PSD in the alpha frequency band (8–12 Hz)
Beta	PSD in the beta frequency band (13–30 Hz)
Gamma	PSD in the gamma frequency band (above 30 Hz)
Spectral entropy	Measure of the complexity or irregularity of the EEG signal

TABLE 5 EEG classification models.

No	Model
1	Random Forest Model
2	Gradient Boost Model
3	K-Nearest Neighbors Model
4	LSTM Model
5	LRCN Model

effective in dealing with the variability often present in EMG data during seizures.

3.4.3 K-nearest neighbors model

The K-Nearest Neighbors model is valuable for EMG classification due to its non-parametric nature. It does not assume any specific distribution of the data, making it adaptable to the complex and often non-linear patterns in EMG signals during seizures. By classifying based on the majority class of nearby data points in the feature space, KNN can effectively capture local patterns in muscle activity. This local decision-making is particularly useful for identifying seizure-related EMG characteristics that may vary across patients or types of seizures.

3.4.4 LSTM model

Long Short-Term Memory networks can process sequential data and record long-term dependencies, they are especially appropriate for EMG data interpretation. EMG signals during seizures often exhibit temporal patterns that evolve over time. LSTM's gating mechanism allows it to selectively remember or forget information, making it adept at identifying relevant temporal features in the EMG signal that distinguish seizure activity from normal muscle function. This temporal modeling capability is crucial for detecting the onset and progression of seizures in EMG data.

3.4.5 LRCN model

The LRCN combines the strengths of both CNNs and LSTMs, making it highly effective for EMG-based seizure detection. The CNN component excels at extracting spatial features from the EMG signal, potentially identifying characteristic frequency patterns or signal morphologies associated with seizures. The LSTM layer then processes these features sequentially, capturing the temporal evolution of muscle activity during seizure events. This dual approach allows LRCN to simultaneously analyze both the spatial and temporal aspects of EMG data, potentially leading to more accurate and robust seizure detection.

4 Results and discussion

In this subsection, we explore the performance of EEG classification for seizure detection using four models: GB, RF, K-NN, LSTM, and LRCN. The objective was to assess and compare their effectiveness in identifying seizures from EEG data. The results are detailed in the accompanying tables and figures, which present the potential of these models in advancing neurological diagnostics. Table 5 outlines the EEG classification models.

4.1 Evaluation matrix

The ML and DL model were evaluated by using evaluation matrix. The Equations 1–5 of evaluation metrics can be defined as follows:

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN} \times 10 \quad (1)$$

$$Sensitivity \frac{True\ Positives}{True\ Positives + False\ positives} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$f1 - score = 2 * \frac{precision \times Sensitivity}{precision + Sensitivity} * 100 \quad (4)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

4.2 Environment setup

All experiments were conducted on a laptop with the following specifications: Intel Core i7 processor, 16GB RAM, and an NVIDIA GeForce RTX 3070 GPU with 8GB VRAM. The software environment consisted of Python 3.9 running within Anaconda, with TensorFlow version 10.1.2 employed for deep learning tasks.

4.3 Results of the GB model

This work classified epileptic and non-epileptic patients using Gradient Boosting (GB) model depending on EEG features. With an accuracy of 0.750, a precision of 0.756, a recall of 0.743, an F1 score of 0.749, and a ROC AUC score of 0.835 the model was able to differentiate between the two groups. With 51,964 true negatives, 51,636 true positives, 16,701 false positives, and 17,850 false negatives, the confusion matrix as shown in Figure 8 further exposed the performance of the model. These findings show how well the model detects trends in EEG data; although there is potential for development in lowering misclassifications, especially in terms of false positives and false negatives, overall the model performs really well.

These results demonstrate the potential of the GB model in accurately classifying epileptic and non-epileptic patients while also highlighting areas for further improvement through feature engineering, hyperparameter tuning, or ensemble methods, as shown in Figure 9.

4.4 Results of the RF model

As shown in Figure 10, the RF model was used with EEG traits to divide people into epileptic and non-epileptic groups. With an accuracy of 0.999, a precision of 1.000, a recall of 0.998, an F1 score of 0.991, and an ROC score of 1.000, the RF model showed extraordinary performance.

The confusion matrix revealed 68,631 true negatives, 69,358 true positives, 34 false positives, and 128 false negatives. These exceptional results demonstrate the efficacy of the RF model in accurately classifying epileptic and non-epileptic patients based on the extracted EEG features, although further validation on independent datasets may be necessary to ensure generalizability, as shown in Figure 11.

4.5 Results of the K-NN

Normal from epileptic EEG data were distinguished using a K-NN classifier. Assigning the class of a data point depending on the majority class of its "k" closest neighbors in the feature space, K-NN is a basic, non-parametric classification method. This work selected K-NN with (k = 5), therefore classifying every EEG sample according on the majority vote of its five closest neighbors in the feature space.

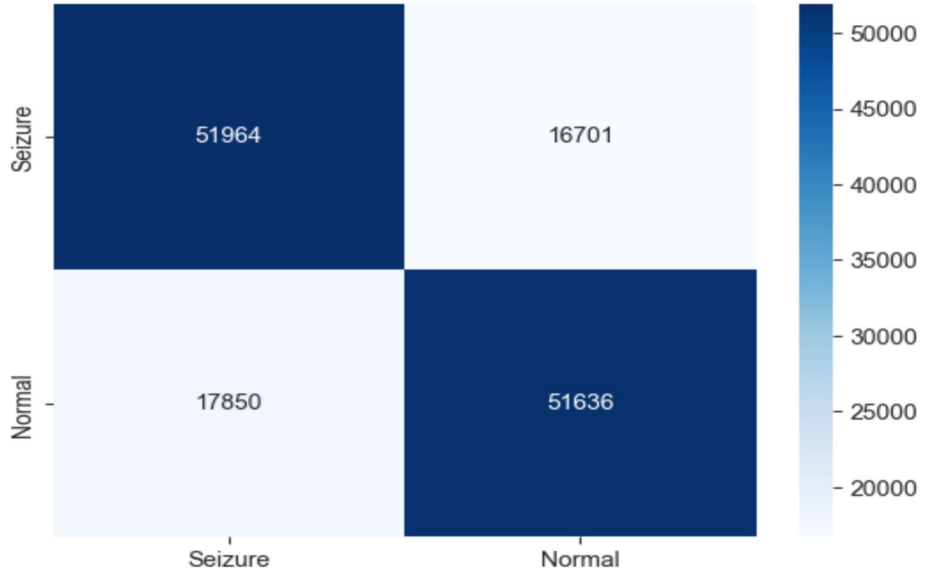


FIGURE 8
Confusion matrix of EEG data using the GB model.

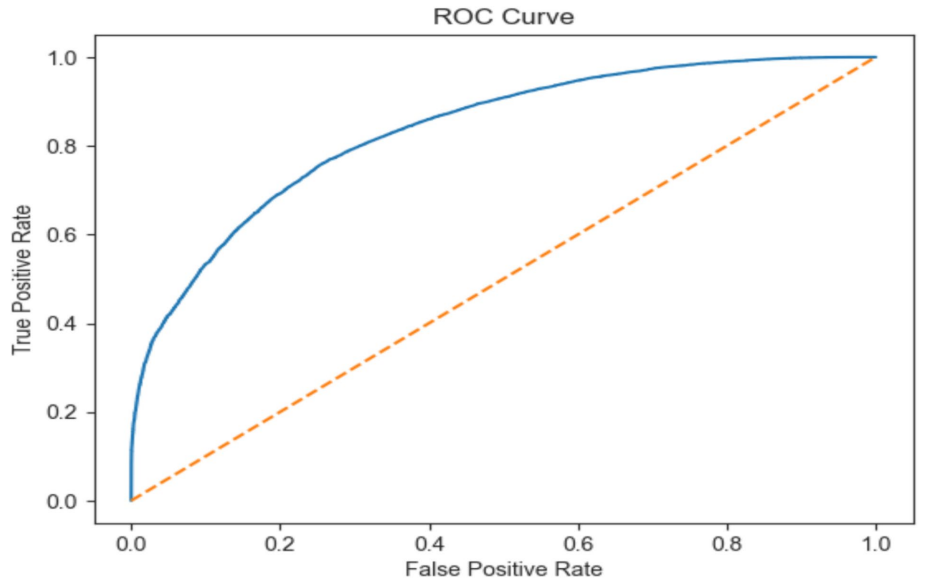


FIGURE 9
ROC AUC score of EEG data using the GB model.

The confusion matrix revealed that, out of the total predictions, 65,924 were true negatives and 67,084 were true positives, indicating that the majority of the normal and seizure cases were correctly identified. However, there were also 2,879 false positives and 2,264 false negatives, as shown in Figure 12.

The KNN model was shown scored with high accuracy (96.3%) indicates that the model correctly classified a substantial majority of the EEG signals. According the precision metric the KNN achieved 95.9% suggests that the model has a low rate of false positives, while recall of 96.7% indicates a low rate of false negatives. The ROC score of 99.02% further validates the model's excellent ability to distinguish between normal and seizure cases, as illustrated in Figure 13.

These results show that although the model is highly accurate, there are still instances of misclassification, which is an area for potential improvement.

4.6 Results of the LSTM model

In this experiment, classified epileptic and non-epileptic patients based on EEG signal characteristics using an LSTM neural network model. The LSTM model turned out with a 0.9906 accuracy. Table 6 gives the LSTM model's parameters. As shown in Figure 14 the confusion matrix indicated 68,190 true

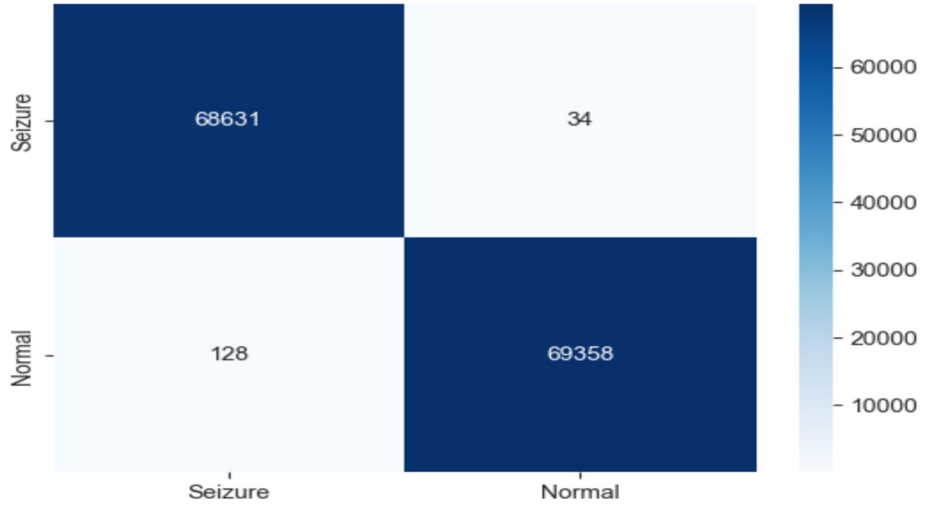


FIGURE 10
Confusion matrix of EEG data using the RF model.

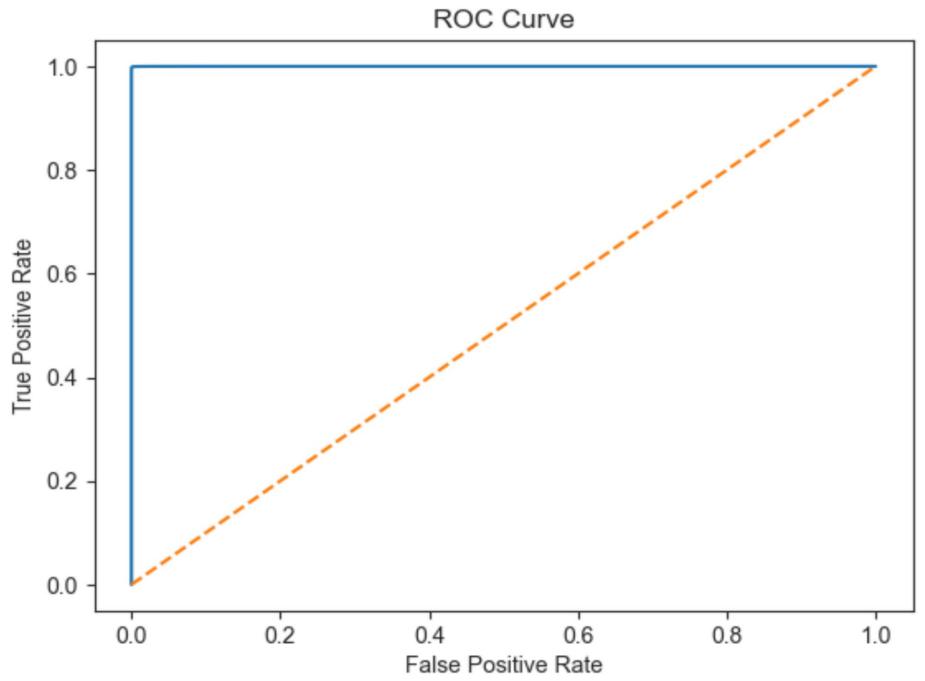


FIGURE 11
ROC AUC score of EEG data using the RF model.

positives, 613 erroneous positives, 68,669 true negatives, and 679 false negatives.

Using collected EEG data, the LSTM model showed remarkable accuracy of 99.06%, a precision of 99.12%, a recall of 99.02% and an F1 score of 99.07% for both epileptic and non-epileptic individuals. These findings demonstrate the great capacity of the model for precisely differentiating between the two classes, therefore stressing its possible uses in EEG-based diagnosis systems. Nevertheless, as Figure 15 shows, the LSTM was optimized and testing across EEG datasets and was shown the improvement in the generalizability of the model and guarantee its resilience in practical conditions.

4.7 Results of the LRCN model

Based on the features of the EEG data, this work categorized people as either epileptic or non-epileptic using an LRCN model. The LRCN model's findings show that the accuracy was 0.9906; the

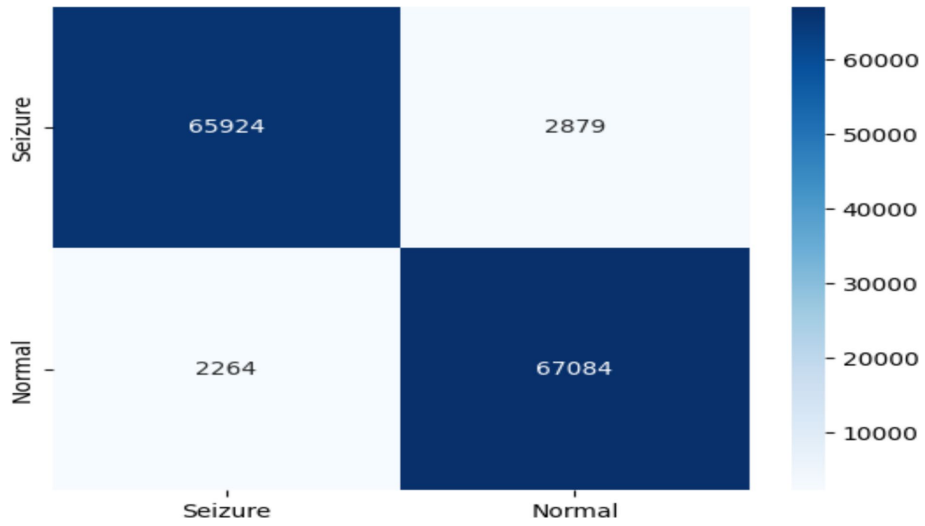


FIGURE 12
Confusion matrix of EEG data using the K-NN model.

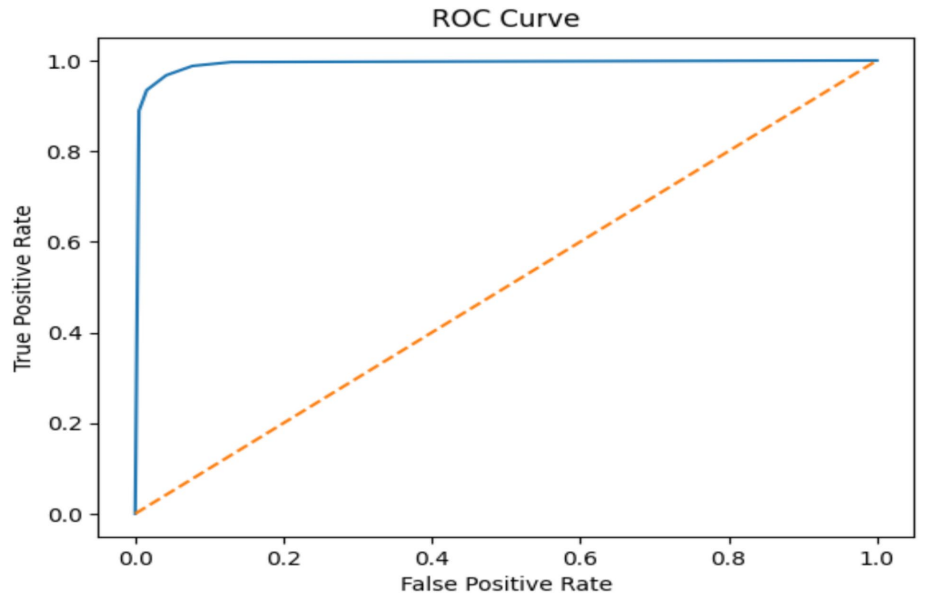


FIGURE 13
ROC AUC score of EEG data using the K-NN model.

precision was 0.9912; the recall was 0.9902; the F1 score was 0.9907. LRCN model characteristics and values (see Table 7).

Strong performance in categorizing seizure and non-seizure episodes from EEG data reveals in the confusion matrix for the LRCN model. The model fairly identifies most instances with 68,678 TP and 69,187 TN. Whereas (FP = 161) reveal minor misclassification of non-seizure events, false negatives (FN = 125) indicate a limited proportion of missed seizures. With low error, the high TP and TN values indicate outstanding sensitivity and accuracy, so the model is very dependable for monitoring epilepsy (see Figure 16).

These results demonstrate the potential of the LRCN model in accurately classifying epileptic and non-epileptic patients based on the extracted EEG features, although further optimization and generalizability testing may be required, as shown in Figure 17.

4.8 Summary of the experimental results of the EEG classification

With almost perfect accuracy, precision, recall, and F1 score, the RF model exceeded the other models based on the testing findings in Section 4.3. Closely matching the RF model, the deep learning models, LSTM and LRCN, also showed outstanding performance with using various evaluation metrics. Though it performed really well, the GB model had somewhat worse measures than the other versions. With

regard to reliably categorizing epileptic and non-epileptic patients based on EEG signal characteristics, the RF, LSTM, and LRCN models shown overall better performance; the RF model ranked highest in this regard in this research. Table 8 and Figure 18 help to show the outcomes.

4.9 EEG monitoring for detecting seizure behavior comparative

With a variety of techniques producing encouraging results, the subject of seizure detection and classification based on EEG data has experienced major developments recently. This review of 23 studies, along with our own research, highlights the diversity of techniques being applied to this critical medical challenge.

TABLE 6 LSTM model parameters using EEG data.

Parameter	Details
LSTM Layer	1,024
LSTM Layer	512, (BatchNormalization ())
LSTM Layer	256
Dense Lyer	34
Dense Lyer	1
Activation Function (Output Layer)	sigmoid
Optimizer	RMSprop
Learning Rate	0.001
Callback	EarlyStopping
Patience for No Improvement (EarlyStopping)	5 epochs
Epoch Training Stopped At	67 epochs
Maximum Epochs	150 epochs
Batch Size	1,024

Traditional machine learning approaches continue to demonstrate their effectiveness, particularly when combined with innovative feature extraction methods. For instance, Rani and Chellam (8) achieved 99.60% accuracy using their Peak Signal Features method with an SVM classifier on the Bonn University dataset. Similarly, Almustafa (9) achieved 97.08% accuracy using a Random Forest classifier. These results underscore the continued relevance of classical machine learning techniques when applied with careful feature engineering.

Deep learning methods have shown remarkable performance in automatically learning relevant features from raw EEG data. Liu et al. (10) achieved a 97.4% F1-score using a hybrid bilinear deep learning network on the Temple University Hospital dataset, while Zhao et al. (11) reached 99.30% accuracy with a Linear Graph Convolution Network on the CHB-MIT dataset. These results demonstrate the power of deep learning in capturing complex patterns in EEG signals without the need for extensive feature engineering.

Hybrid and novel approaches have also yielded impressive results. Brari and Belghith (17) achieved 100% accuracy on the Bonn University dataset using a framework leveraging chaos and fractal theories. Kantipudi et al. (19) reported 99.6% detection performance with their complex model integrating wavelet-based filtering, bio-inspired optimization, and a specialized neural network. These innovative approaches show the potential for pushing the boundaries of seizure detection performance.

Our study, which achieved 99.9% accuracy using a Random Forest Classifier on a standard online dataset, aligns with and even surpasses many of the high-performing methods in the literature. This result underscores the potential of ensemble methods like Random Forest when applied to well-preprocessed EEG data.

The variability in datasets used across studies presents a challenge in directly comparing results. While some datasets like CHB-MIT and Bonn University are frequently used, allowing for some comparison, differences in preprocessing, feature extraction, and evaluation metrics can still make direct comparisons difficult. This highlights the

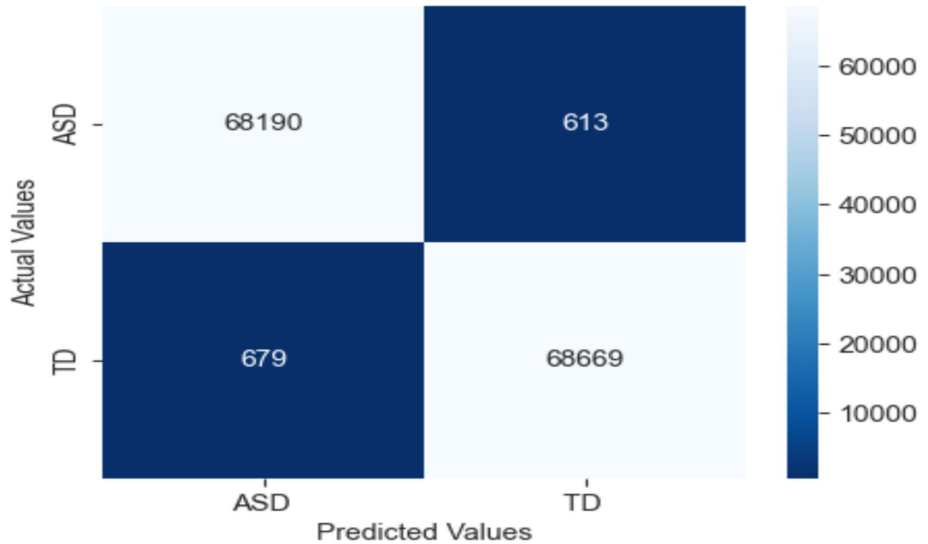


FIGURE 14
Confusion matrix of EEG data using the RF model.

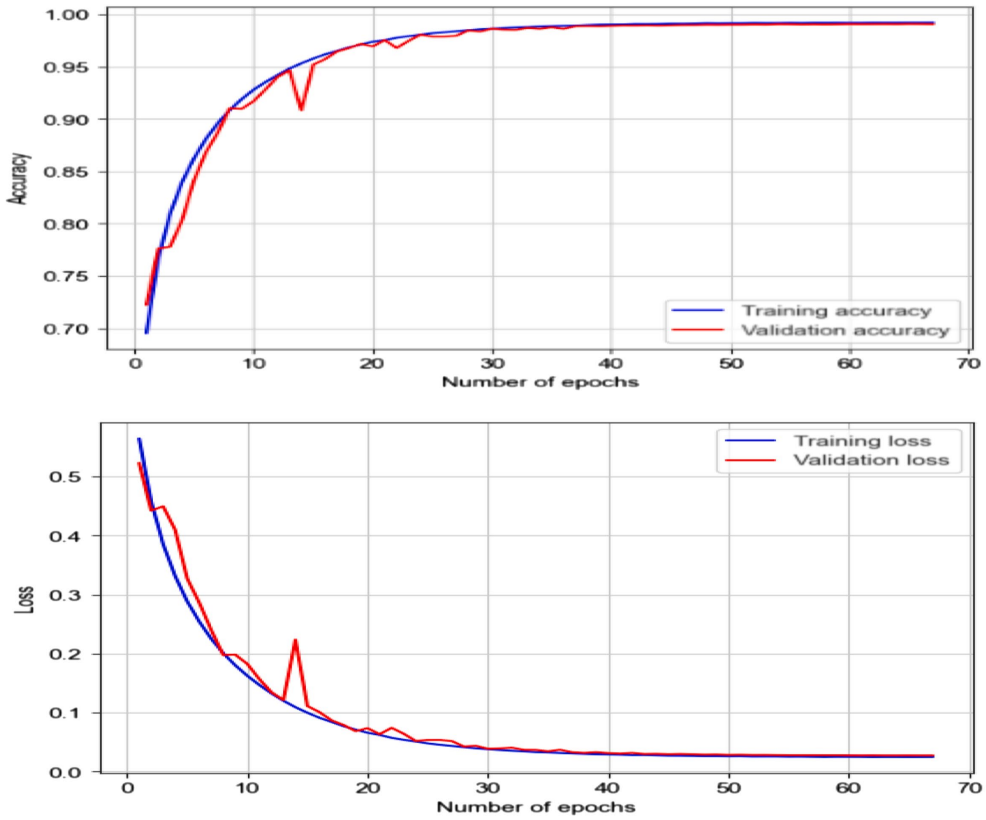


FIGURE 15
Accuracy and loss of EEG data using the LSTM model.

TABLE 7 LRCN model parameters using EEG data.

Parameter	Details
ConvD1	filters = 64, kernel = 3, activation = 'relu'
Custom Layer	Max Pooling
LSTM Lyer	1,024
LSTM Layer	512
LSTM	128
Dense Lyer	1
Activation Function (Output Layer)	sigmoid
Optimizer	RMSprop
Learning Rate	0.001
Callback	EarlyStopping
Patience for No Improvement (EarlyStopping)	5 epochs
Epoch Training Stopped At	69 epochs
Maximum Epochs	150 epochs
Batch Size	128

need for standardized benchmarks and evaluation protocols in the field.

It's noteworthy that while many studies report very high accuracies (>99%), real-world performance may differ due to factors such as inter-patient variability, noise in clinical settings, and the challenge of detecting seizure onset rather than ongoing seizure activity. Future research should focus on validating these high-performing models in diverse clinical settings and on larger patient populations.

The trend towards multimodal approaches, as seen in Hamlin et al. (26), and privacy-preserving methods, as in Ein Shoka et al. (20), points to future directions for the field. Integrating data from multiple sensor types and ensuring patient privacy will be crucial for the widespread adoption of automated seizure detection systems in clinical practice.

While the study demonstrates high accuracy (99.9%) in seizure detection, translating these models to wearable devices faces critical hurdles. Computational efficiency demands significant processing power, conflicting with the resource constraints of wearables. Real-time implementation requires low-latency pipelines, necessitating streamlined preprocessing and hardware-accelerated signal processing. Power consumption, patient-specific variability, and ambulatory noise (e.g., motion artifacts) further complicate reliability. Regulatory compliance, cost barriers, and the need for fail-safe mechanisms to minimize false alarms add layers of complexity. Addressing these challenges hinges on hardware sensor systems to balance accuracy with practicality for clinical adoption.

In conclusion, while significant progress has been made in seizure detection and classification, with our study contributing to the high-performance benchmarks, there remains room for improvement in areas such as real-time detection, generalizability across patients, and interpretability of complex models. Future work should focus on these

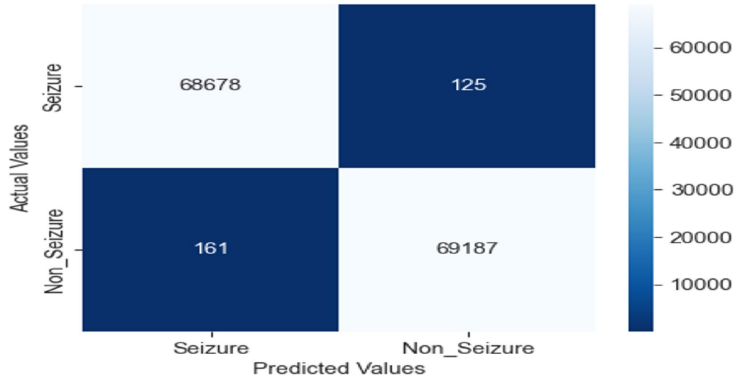


FIGURE 16
Confusion matrix of EEG data using the LRCN model.

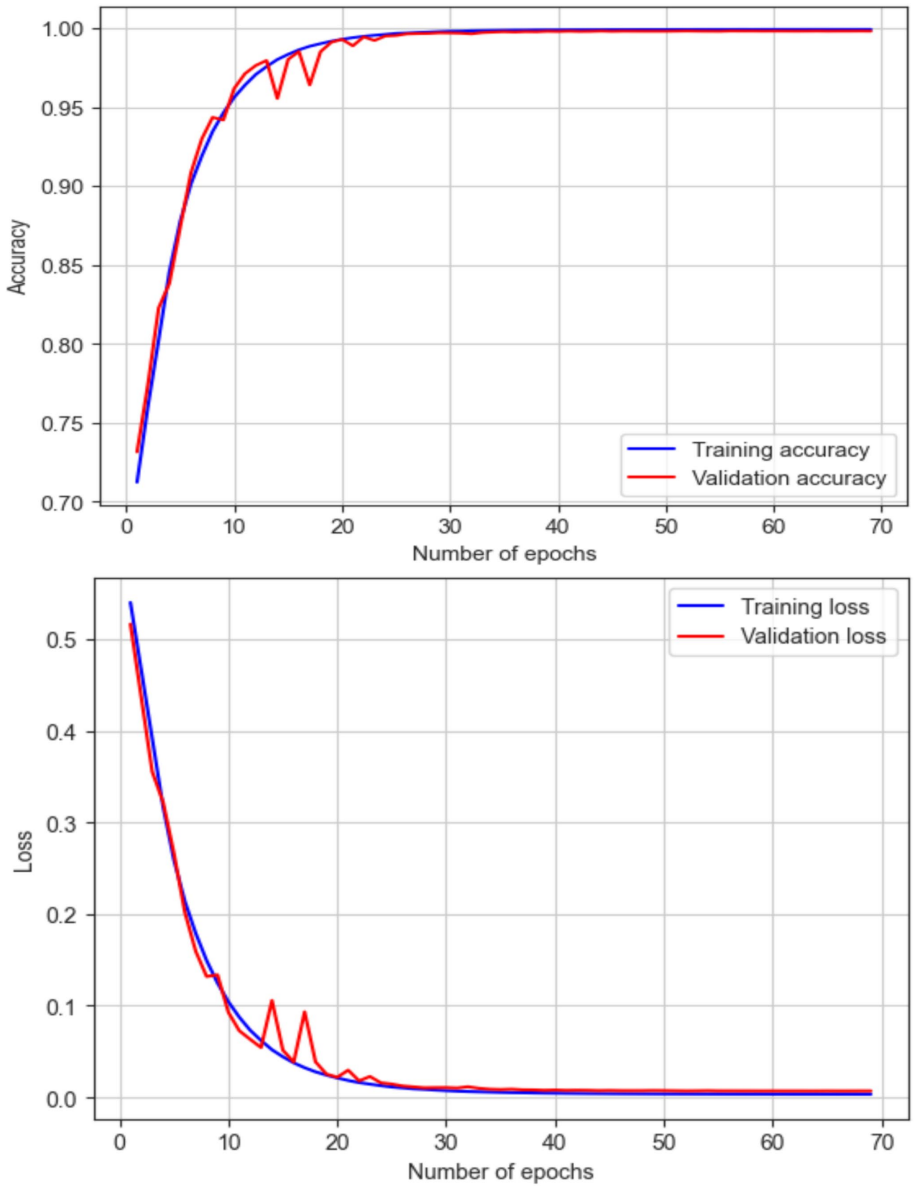


FIGURE 17
Accuracy and loss of EEG data using the LRCN model.

TABLE 8 EEG classification results summary.

Model	Accuracy %	Precision %	Recall %	F1 score %
GB	75.0	75.6	74.3	74.9
KNN	96.3	95.9	96.7	96.3
RFC	99.8	99.9	99.8	99.8
LSTM	99.0	99.1	99.0	99.0
LRCN	99.7	99.8	99.7	99.7

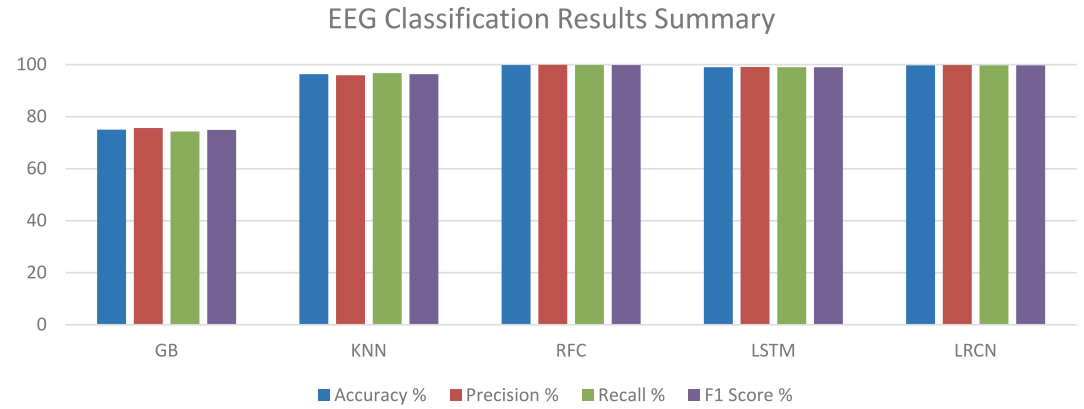


FIGURE 18
EEG classification results summary.

TABLE 9 EEG monitoring for detecting seizure behavior comparative.

Study	Model	Results
Our study	Random forest classifier	99.9% accuracy
Liu et al. (10)	Hybrid bilinear deep learning network	97.4% F1-score (TUH), 97.2% F1-score (EPILEPSIAE)
Fergus et al. (6)	k-NN classifier	88% sensitivity and specificity
Raghu et al. (7)	SVM with SDI feature	95.80–97.53% sensitivity, 0.4–0.57/h false detection rate
Rani and Chellam (8)	SVM with Peak Signal Features	99.60% accuracy
Almustafa (9)	Random Forest	97.08% accuracy
Zhao et al. (11)	Linear Graph Convolution Network	99.30% accuracy
Gabeff et al. (12)	CNN	0.873 F1-score, 90% seizure detection
Chou et al. (13)	CNN (various architectures)	97.7% accuracy (best model)
Kunekar et al. (15)	LSTM	97% validation accuracy
Mert and Akan (3)	Novel EEG analysis methodologies	97.89% accuracy
Brari and Belghith (17)	Chaos and fractal theory-based ML	100% accuracy
Shah et al. (18)	Random Neural Networks with DWT	93.27% (CHB-MIT), 99.84% (Bonn) accuracy
Kantipudi et al. (19)	FLHF, GBSO, and TAENN	99.6% detection performance
Zeng et al. (21)	Hybrid deep and shallow learning	Nearly 100% accuracy
Polat and Nour (24)	SVM with various kernels	76.70–82.50% accuracy
Hamlin et al. (26)	LDA with non-cerebral sensors	96% mean ROC value

challenges to bridge the gap between research performance and clinical applicability (see Table 9).

To contextualize the performance of our proposed framework, we provide a detailed comparison with recent state-of-the-art methods in EEG-based seizure detection. Table 10 summarizes key metrics, datasets, and methodologies, emphasizing the strengths of our approach.

Deploying EEG-based seizure detection in clinical settings faces computational and practical hurdles. While our Random Forest (RF)

TABLE 10 Comparative analysis with state-of-the-art seizure detection approaches.

Study	Model/approach	Dataset	Accuracy	Sensitivity	Specificity	F1-Score
Our study	Random Forest (RF)	102 patients	99.9%	99.8%	99.9%	99.8%
Liu et al. (10)	Hybrid Bilinear CNN + RNN	TUH, EPILEPSIAE	97.2	–	–	97.4%
Zhao et al. (11)	Linear Graph ConvNet (LGCN)	CHB-MIT	99.3%	99.4%	98.8%	–
Kantipudi et al. (19)	GBSO-TAENN (Bio-inspired NN)	Undisclosed	99.6%	–	–	99.0%
Gabeff et al. (12)	CNN	REPO2MSE	90%	–	–	87.3%

model achieves 99.9% accuracy with low latency (<10 ms) on CPUs, deep learning (DL) models like LSTM/LRCN require GPUs and exhibit higher latency (80–120 ms), limiting real-time use in wearables. Scalability and power constraints further favor RF, which processes 100 + EEG streams efficiently (~2 W) compared to DL's GPU-dependent demands (~150 W). Additionally, long-term EEG monitoring poses comfort challenges, as patients must wear sensor caps for extended periods—a barrier for ambulatory use but manageable for admitted patients under supervision. For hospitalized individuals, continuous EEG provides critical insights despite discomfort, enabling timely interventions. Future work must address hardware miniaturization (e.g., flexible, wireless electrodes) and hybrid models to balance accuracy, comfort, and regulatory compliance (e.g., IEC 62304). These steps are vital to translate lab advancements into bedside solutions.

5 Conclusion

This study demonstrates that EEG signals remain a robust source for epileptic seizure detection, with the RF classifier achieving a remarkable 99.9% accuracy. Although deep learning models, such as LSTM and LRCN, also performed well, the superior results of RF underscore the relevance of traditional machine learning approaches in clinical seizure detection. These findings indicate that RF offers a viable solution for practical EEG-based seizure monitoring due to its accuracy and generalizability. However, the practical challenges associated with continuous, long-term EEG monitoring necessitate further exploration of alternative non-invasive monitoring techniques. Future research should focus on reducing the number of electrodes required for EEG-based detection without compromising accuracy, investigate dry electrode technologies, and integrate EEG with other modalities, such as video and EMG, for more comprehensive seizure monitoring solutions. Moreover, addressing the challenges of real-time detection and generalizability across diverse patient populations remains paramount for the widespread clinical adoption of EEG-based seizure detection systems.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://datadryad.org/stash/dataset/doi:10.5061/dryad.xsj3tx99w.

Author contributions

MA-A: Conceptualization, Investigation, Software, Writing – original draft, Writing – review & editing. SA: Data curation, Methodology, Supervision, Writing – original draft, Writing – review & editing, Resources. AA: Formal analysis, Project administration, Validation, Writing – original draft, Writing – review & editing. MA: Funding acquisition, Resources, Visualization, Writing – original draft, Writing – review & editing. MD: Conceptualization, Data curation, Formal analysis, Funding acquisition, Writing – original draft, Writing – review & editing. TA: Conceptualization, Methodology, Project administration, Resources, 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-2024-471.

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

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. Ein Shoka MM, Dessouky AE-S, Hemdan EED. EEG seizure detection: concepts, techniques, challenges, and future trends. *Multimed Tools Appl.* (2023) 82:42021–51. doi: 10.1007/s11042-023-15052-2

2. Gupta S, Ranga V, Agrawal P. Epil net: a novel approach to IoT based epileptic seizure prediction and diagnosis system using artificial intelligence. *ADCAIJ Adv Distrib Comput Artif Intell J.* (2022) 10:435–52. doi: 10.14201/adcaij2021104435452

3. Mert A, Akan A. Seizure onset detection based on frequency domain metric of empirical mode decomposition. *SIViP*. (2018) 12:1489–96. doi: 10.1007/s11760-018-1304-y

4. Acharya UR, Hagiwara Y, Adeli H. Automated seizure prediction. *Epilepsy Behav*. (2018) 88:251–61. doi: 10.1016/j.yebeh.2018.09.030

5. Van de Vel A, Cuppens K, Bonroy B, Milosevic M, Jansen K, Van Huffel S, et al. Non-EEG seizure detection systems and potential SUDEP prevention: state of the art: review and update. *Seizure*. (2016) 41:141–53. doi: 10.1016/j.seizure.2016.07.012

6. Fergus P, Hussain A, Hignett D, Al-Jumeily D, Abdel-Aziz K, Hamdan H. A machine learning system for automated whole-brain seizure detection. *Appl Comput Informatics*. (2016) 12:70–89. doi: 10.1016/j.aci.2015.01.001

7. Raghu S, Sriraam N, Vasudeva Rao S, Hegde AS, Kubben PL. Automated detection of epileptic seizures using successive decomposition index and support vector machine classifier in long-term EEG. *Neural Comput Appl*. (2020) 32:8965–84. doi: 10.1007/s00521-019-04389-1

8. Rani TP, Chellam GH. A novel peak signal feature segmentation process for epileptic seizure detection. *Int J Inf Technol*. (2021) 13:423–31. doi: 10.1007/s41870-020-00524-7

9. Almustafa KM. Classification of epileptic seizure dataset using different machine learning algorithms. *Inform Med Unlocked*. (2020) 21:100444. doi: 10.1016/j.imu.2020.100444

10. Liu T, Truong ND, Nikpour A, Zhou L, Kavehei O. Epileptic seizure classification with symmetric and hybrid bilinear models. *IEEE J Biomed Heal Inform*. (2020) 24:2844–51. doi: 10.1109/JBHI.2020.2984128

11. Zhao Y, Dong C, Zhang G, Wang Y, Chen X, Jia W, et al. EEG-based seizure detection using linear graph convolution network with focal loss. *Comput Methods Prog Biomed*. (2021) 208:106277. doi: 10.1016/j.cmpb.2021.106277

12. Gabeff V, Teijeiro T, Zapater M, Cammoun L, Rheims S, Ryvlin P, et al. Interpreting deep learning models for epileptic seizure detection on EEG signals. *Artif Intell Med*. (2021) 117:102084. doi: 10.1016/j.artmed.2021.102084

13. Chou CH, Shen TW, Tung H, Hsieh PF, Kuo CE, Chen TM, et al. Convolutional neural network-based fast seizure detection from video electroencephalograms. *Biomed Signal Process Control*. (2023) 80:104380. doi: 10.1016/j.bspc.2022.104380

14. Sun Y, Chen X. Automatic detection of epilepsy based on entropy feature fusion and convolutional neural network. *Oxidative Med Cell Longev*. (2022) 2022:1–13. doi: 10.1155/2022/1322826

15. Kunekar P, Gupta MK, Gaur P. Detection of epileptic seizure in EEG signals using machine learning and deep learning techniques. *J Eng Appl Sci*. (2024) 71:1–15. doi: 10.1186/s44147-023-00353

16. Bandarabadi M, Rasekhi J, Teixeira CA, Karami MR, Dourado A. On the proper selection of preictal period for seizure prediction. *Epilepsy Behav*. (2015) 46:158–66. doi: 10.1016/j.yebeh.2015.03.010

17. Brari Z, Belghith S. A novel machine learning model for the detection of epilepsy and epileptic seizures using electroencephalographic signals based on Chaos and fractal theories. *Math Probl Eng*. (2021) 2021:1–10. doi: 10.1155/2021/2107113

18. Shah SY, Larijani H, Gibson RM, Liarokapis D. Epileptic seizure classification based on random neural networks using discrete wavelet transform for electroencephalogram signal decomposition. *Appl Sci*. (2024) 14:599. doi: 10.3390/app14020599

19. Kantipudi MVVP, Kumar NSP, Aluvalu R, Selvarajan S, Kotecha K. An improved GBSO-TAENN-based EEG signal classification model for epileptic seizure detection. *Sci Rep*. (2024) 14:843. doi: 10.1038/s41598-024-51337-8

20. Ein Shoka AA, Dessouky MM, El-Sayed A, El-Din Hemdan E. An efficient CNN based epileptic seizures detection framework using encrypted EEG signals for secure telemedicine applications. *Alex Eng J*. (2023) 65:399–412. doi: 10.1016/j.aej.2022.10.014

21. Zeng W, Shan L, Su B, Du S. Epileptic seizure detection with deep EEG features by convolutional neural network and shallow classifiers. *Front Neurosci*. (2023) 17:1145526. doi: 10.3389/fnins.2023.1145526

22. Bhandari HC, Pandeya YR, Jha K, Jha S, Ahmad S. Exploring non-Euclidean approaches: a comprehensive survey on graph-based techniques for EEG signal analysis. *J Adv Inf Technol*. (2024) 15:1089–105. doi: 10.12720/jait.15.10.1089-1105

23. Singh S, Kaur H. An intelligent method for epilepsy seizure detection based on hybrid nonlinear EEG data features using adaptive signal decomposition methods. *Circuits Syst Signal Process*. (2022) 42:2782–803. doi: 10.1007/s00034-022-02223-z

24. Polat K, Nour M. Epileptic seizure detection based on new hybrid models with electroencephalogram signals. *Irbm*. (2020) 41:331–53. doi: 10.1016/j.irbm.2020.06.008

25. Farooq MS, Zulfiqar A, Riaz S. Epileptic seizure detection using machine learning: taxonomy, opportunities, and challenges. *Diagnostics*. (2023) 13:1058. doi: 10.3390/diagnostics13061058

26. Hamlin A, Kobylarz E, Lever JH, Taylor S, Ray L. Assessing the feasibility of detecting epileptic seizures using non-cerebral sensor data. *Comput Biol Med*. (2021) 130:104232. doi: 10.1016/j.compbiomed.2021.104232

27. Rahmani MKI, Ahmad S, Hussain MR, Ameen AK, Ali A, Shaman F, et al. Enhanced Nanoelectronic detection and classification of motor imagery electroencephalogram signal using a hybrid framework. *J Nanoelectron Optoelectron*. (2023) 18:1254–63. doi: 10.1166/jno.2023.3504

28. Aayesha, Qureshi MB, Afzaal M, Qureshi MS, Fayaz M. Machine learning-based EEG signals classification model for epileptic seizure detection. *Multimed Tools Appl*. (2021) 80:17849–77. doi: 10.1007/s11042-021-10597-6

29. George ST, Subathra MSP, Sairamya NJ, Susmitha L, Joel Premkumar M. Classification of epileptic EEG signals using PSO based artificial neural network and tunable-Q wavelet transform. *Biocybern Biomed Eng*. (2020) 40:709–28. doi: 10.1016/j.bbe.2020.02.001

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 19 June 2025
DOI 10.3389/fmed.2025.1618550

Check for updates

OPEN ACCESS

EDITED BY
Deepti Deshwal,
Maharaja Surajmal Institute of Technology, India

REVIEWED BY
Yang Li,
Shihezi University, China
Najib Ben Aoun,
Al Baha University, Saudi Arabia

*CORRESPONDENCE
Thomas I. Nathaniel
✉ nathanit@greenvillemed.sc.edu
Ahmad Almadhor
✉ aaalmadhor@ju.edu.sa

RECEIVED 26 April 2025
ACCEPTED 03 June 2025
PUBLISHED 19 June 2025

CITATION
Alsubai S, Ojo S, Nathaniel TI, Ayari M, Baili J, Almadhor A and Al Hejaili A (2025) Transfer deep learning and explainable AI framework for brain tumor and Alzheimer's detection across multiple datasets.
Front. Med. 12:1618550.
doi: 10.3389/fmed.2025.1618550

COPYRIGHT
© 2025 Alsubai, Ojo, Nathaniel, Ayari, Baili, Almadhor and Al Hejaili. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Transfer deep learning and explainable AI framework for brain tumor and Alzheimer's detection across multiple datasets

Shtwai Alsubai[1], Stephen Ojo[2], Thomas I. Nathaniel[3]*, Mohamed Ayari[4], Jamel Baili[5], Ahmad Almadhor[6]* and Abdullah Al Hejaili[7]

[1]College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia, [2]Department of Electrical and Computer Engineering, College of Engineering, Anderson University, Anderson, SC, United States, [3]School of Medicine Greenville, University of South Carolina, Columbia, SC, United States, [4]Faculty of Computing and Information Technology, Northern Border University, Arar, Saudi Arabia, [5]Department of Computer Engineering, College of Computer Science, King Khalid University, Abha, Saudi Arabia, [6]Department of Computer Engineering and Networks, College of Computer and Information Sciences, Jouf University, Sakaka, Saudi Arabia, [7]Faculty of Computers and Information Technology, Computer Science Department, University of Tabuk, Tabuk, Saudi Arabia

Introduction: The pressing need for accurate diagnostic tools in the medical field, particularly for diseases such as brain tumors and Alzheimer's, poses significant challenges to timely and effective treatment.

Methods: This study presents a novel approach to MRI image classification by integrating transfer learning with Explainable AI (XAI) techniques. The proposed method utilizes a hybrid CNN-VGG16 model, which leverages pre-trained features from the VGG16 architecture to enhance classification performance across three distinct MRI datasets: brain tumor classification, Alzheimer's disease detection, and a third dataset of brain tumors. A comprehensive preprocessing pipeline ensures optimal input quality and variability, including image normalization, resizing, and data augmentation.

Results: The model achieves accuracy rates of 94% on the brain tumor dataset, 81% on the augmented Alzheimer dataset, and 93% on the third dataset, underscoring its capability to differentiate various neurological conditions. Furthermore, the integration of SHapley Additive exPlanations (SHAP) provides a transparent view of the model's decision-making process, allowing clinicians to understand which regions of the MRI scans contribute to the classification outcomes.

Discussion: This research demonstrates the potential of combining advanced deep learning techniques with explainability to improve diagnostic accuracy and trust in AI applications within healthcare.

KEYWORDS

MRI image classification, transfer learning, explainable AI (XAI), hybrid CNN-VGG16 model, brain tumors, Alzheimer's disease, SHAP, medical imaging

1 Introduction

Brain tumors constitute a critical subset of central nervous system (CNS) disorders, with pathologies ranging from slow-growing benign masses to highly aggressive malignant neoplasms (1). Malignant types such as glioblastomas and anaplastic astrocytomas are particularly concerning due to their rapid proliferation, high invasiveness, and poor prognosis (2). The five-year relative survival rate for adults remains around 35.6%. These metastatic tumors are especially challenging due to their rapid infiltration into brain parenchyma and resistance to conventional therapies (3). The World Health Organization (WHO) classifies CNS tumors into grades I–IV based on histopathological, immunohistochemical, and molecular features (4), underscoring the need for early and accurate grading to guide clinical interventions.

Magnetic Resonance Imaging (MRI) remains the gold standard for brain tumor diagnosis and grading due to its superior soft tissue contrast and non-invasive nature (5). Advanced MRI modalities: such as T1-weighted (T1), contrast-enhanced T1 (T1C), T2-weighted (T2) (6), Fluid Attenuated Inversion Recovery (FLAIR) (7), Diffusion Tensor Imaging (DTI), Perfusion MRI, and MR Spectroscopy (MRS) (8) offer rich, multi-parametric information on tumor morphology, oedema, necrosis, vascularity, and infiltration (9). However, the manual interpretation of these high-dimensional images is time-consuming, prone to inter-observer variability, and particularly burdensome in resource-constrained settings with radiologist shortages (10). Tumor heterogeneity and overlapping imaging phenotypes further complicate diagnosis, prompting increased adoption of automated analysis tools powered by AI (11).

In parallel, neurodegenerative disorders like Alzheimer's disease (AD) pose unique diagnostic challenges. AD is characterized by progressive cognitive decline and structural brain changes such as cortical thinning and hippocampal atrophy, visible in MRI scans (12). Due to the limited availability of labeled data for early AD diagnosis, data augmentation techniques such as affine transformations, intensity scaling, noise injection, and GAN-based synthesis have been employed to improve model robustness (13). These enriched datasets also facilitate sequential transfer learning, enabling the repurposing of knowledge from AD-related imaging to other neurological domains, including brain tumor classification (14). Convolutional Neural Networks (CNNs) have tremendously succeeded in medical image classification, segmentation, and anomaly detection. Pre-trained architectures such as VGG16, ResNet, and DenseNet, initially developed for natural image datasets like ImageNet, can be fine-tuned via transfer learning to perform effectively in medical contexts (15).

This work proposes a novel hybrid framework that integrates a pre-trained VGG16 backbone with custom CNN layers and applies a sequential transfer learning strategy across three structurally distinct MRI datasets: a brain tumor, Alzheimer's disease, and an independent validation set. This approach leverages domain-relatedness in neuroimaging to enhance feature generalization and classification accuracy across multiple brain pathologies. Despite their high predictive performance, deep learning models are often criticized for their "black-box" nature, which limits interpretability and clinical trust (16). To overcome this limitation, we incorporate SHapley Additive exPlanations (SHAP), an explainable AI (XAI) method that attributes the model's output to specific pixels or regions in the input image. SHAP values offer visual insight into the regions most influential to model decisions, aligning them with anatomical structures and facilitating clinician interpretation. By striking a balance between high performance and interpretability, our framework presents a promising solution for real-world deployment in neuroimaging diagnostics.

This work proposes a novel hybrid framework that integrates a pre-trained VGG16 backbone with custom CNN layers and applies a sequential transfer learning strategy across three structurally distinct MRI datasets: a brain tumor, Alzheimer's disease, and an independent validation set. This approach leverages domain-relatedness in neuroimaging to enhance feature generalization and classification accuracy across multiple brain pathologies. Despite their high predictive performance, deep learning models are often criticized for their "black-box" nature, which limits interpretability and clinical trust (16). To overcome this limitation, we incorporate SHapley Additive exPlanations (SHAP), an explainable AI (XAI) method that attributes the model's output to specific pixels or regions in the input image. SHAP values offer visual insight into the regions most influential to model decisions, aligning them with anatomical structures and facilitating clinician interpretation. By striking a balance between high performance and interpretability, our framework presents a promising solution for real-world deployment in neuroimaging diagnostics.

The proposed method begins with preprocessing all datasets, including normalization, resizing, augmentation, and partitioning into train/validation/test splits. A hybrid CNN architecture is then constructed by combining frozen VGG16 features with custom convolutional and dense layers. The model is trained on a brain tumor dataset and then fine-tuned sequentially on an Alzheimer's dataset and a third validation dataset using transfer learning. Each stage involves model reconfiguration and controlled unfreezing of layers. Finally, SHAP-based explainability is applied to visualize model decisions, and performance is evaluated using standard metrics such as accuracy, precision, recall, F1-score, and confusion matrices.

Figure 1 illustrates the concept of transfer learning, a technique in machine learning where knowledge gained from a source domain is utilized to enhance learning in a target domain. The source domain comprises a large dataset, such as ImageNet, which contains over a million images. A pre-trained model is developed using this extensive dataset, comprising three key components: early layers for feature extraction, middle layers, and task-specific layers. In transfer learning, the early layers that capture general features like edges and textures are transferred to the model for the target domain, where data is limited, such as a medical image dataset with only hundreds of samples. These layers become "frozen" in the fine-tuned model, meaning they are not updated during training on the small dataset. The middle layers are fine-tuned, and the adjustments are based on the new data to capture domain-specific features better. Finally, the task-specific layers from the source model are replaced with new ones tailored to the target domain's specific task.

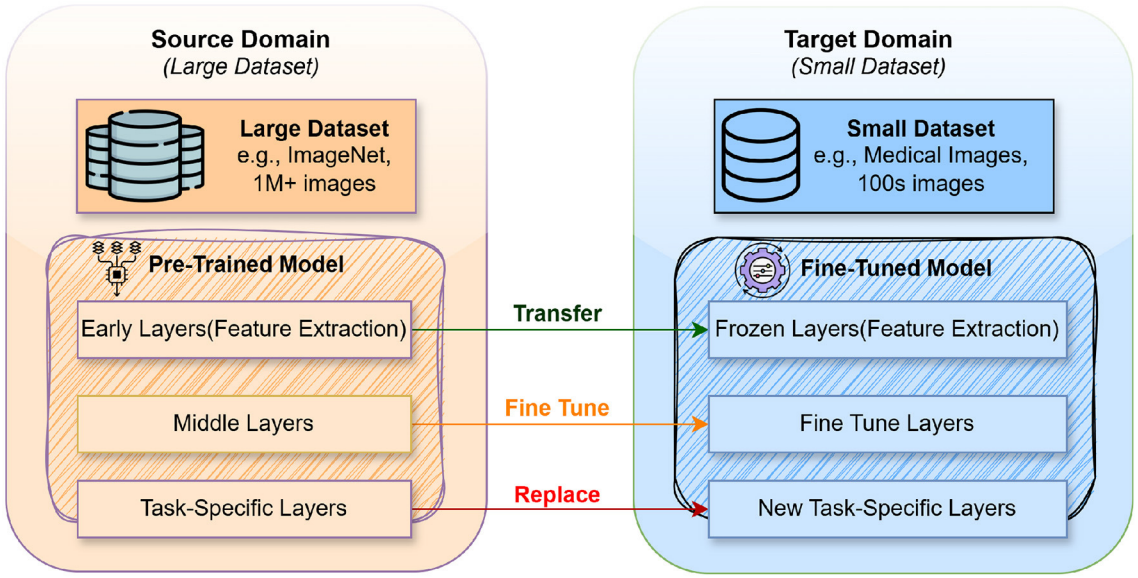


FIGURE 1
How transfer learning works.

1.1 Research contributions

The major research contributions of this study are the following:

- A novel approach that leverages a pre-trained VGG16 model combined with custom CNN layers, using sequential transfer learning across three distinct MRI datasets (brain tumor, Alzheimer's, and validation) to improve classification accuracy while requiring minimal training data.
- This study demonstrates effective knowledge transfer between different neurological conditions (from brain tumor classification to Alzheimer's detection), showing that features learned from one medical imaging domain can enhance performance in related but distinct diagnostic tasks. A comprehensive preprocessing pipeline, including image normalization, resizing, and data augmentation, is implemented to improve model robustness and generalizability across datasets with varying characteristics.
- This research incorporates SHapley Additive exPlanations (SHAP) analysis to provide transparent, pixel-level attribution of model decisions, addressing the "black box" problem of deep learning in healthcare by enabling clinicians to understand which regions of MRI scans influence diagnostic classifications.

1.2 Research organization

This research is organized into the following main sections. Section 2 presents related work, discussing recent advances in deep learning for medical imaging, the effectiveness of transfer learning, and the growing importance of XAI in healthcare. Section 3 outlines the proposed framework, detailing integrating pre-trained convolutional neural networks with XAI methods, such as Grad-CAM, to enhance performance and interpretability. Section 4 presents the experimental analysis, which includes dataset description, evaluation metrics, and results comparing the proposed model with existing techniques. Finally, Section 5 concludes the study by summarizing key findings and suggesting directions for future research.

2 Related work

This section presents related work, discussing recent advances in deep learning for medical imaging, the effectiveness of transfer learning, and the growing importance of XAI in healthcare. Tuncer et al. (17) proposed a lightweight convolutional neural network named FiboNeXt for Alzheimer's disease classification using MRI images. The model was designed by integrating ConvNeXt architecture elements, attention, and concatenation layers. The dataset was divided into four classes and included both original and augmented versions, where the augmented data was used for training and the original for testing. The primary aim was to achieve high accuracy with fewer trainable parameters. Experimental results demonstrated that FiboNeXt achieved 95.40 and 95.93% validation accuracy on two datasets, while test accuracy reached 99.66 and 99.63%, respectively, highlighting the model's efficiency and generalization capability. An optimized hybrid transfer learning (TL) framework was introduced by Lasagni et al. (18) to classify brain tumors using MRI images. The approach combined advanced preprocessing techniques, such as noise reduction and contrast enhancement, with an ensemble of pretrained deep learning models, VGG16 and ResNet152V2. The framework achieved an impressive classification accuracy of 99.47% on a complex four-class dataset. Explainable AI (XAI) methods like SHAP and Grad-CAM were employed to ensure transparency and clinical trust. These tools provided visual and

quantitative insights into model predictions, facilitating better interpretability and making the model more suitable for real-world clinical applications.

Bhaskaran and Datta (19) investigated the use of 3D convolutional neural networks (3D-CNNs) for detecting focal cortical dysplasia (FCD) from a dataset containing MRI scans of 170 individuals (85 patients and 85 controls). They studied the advantages of cross-modality transfer learning using pretrained ResNet variants (ResNet-18, -34, and -50, trained initially on segmentation tasks). Transfer learning significantly improved classification performance to up to 80.3%. Moreover, they also introduced a novel Heat-Score, a combination of Grad-CAM, to evaluate the model interpretability. The model was able to fill the gap between AI predictions and expert diagnostic insights by using this metric, showing the model's effectiveness in identifying clinically relevant seizure zones. Tonni et al. (20) used the InceptionV3 architecture to classify brain MRI images into three tumor types (meningioma, glioma and pituitary) with different embeddings initialization for imagenet and the studied data. Several open-source XAI tools were integrated to address the challenge of model interpretability, including LIME, SHAP, and Grad-CAM. The model attained a classification accuracy of 93% and an F1-score of 0.93. Among the XAI tools, SHAP provided the highest level of explainability at ∼60%, aligning better with expert-identified tumor regions. In contrast, LIME and Grad-CAM explained <50% of the cases. The findings revealed that non-tumor-related features had a notable impact on model predictions, suggesting a need for further refinement in feature attribution techniques.

Nahiduzzaman et al. (21) proposed a novel framework that integrates a lightweight parallel depthwise separable convolutional neural network (PDSCNN) with a hybrid ridge regression extreme learning machine (RRELM) for classifying four brain tumor types (glioma, meningioma, pituitary, and no tumor) using MRI images. The approach utilizes contrast-limited adaptive histogram equalization (CLAHE) to enhance tumor feature visibility, followed by PDSCNN for efficient tumor-specific feature extraction with reduced computational cost. To improve classification performance, a ridge regression-enhanced ELM (RRELM) is introduced, addressing the limitations of traditional ELMs. Comparative analysis with state-of-the-art models revealed that the proposed PDSCNN-RRELM achieved superior results, with average precision, recall, and accuracy reaching 99.35%, 99.30%, and 99.22% through five-fold cross-validation. Vanaja et al. (22) proposed a diagnostic framework for Alzheimer's Disease (AD) by leveraging machine learning and a customized deep convolutional neural network (cDCNN) with three convolutional layers applied to MRI data. The analysis incorporates two datasets, Alzheimer's Disease Neuroimaging Initiative (ADNI) and a Kaggle dataset, to examine diverse subject groups and imaging characteristics linked to AD pathology. To mitigate class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is employed. Traditional machine learning classifiers such as support vector machine, k-nearest neighbor, random forest, decision trees, and XGBoost are evaluated alongside the cDCNN model, which focuses on key MRI biomarkers of AD. The cDCNN achieved 87% accuracy on the ADNI dataset despite preprocessing challenges due to converting DICOM images to JPEG, which affected image quality.

Joshi et al. (23) introduced a transfer learning approach for classifying Parkinson's disease using the imbalanced PPMI dataset, leveraging Big Transfer (BiT) models. These pre-trained models utilize Group Normalization with Weight Standardization and adopt BiT-HyperRule for effective fine-tuning across diverse datasets. Various BiT architectures, including BiT-S and BiT-M variants, were evaluated. The best-performing model, BiT-M152x4, achieved 86.71% accuracy, surpassing the previous state-of-the-art RA-GCN model (76%). Additionally, the same BiT models were applied to the imbalanced BCCD dataset, where BiT-M152x4 again outperformed VGG16 (98.52% vs. 74%), demonstrating the versatility and robustness of the proposed approach. Bin Shabbir Mugdha and Uddin (24) conducted a comparative analysis between a newly developed Convolutional Neural Network (CNN) model and several pre-trained models using transfer learning, including VGG-16, ResNet-50, AlexNet, and Inception-v3. VGG-16 achieved the best performance among all models with a test accuracy of 95.52%, training accuracy of 99.87%, and a validation loss of 0.2348. ResNet-50 followed with 93.31% test accuracy, 98.78% training accuracy, and 0.6327 validation loss. The custom CNN model achieved 92.59% test accuracy, 98.11% training accuracy, and a validation loss of 0.2960. Inception-v3 showed the lowest performance with 89.40% test accuracy and a validation loss of 0.4418.

Khedgaonkar et al. (25) proposed a Graph Neural Network (GNN)-based approach for brain MRI classification, addressing the limitations of traditional methods in integrating spatial and frequency domain features. By applying Fourier, Gabor, and convolutional transformations, key features are extracted and fused into a unified representation. MRI images are modeled as nodes in a graph, capturing structural and semantic relationships. The GNN leverages this graph structure to learn discriminative features through neighborhood aggregation. The method demonstrated superior performance across precision, accuracy, recall, specificity, AUC, and delay, outperforming conventional techniques. Ilani et al. (26) focused on classifying brain tumors glioma, meningioma, and pituitary using MRI scans, leveraging the U-Net architecture for segmentation alongside transfer learning-based CNN models such as Inception-V3, EfficientNetB4, and VGG19. Model performance was evaluated using F-score, recall, precision, and accuracy metrics. U-Net outperformed other models, achieving 98.56% accuracy, a 99% F-score, 99.8% AUC, and 99% recall and precision. It also maintained strong generalization with 96.01% accuracy in cross-dataset validation using an external cohort. The results highlight U-Net's effectiveness in precise brain tumor segmentation, supporting early diagnosis and treatment planning.

Rasool et al. (27) proposed ResMHA-Net, a deep learning framework combining ResNet residual blocks with multi-head attention to enhance glioma segmentation in 3D MRI. This architecture captured long-range dependencies and emphasized informative regions, improving the segmentation of complex glioma sub-regions. It was trained and validated on BraTS 2018–2021 datasets, with the best performance observed on BraTS 2021, demonstrating strong adaptability. Predicted masks from three datasets were used to extract radiomic features, which, along with clinical data, trained an ensemble model for survival prediction. This model employed a voting mechanism across multiple learners and achieved a 73% overall survival prediction accuracy. Gasmi et al. (28) developed an ensemble classification model integrating

Vision Transformers (ViT) and EfficientNet-V2 to capture both global and local features from brain MRI. Model outputs were combined using a genetic algorithm-optimized weighted ensemble, which selected the best combination to maximize accuracy. Trained on a labeled MRI dataset, the ensemble model outperformed individual and traditional classifiers, achieving a 95% classification accuracy with improved precision, recall, and F1-score.

While these studies have achieved high accuracy through various architectures and optimization techniques, many face limitations such as reliance on single-domain datasets, limited transferability across neurological disorders, or insufficient interpretability. Most existing works focus on enhancing performance or providing visual explanations, but few offer a unified framework that balances generalization, accuracy, and explainability across diverse brain pathologies. Furthermore, many methods lack rigorous evaluation of independent datasets, raising concerns about overfitting and real-world applicability. Our work addresses these gaps by proposing a multi-stage transfer learning strategy that spans distinct MRI datasets and integrating SHAP for transparent, clinically meaningful explanations.

3 Proposed framework

This section explains the proposed framework, detailing the integration of pre-trained convolutional neural networks with XAI methods like Grad-CAM to improve performance and interpretability. The workflow of the proposed framework is illustrated in Figure 2. The figure presents a comprehensive pipeline for a Hybrid CNN-VGG16 model designed for MRI image classification, which leverages transfer learning and explainable artificial intelligence (XAI) techniques. The process is divided

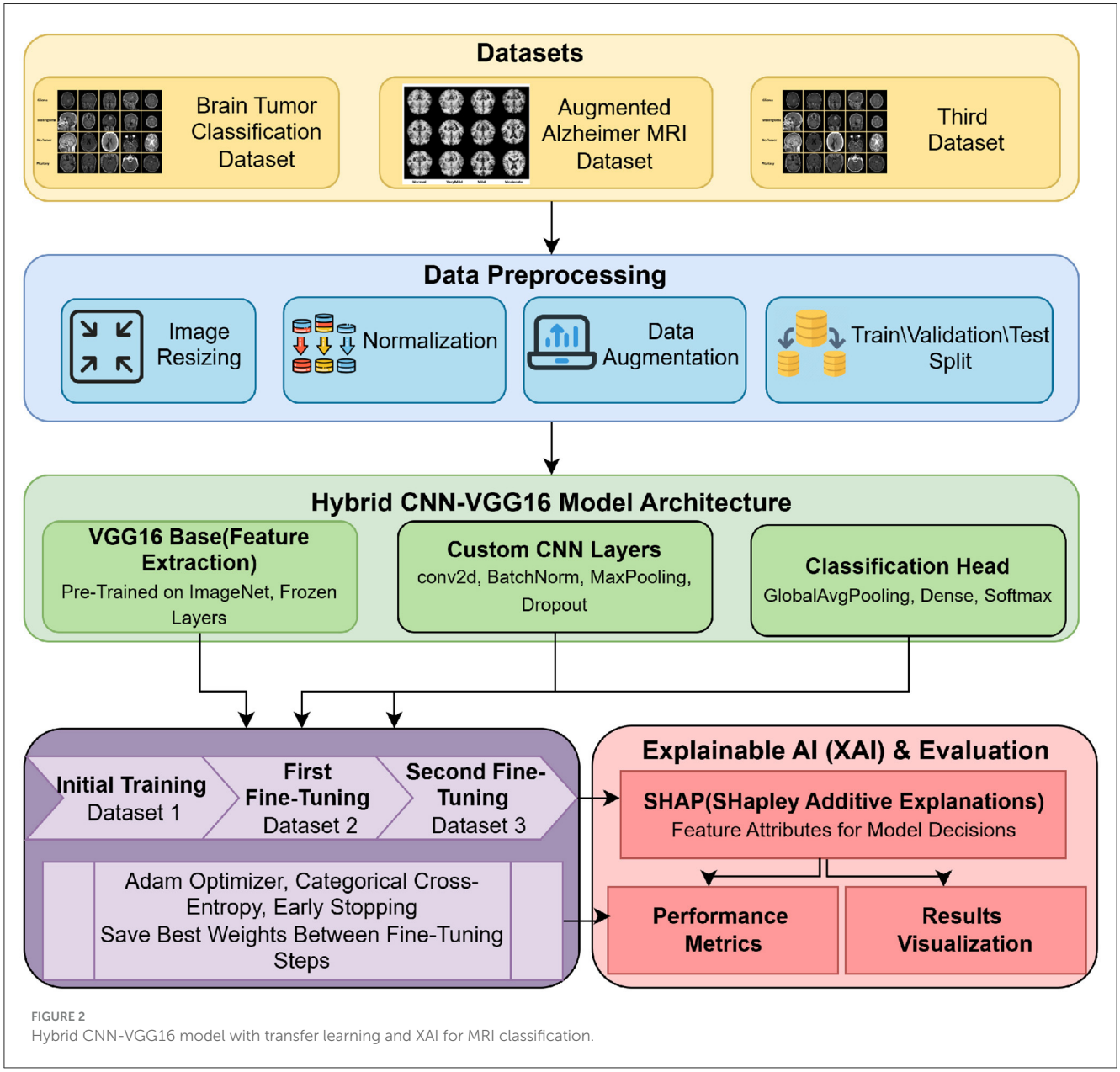


FIGURE 2
Hybrid CNN-VGG16 model with transfer learning and XAI for MRI classification.

into five primary stages: datasets, data preprocessing, model architecture, training, and evaluation with XAI. The first stage highlights the use of three distinct datasets: the Brain Tumor Classification Dataset (with classes like glioma, meningioma, no tumor, and pituitary), the Augmented Alzheimer MRI Dataset (including mild, moderate, non-demented, and very mild demented classes), and a third dataset which again covers brain tumor categories. These datasets undergo different preprocessing steps, such as image resizing, normalization, augmentation, and dataset splitting into training, validation, and testing sets. Next, the Hybrid CNN-VGG16 model architecture is detailed. It begins with the VGG16 base model pretrained on ImageNet with frozen layers used for feature extraction. On top of this base, custom convolutional layers (including Conv2D, batch normalization, max pooling, and dropout) are added to enhance learning. The final part of the model is the classification head, which includes global average pooling, dense layers, and a softmax layer for multi-class output. The training process is conducted in three sequential phases. It starts with initial training on the brain tumor dataset, followed by two fine-tuning stages on the Alzheimer dataset and then on the third dataset. The training uses the Adam optimizer, categorical cross-entropy loss, and early stopping, with the best-performing model weights preserved between each stage. Finally, the Explainable AI (XAI) & Evaluation block involves model interpretation and performance assessment. SHapley Additive exPlanations (SHAP) provides feature attributions, allowing insight into how the model makes decisions. Additionally, several performance metrics such as accuracy, F1-score, precision, and recall are used, and visual results are presented via confusion matrices and SHAP plots.

Algorithm 1 defines a general process to adapt a pre-trained source model M_S to a new target task using the target dataset D_T. The source model is cloned to create the target model M_T, after which selected layers are frozen based on the strategy ϕ. The final output layer is replaced to align with the target labels, and the dataset D_T is split into training, validation, and test subsets. Fine-tuning is performed over E epochs using gradient descent on trainable parameters, with early stopping optionally applied. The algorithm also supports the progressive unfreezing of layers for staged fine-tuning. The final model is evaluated on the D_T^{test} test set. Specifically, the following terms are: M_S denotes the pre-trained source model, and M_T is the target model initialized as a clone of M_S. The target dataset is represented as $D_T = \{(x_i, y_i)\}_{i=1}^{N_T}$, where x_i is an input sample, y_i is the corresponding target label, and N_T is the total number of samples. The learning rate is denoted by α, and E represents the number of training epochs. The strategy ϕ defines which layers in M_T will be frozen or trainable during fine-tuning. Each mini-batch is represented by $B = \{(x_j, y_j)\}_{j=1}^{b}$, where b is the batch size. For each sample x_j in the batch, $\hat{y}_j$ is the predicted output by M_T. The loss for a batch is computed as $\mathcal{L} = \frac{1}{b}\sum_j \ell(\hat{y}_j, y_j)$, where ℓ is a loss function such as cross-entropy. The model parameters are denoted by θ, and gradient descent updates them via $\theta \leftarrow \theta - \alpha\nabla_\theta\mathcal{L}$. The dataset D_T is split into training, validation, and test sets, denoted by D_T^{train}, D_T^{val}, and D_T^{test}, respectively. Additionally, if progressive unfreezing is enabled, layers are incrementally unfrozen in S stages, with each stage using its learning rate α_s and epoch count E_s.

Algorithm 2 details a pipeline for MRI image classification using three datasets. The datasets are defined as follows: $D_1 = \{(x_i^1, y_i^1)\}_{i=1}^{N_1}$ corresponds to the Brain Tumor Dataset (BTD), $D_2 = \{(x_i^2, y_i^2)\}_{i=1}^{N_2}$ is the Alzheimer Dataset (AD), and $D_3 = \{(x_i^3, y_i^3)\}_{i=1}^{N_3}$ is the third validation dataset (VD). Here, x_i^k is an MRI image, and y_i^k is its corresponding label for dataset D_k with N_k samples. The learning rate, batch size, and number of epochs for training on dataset D_k are represented by α_k, B_k, and E_k, respectively. During preprocessing, each image x_i is normalized by subtracting the mean μ and dividing by the standard deviation σ, then resized to a fixed height h and width w. Augmentation is applied through transformation functions $T(x_i)$, and the dataset is split into training, validation, and test subsets. The model is constructed using a pretrained VGG16 backbone denoted as V, from which features F are extracted. These features are frozen and connected to additional convolutional, batch normalization (BN), max pooling, dropout, global average pooling (GAP), and dense layers, ending with a final dense output layer with C units representing the number of classes. The function `Train` compiles the model with the Adam optimizer (learning rate α) and categorical cross-entropy (CCE) loss, then fits it on the training set and evaluates it on the

Require: Source model M_S, source dataset D_S, target dataset $D_T = \{(x_i, y_i)\}_{i=1}^{N_T}$, learning rate α, epochs E, freezing strategy ϕ
Ensure: Fine-tuned model M_T
function TransferLearn(M_S, D_T, α, E, ϕ)
  $M_T \leftarrow M_S$
  for each layer l in M_T **do**
    if $l \in \phi$ **then**
      Freeze l
    else
      Make l trainable
    end if
  end for
  Replace output layer of M_T to match classes in D_T
  Split D_T into D_T^{train}, D_T^{val}, D_T^{test}
  for $e = 1$ to E **do**
    for each batch $B = \{(x_j, y_j)\}_{j=1}^{b} \subset D_T^{train}$ **do**
      $\hat{y}_j \leftarrow M_T(x_j)$ for all $x_j \in B$
      $\mathcal{L} \leftarrow \frac{1}{b}\sum_j \ell(\hat{y}_j, y_j)$
      Update $\theta \leftarrow \theta - \alpha\nabla_\theta\mathcal{L}$ for all trainable θ
    end for
    Evaluate M_T on D_T^{val}
    if early stopping criteria met **then**
      break
    end if
  end for
  if progressive unfreezing enabled **then**
    for $s = 1$ to S **do**
      Unfreeze new layers per strategy ϕ
      Fine-tune with reduced α_s for E_s epochs
    end for
  end if
  Evaluate M_T on D_T^{test}
  return M_T
end function

Algorithm 1. Transfer learning for neural network models.

Require: $D_1 = \{(x_i^1, y_i^1)\}_{i=1}^{N_1}$ (BTD), $D_2 = \{(x_i^2, y_i^2)\}_{i=1}^{N_2}$ (AD), $D_3 = \{(x_i^3, y_i^3)\}_{i=1}^{N_3}$ (VD)
Require: $\alpha_k, B_k, E_k \quad \forall k \in \{1, 2, 3\}$
function PREPROCESS(D)
  $x_i \leftarrow \frac{x_i - \mu}{\sigma}$, $x_i \leftarrow \text{resize}(x_i, h, w)$
  $D^{aug} \leftarrow D \cup \{(T(x_i), y_i)\}$
  $D^{train}, D^{val}, D^{test} \leftarrow \text{split}(D^{aug})$
  return $D^{train}, D^{val}, D^{test}$
end function
function BUILD(C)
  $V \leftarrow \text{VGG16(pretrained)}$, $F \leftarrow \text{extract}(V)$
  freeze(F)
  $M \leftarrow F \rightarrow \text{Conv2D}(256) \rightarrow \text{BN} \rightarrow \text{MaxPool} \rightarrow \text{Drop}(0.3)$
  $M \leftarrow M \rightarrow \text{GAP} \rightarrow \text{Dense}(512) \rightarrow \text{Drop}(0.5) \rightarrow \text{Dense}(C)$
  return M
end function
function TRAIN($M, D^{train}, D^{test}, \alpha, B, E$)
  compile(M, Adam(α), CCE)
  fit(M, D^{train}, E, B)
  eval(M, D^{test})
  return M
end function
function FINETUNE($M, D^{new}, C, \alpha, B, E$)
  replace_head(M, C)
  unfreeze(M.tail)
  compile(M, Adam(α), CCE)
  fit(M, D^{new}, E, B)
  return M
end function
function EXPLAIN(M, X)
  $E \leftarrow \text{DeepExplainer}(M, X_{bg})$
  for $x_i \in X$ **do**
    $\hat{y}_i \leftarrow \arg\max M(x_i)$
    $S_i \leftarrow E(x_i)$
    plot(S_i, x_i)
  end for
end function
function EVAL(M, D)
  Compute: Acc, F1, Prec, Rec, conf_mat(M, D)
end function
$D_1^* \leftarrow$ PREPROCESS(D_1), $D_2^* \leftarrow$ PREPROCESS(D_2), $D_3^* \leftarrow$ PREPROCESS(D_3)
$M \leftarrow$ BUILD($|C_1|$)
$M \leftarrow$ TRAIN($M, D_1^{train}, D_1^{test}, \alpha_1, B_1, E_1$)
$M \leftarrow$ FINETUNE($M, D_2^{train}, |C_2|, \alpha_2, B_2, E_2$)
$M \leftarrow$ FINETUNE($M, D_3^{train}, |C_3|, \alpha_3, B_3, E_3$)
EXPLAIN(M, D_3^{test})
EVAL(M, D_1^{test}), EVAL(M, D_2^{test}), EVAL(M, D_3^{test})

Algorithm 2. Hybrid CNN-VGG16 with TL and XAI for MRI classification.

test set. The function FineTune replaces the output head with C classes, unfreezes the last layers for fine-tuning, recompiles the model, and continues training. The Explain function employs DeepExplainer from SHAP to generate saliency maps S_i for test samples x_i, where X_{bg} is a background dataset used for explanations. The predicted label for a sample is given by $\hat{y}_i = \arg\max M(x_i)$. The evaluation function computes standard metrics: accuracy (Acc), F1-score (F1), precision (Prec), recall (Rec), and confusion matrices. These changes have been incorporated to improve the transparency of the algorithm.

3.1 Experimental dataset

In this research, we utilized three datasets for classifying MRI images by training deep learning models. The first dataset (https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumorclassification-mri, accessed March 25, 2025) is based on Brain tumors, among the most aggressive diseases affecting children and adults, comprising 85%–90% of all primary Central Nervous System (CNS) tumors. Annually, ∼11,700 new brain tumor cases are reported, with a 5-year survival rate of 34% for men and 36% for women. Tumors are categorized into the following types: Glioma Tumor, Meningioma Tumor, No Tumor, and Pituitary Tumor. The second dataset (https://www.kaggle.com/datasets/uraninjo/augmented-alzheimer-mri-dataset, accessed March 25, 2025) used in this research is the Augmented Alzheimer's MRI Dataset. It contains brain MRI images classified into four categories: Non-Demented, Very Mild Demented, Mild Demented, and Moderate Demented. The dataset is organized into two main folders, one containing the original images and the other containing augmented versions to increase data variability. Both training and testing sets include samples from all four classes. Augmented data helps improve deep learning models' performance and generalization capability in classifying different stages of Alzheimer's disease. The third dataset (https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri, accessed March 25, 2025) used in this research is a combined brain tumor MRI dataset derived from three sources: Figshare, the SARTAJ dataset, and the Br35H dataset. It contains 7,023 MRI images classified into four categories: Glioma, Meningioma, Pituitary, and No Tumor. Images for the "No Tumor" class were taken from the Br35H dataset. Due to misclassification issues observed in the Glioma class of the SARTAJ dataset, which was identified through inconsistent model performance and validation against other research, those images were removed and replaced with correctly labeled images from the Figshare dataset. This curated dataset supports the classification of brain tumors, which can be either benign or malignant and is critical for early diagnosis, given the life-threatening nature of tumor-induced pressure within the skull.

3.2 Data preprocessing

The preprocessing process begins with loading each MRI image and converting it from the default Blue-Green-Red (BGR) color format to the standard Red-Green-Blue (RGB) format to ensure compatibility with deep learning models. This conversion maintains consistency in color representation across all images, preventing misinterpretation of visual features during training and improving the accuracy of tumor classification. After converting

the image to RGB, the next preprocessing step involves resizing each image to a fixed dimension of 128 × 128 pixels. Neural networks require input data to have a consistent shape, and resizing ensures that all images, regardless of their original resolution, meet the input requirements of the model. Specifically, resizing transforms an image $I \epsilon R^{H \times W \times 3}$ into a standardized format $I^{'} \epsilon R^{128 \times 128 \times 3}$, where H and W represent the original height and width of the image, respectively. This step ensures that all input images are of uniform size, allowing for efficient model training and processing. After resizing, the pixel values of the images are normalized by scaling them from the original range of [0, 255] to [0, 1]. This is achieved by dividing each pixel value by 255 (see Equation 1):

$$I_{norm} = \frac{I}{255} \tag{1}$$

Normalization helps stabilize and accelerate the neural network's learning process by ensuring the input data has a smaller, more uniform range of values. It also helps reduce the internal covariate shift, thus enabling more effective weight updates during training.

The class labels, initially string values such as "glioma_tumor," "meningioma_tumor," etc., are converted into a numerical format using a label map. Each label is then one-hot encoded using the to_categorical() function. One-hot encoding transforms categorical labels into a binary matrix where only the index of the class is marked as 1, and all others are 0. For instance, the label "glioma_tumor" becomes [1, 0, 0, 0]. This format is compatible with multi-class classification models. In mathematical terms, for a class $C \in \{0, 1, 2, 3\}$, the one-hot encoded vector y is defined as shown in Equation 2:

$$y_i = \begin{cases} 1, & \text{if } i = C \\ 0, & \text{otherwise} \end{cases} \quad \text{for } i \in \{0, 1, 2, 3\} \tag{2}$$

It is essential to evaluate model performance and prevent overfitting; therefore, this study utilized train_test_split() to divide the dataset into training and validation sets. Specifically, 80% of the data was allocated for training and 20% for validation. The use of a fixed random_state ensured reproducibility. This separation allowed the model to be assessed on unseen data, providing a more precise measure of its generalization capability.

3.3 Data augmentation

The model generalization should be improved together with mitigating overfitting if we have a small dataset. For this, this study expanded the training data using data augmentation techniques, which artificially increased the training dataset by generating simple variations of the images. These variations make the model more robust and well-performing for real-world transformations that may occur in medical imaging.

The operations applied in this study for augmentation are random rotations in a ±20-degree range, horizontal and vertical translations of 20% of the Image dimensions, shear transformation with moderate intensity, zooming in ± 20% and random horizontal flips to simulate different orientations. These transformations were chosen carefully to resemble variations in MRI scans that occur naturally and do not modify the underlying anatomical structures. The augmentation process can be formally described as applying a transformation function T to an input image x, resulting in an augmented image x' (see Equation 3):

$$x^{'} = T(x) \tag{3}$$

Where the transformation function T is a composition of individual operations, such as (see Equation 4):

$$\begin{aligned} x' &= R_\theta(x) && \text{(Rotation)} \\ x' &= T_{dx,dy}(x) && \text{(Translation)} \\ x' &= S_\alpha(x) && \text{(Shear)} \\ x' &= Z_s(x) && \text{(Zoom)} \\ x' &= F(x) && \text{(Flip)} \end{aligned} \tag{4}$$

These operations ensure that the data is represented in diverse ways during the training process, thus increasing the chances for it to generalize better to unseen inputs. The augmentation parameters were fit to the training dataset before training, and these fit parameters were used in the training process. Hence, the behavior of transformation is consistent during the time of learning.

3.4 Model architecture

The details of the model architecture of CNN, Custom CNN, VGG16, ResNet and Hybrid CNN-VGG16 are discussed in this section.

3.4.1 CNN model

The first model architecture specifically for the MRI image classification is the Convolutional Neural Network (CNN) (29). It takes input images of size 128 × 128 × 3 and starts with a Conv2D layer of 32 filters (3 × 3 kernel, ReLU) and, as usual, MaxPooling2D (2 × 2) to reduce spatial dimensions. It is followed by a Conv2D layer with 64 filters (3 × 3, ReLU) and another MaxPooling2D (2 × 2). Then, a third Conv2D layer with 128 filters (3 × 3, ReLU) and another MaxPooling2D layer (2 × 2) is added. It is then flattened and passed through a Dense layer with 128 neurons (ReLU) and a Dropout layer of 0.5 dropout rate to prevent overfitting. The Dense output layer with a softmax activation is used to classify the input into one of four classes: glioma, meningioma, no tumor and pituitary tumor. Lastly, we compile the model using the Adam optimizer and the categorical cross-entropy loss, which fit the multi-class classification correctly.

3.4.2 Custom CNN model

The custom CNN model shares the core structure of the basic CNN three convolutional layers followed by max-pooling, flattening, a dense layer, dropout, and a softmax output for multi-class classification. However, it enhances the architecture by

integrating Batch Normalization after each convolutional layer. This addition helps stabilize learning, speeds up convergence, and improves generalization. While the layer progression and classification targets remain the same, the inclusion of batch normalization distinguishes this model by offering better training dynamics and potentially higher performance (30).

3.4.3 VGG16

The third model utilizes VGG16, a well-known deep CNN architecture pre-trained on the ImageNet dataset, as a feature extractor (31). Unlike the previous custom models, VGG16's convolutional layers are frozen to retain learned features, reducing training time and preventing overfitting on small datasets. On top of the frozen base, custom classification layers are added: a global average pooling layer to reduce feature maps, a dense layer with ReLU activation, a dropout layer for regularization, and a softmax output layer to classify MRI images into four tumor categories. This transfer learning approach combines the power of a proven model with task-specific tuning for improved accuracy and generalization.

3.4.4 ResNeT

The fourth is a ResNet model that integrates residual connections for more efficient learning, especially in deeper networks (32). It starts with a convolutional layer followed by max-pooling, similar to previous models. The main distinction in this model is the use of residual blocks, which include two convolutional layers per block. The shortcut connections are added to the output of these blocks, enabling the model to bypass specific layers and help mitigate the vanishing gradient issue. In the second block, a 1×1 convolution is used to match the output dimensions of the shortcut. The rest of the architecture follows the same structure, with global average pooling, a dense layer, and a softmax output for classification. The model is optimized using Adam with a learning rate of 0.0001 and uses categorical cross-entropy for loss.

3.4.5 Hybrid VGG16-CNN

The Hybrid CNN + VGG16 model integrates a pre-trained VGG16 model for feature extraction with a custom CNN designed to learn additional task-specific features (33). The VGG16 model, with its convolutional layers frozen, leverages the pre-learned features from the ImageNet dataset without any further updates during training. A Global Average Pooling layer processes its output to create a more compact representation of the features. The custom CNN learns additional features directly relevant to tumor classification. This CNN includes several convolutional layers followed by max-pooling layers to reduce the spatial dimensions of the feature maps. The resulting output is flattened and passed through a fully connected layer, with ReLU activation and a dropout layer for regularization. The features from both models are merged using the concatenate operation, followed by another fully connected layer with ReLU activation and a dropout layer. The final output layer uses softmax activation to produce a probability distribution over the four tumor categories: glioma tumor, meningioma tumor, no tumor, and pituitary tumor. The model is compiled with the Adam optimizer and categorical cross-entropy as the loss function, which is suitable for multi-class classification. It is trained for 50 epochs with a batch size of 32, using training and validation data.

3.5 Fine tuning models

The previously trained Hybrid CNN + VGG16 model was fine-tuned for the second experimentation phase using the Augmented Alzheimer's MRI dataset. This dataset includes four categories: Mild Demented, Moderate Demented, Non Demented, and Very Mild Demented. The hybrid model combines the VGG16 architecture, which was pre-trained on the ImageNet dataset and used as a frozen feature extractor, with a custom CNN trained to extract domain-specific features. To adapt the model for this new classification task, the final dense layer was replaced to match the four output classes. While the VGG16 layers remained frozen to retain their generalized feature representations, the custom CNN layers were set as trainable to learn patterns specific to Alzheimer's stages. Additionally, dropout and L2 regularization were applied to mitigate overfitting. The model was compiled using the Adam optimizer with a learning rate of 0.0005 and trained using augmented image data. To further validate our hybrid CNN+VGG16 model, we evaluated its performance on a third publicly available MRI brain tumor dataset consisting of four categories: glioma, meningioma, pituitary, and no tumor. The model architecture and training methodology remained consistent with previous experiments, incorporating dual-input feature fusion and transfer learning. After minor data augmentation and preprocessing adjustments, the model was retrained using a two-input pipeline and evaluated on stratified splits. The model demonstrated strong generalization to this new dataset, maintaining high accuracy across all classes. These results further reinforce the robustness and adaptability of our proposed hybrid model to varying data distributions.

To evaluate the generalization performance of the proposed Hybrid CNN + VGG16 model without relying on data augmentation, we conducted additional experiments on the unaltered original Alzheimer's MRI dataset. While the model architecture and configuration remained consistent, the training set consisted solely of original images, with no synthetic augmentation applied. The output layer was modified to match the four-class structure of this dataset. Only the custom CNN layers were updated during fine-tuning, while the VGG16 backbone remained frozen. The training used the same optimizer (Adam) and loss function (categorical cross-entropy) as in the augmented experiments. This experiment provides insight into how well the model performs in a more constrained, real-world scenario.

4 Experimental analysis and results

In this section, the accuracy, precision, recall, and F1 scores are used to assess the performance of the models. More specifically, it describes systematic experimental outcomes. This subsection defines all performance measurements, such as accuracy, precision,

recall, and F1-score and indicates how these measurements must be used.

The number of correctly classified instances (TP + TN) is the total number of instances of the data set. By applying Equation 5, we can calculate this value:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

It is the ratio of the number of times the model accurately predicted a product to the total number of times it has predicted it positively. Applying Equation 6 in this way will provide this result:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

The ratio of positive predictions to the data's actual number of positive instances. It reflects the model's ability to capture all positive instances. Use Equation 7 in the following manner to find this value:

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

The harmonic mean of precision and recall provides a single metric to balance both. It is beneficial when an imbalance between classes is calculated using Equation 8.

$$F1 - score = 2 \times \frac{Precision + Recall}{Precision + Recall} \tag{8}$$

Figure 3a illustrates a model's training and validation accuracy over 45 epochs. The training accuracy commences at ~0.460 at the 0_{th} epoch and shows a steady upward trajectory, reaching about 0.800 by the 40_{th} epoch. Similarly, the validation accuracy begins at around 0.500 and follows a comparable increasing trend, surpassing the training accuracy at several points and culminating at ~0.805 at the final epoch. Figure 3b presents the corresponding loss values for training and validation over the same number of epochs. The training loss starts at around 1.17 at the 0_{th} epoch and declines progressively, reaching about 0.47 by the 40_{th} epoch. The validation loss follows a similar pattern, beginning near 1.02 and steadily decreasing to ~0.50 at the final epoch.

Figure 4a illustrates a model's training and validation accuracy over 17 epochs. The training accuracy begins at ~0.790 at the 0_{th} epoch and exhibits a consistent upward trend, reaching about 0.955 by the 17_{th} epoch. The validation accuracy initiates at around 0.880 and fluctuates slightly throughout the training process, peaking around the 14_{th} epoch near 0.935 before ending at ~0.920. Figure 4b presents the corresponding training and validation loss across the same epoch range. The training loss starts relatively high at ~0.61 in the 0_{th} epoch and shows a steady decline, reaching around 0.13 by the 17_{th} epoch. The validation loss follows a more irregular pattern, beginning near 0.40, spiking intermittently, and settling at around 0.33 in the final epoch.

Figure 5a shows a model's training and validation accuracy over 15 epochs. The training accuracy starts at ~0.310 at the 0_{th} epoch and rises steadily throughout the training process, reaching about 0.905 by the 15_{th} epoch. The validation accuracy initially starts higher at around 0.390, increases with some fluctuations, and peaks around 0.890 near the 11_{th} epoch before settling slightly lower at ~0.875 by the final epoch. Figure 5b presents the corresponding loss values over the same epoch range. The training loss begins at a relatively high value of around 1.38 at the 0_{th} epoch and decreases consistently, dropping to ~0.28 by the 15_{th} epoch. The validation loss starts at about 1.22 and fluctuates more than the training loss, reaching a peak around 1.48 at the 3_{rd} epoch but then follows a general downward trend to around 0.40 at the final epoch.

Table 1 presents the classification performance across three datasets: Brain Tumor MRI, Augmented Alzheimer MRI, and a third tumor classification dataset. The Brain Tumor MRI dataset includes four tumor classes: glioma_tumor, meningioma_tumor, no_tumor, and pituitary_tumor. The model achieves the highest F1-score of 0.98 for the pituitary_tumor class, with corresponding precision and recall values of 0.97 and 0.99, respectively. The

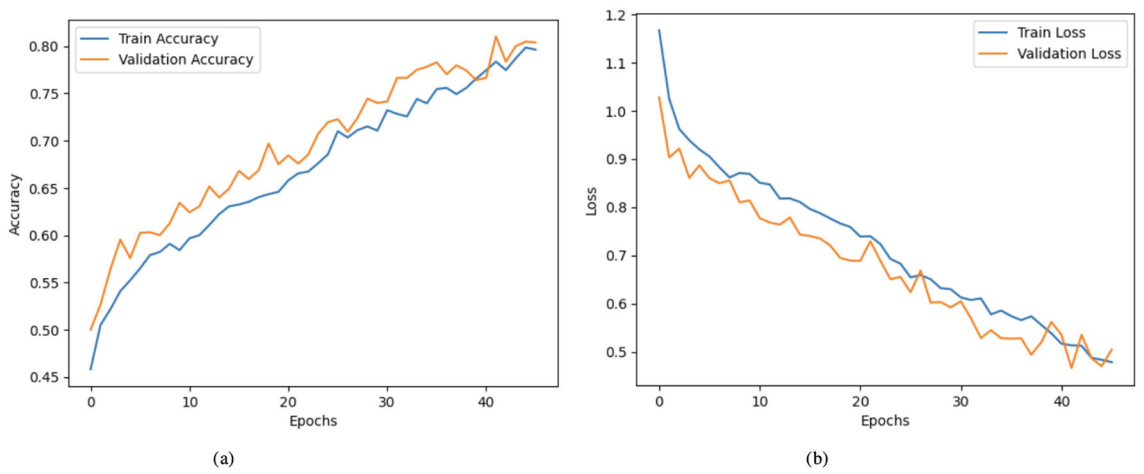


FIGURE 3
Graphical representation of hybrid CNN-VGG16 model with XAI on second dataset. **(a)** Accuracy graph. **(b)** Loss graph.

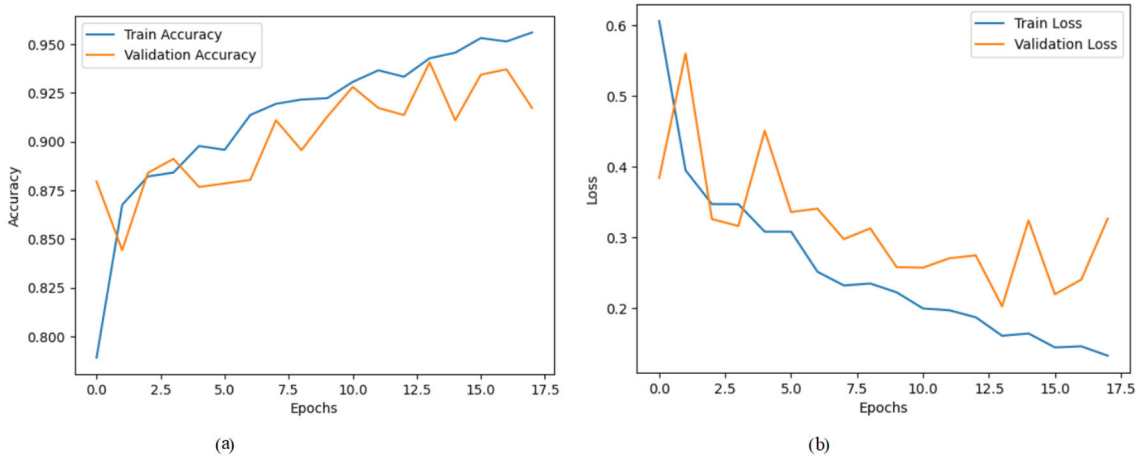


FIGURE 4
Graphical representation of hybrid CNN-VGG16 model with transfer learning on third dataset. **(a)** Accuracy graph. **(b)** Loss graph.

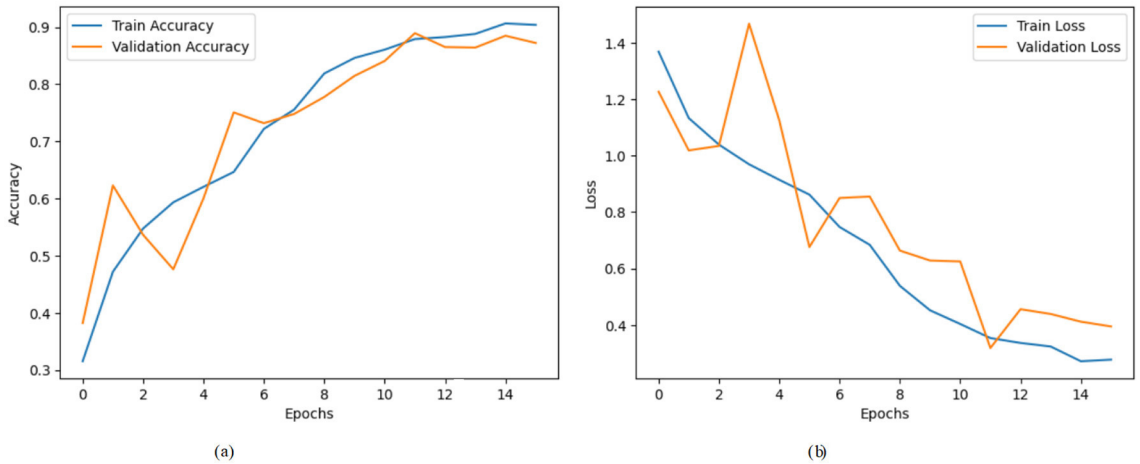


FIGURE 5
Graphical representation of hybrid CNN-VGG16 model with XAI on third dataset. **(a)** Accuracy graph. **(b)** Loss graph.

glioma_tumor class also performs strongly with all three metrics: precision, recall, and F1-score at 0.96. The no_tumor class has a slightly lower recall of 0.87, contributing to an F1-score of 0.90. Overall, the model demonstrates high classification effectiveness with a total accuracy of 94%. In the Augmented Alzheimer MRI dataset. This dataset includes four classes: MildDemented, ModerateDemented, NonDemented, and VeryMildDemented. Among these, the NonDemented class achieves the highest F1-score of 0.87, driven by a strong recall of 0.89. Although the ModerateDemented class attains a perfect precision of 1.00, its low recall of 0.54 results in a moderate F1-score of 0.70, indicating potential challenges in correctly identifying all instances of this class. The overall model accuracy for this dataset is 81%, which suggests reasonable but improvable classification performance. The third dataset consists of the following classes: glioma, meningioma, no tumor, and pituitary. The tumor class performs the best with an F1-score of 0.97, bolstered by a precision of 0.96 and a recall of 0.98. The pituitary class also achieves high recall (0.99), although its precision is relatively lower at 0.88, yielding an F1-score of 0.93. The overall model accuracy stands at 93%, indicating a strong performance across multiple tumor categories.

For multi-class classification, SHAP values were calculated per class and reshaped for visualization. Summary plots were generated to identify globally important regions across all samples.

4.1 Model explainability using SHAP

To better understand how our Hybrid CNN + VGG16 model makes decisions, we used SHapley Additive explanations (SHAP). This method explains model predictions by highlighting which parts of the input image contribute most to the final output. Since our model has a dual-input architecture with the same MRI image passing through two branches for enhanced

TABLE 1 Classification metrics across three datasets.

Dataset	Class	Precision	Recall	F1-Score
Brain tumor MRI	Glioma_tumor	0.96	0.96	0.96
	Meningioma_tumor	0.91	0.91	0.91
	No_tumor	0.92	0.87	0.90
	Pituitary_tumor	0.97	0.99	0.98
	Accuracy	94%		
Augmented Alzheimer MRI	MildDemented	0.81	0.64	0.72
	ModerateDemented	1.00	0.54	0.70
	NonDemented	0.84	0.89	0.87
	VeryMildDemented	0.76	0.77	0.76
	Accuracy	81%		
Third dataset	Glioma	0.96	0.89	0.92
	Meningioma	0.92	0.83	0.87
	Notumor	0.96	0.98	0.97
	Pituitary	0.88	0.99	0.93
	Accuracy	93%		

feature learning, we adapted SHAP's DeepExplainer to handle this structure accordingly. We selected a sample batch from the validation set and computed SHAP values for both inputs. Summary plots were generated to identify which features (or pixel regions) are typically important over the dataset and image plots for each pixel that mattered in discriminating a given prediction from the others. This allowed these visualizations to show that no matter the input, the model always attends to brain regions involved in Alzheimer's disease. This provides valuable guidance for building trust in AI-based clinical tools, and the model is strengthened in terms of interpretability and communicates that it is learning meaningful patterns.

In order to increase the interpretability of the hybrid CNN+VGG16 model trained for the brain tumor classification task, we combined the SHapley Additive exPlanations (SHAP) technique that allows explainable AI. The model has a multi-input structure (perhaps there is a better term for this), so SHAP's DeepExplainer was used on batches of validation images to compute pixel-based contributions for each prediction. The SHAP values revealed which areas of the MRI scans were the most important in allowing the model to decide. We would find through summary plots that the model consistently locked in on key tumor areas irrespective of the different categories, thus showing that it accurately emphasized those features. However, this transparency not only supports the credibility of the model but, additionally, is of the essence for the reliability of AI-based diagnostics in other medical applications.

4.1.1 SHAP summary plots of second dataset

SHAP values were successfully computed for a multi-input model using DeepExplainer, with each input consisting of 32 RGB images (128 × 128). The resulting SHAP tensors had a shape of (32, 128, 128, 3, 4), indicating class-specific attributions. Separate summary plots were generated for the four classes across both inputs, highlighting important spatial regions contributing to the model's predictions.

Figure 6a presents a SHAP summary plot that visualizes the influence of Features labeled numerically from 20551 to 36331. The *x*-axis represents SHAP values, where positive values indicate features that push the prediction higher, and negative values indicate the opposite. Color gradients reflect feature magnitudes. Pink denotes high values, and blue denotes low values. In this case, certain features like 20551 and 28222 exhibit a more pronounced impact on the model's predictions, evidenced by their wider spread along the SHAP value axis compared to others. On the other hand, features such as 20548 and 20549 show minimal impact, clustering closer to zero. Figure 6b presents a SHAP summary plot that illustrates the influence of features from "Feature 35950" to "Feature 33595" on the model's output. Notably, 35950 and 35184 are significantly influenced by their pronounced spread along the SHAP value axis, suggesting they contribute meaningfully to the model's output. In contrast, features like 21767 and 35569 cluster closer to zero, indicating a minimal effect on the predictive performance. Figure 6c presents a SHAP summary plot that illustrates the features that influence the model's output, ranging from "Feature 20158" to "Feature 27859." Notably, features such as "Feature 20158" and "Feature 34381" significantly impact the model's predictions, as indicated by the broader distribution of SHAP values. This suggests that variations in these features can lead to more pronounced effects on the predictions. In contrast, features like "Feature 34348" and "Feature 18958" cluster closer to the zero line, indicating a lesser impact on model predictions. This clustering reveals that changes in these features do not significantly influence the overall model output. Figure 6d presents a SHAP summary plot that visualizes the influence of features ranging from "Feature 20158" to "Feature 24772" on the model's output. For instance, Features 20158 and 33604 exhibit strong positive contributions when their values are high, whereas Features

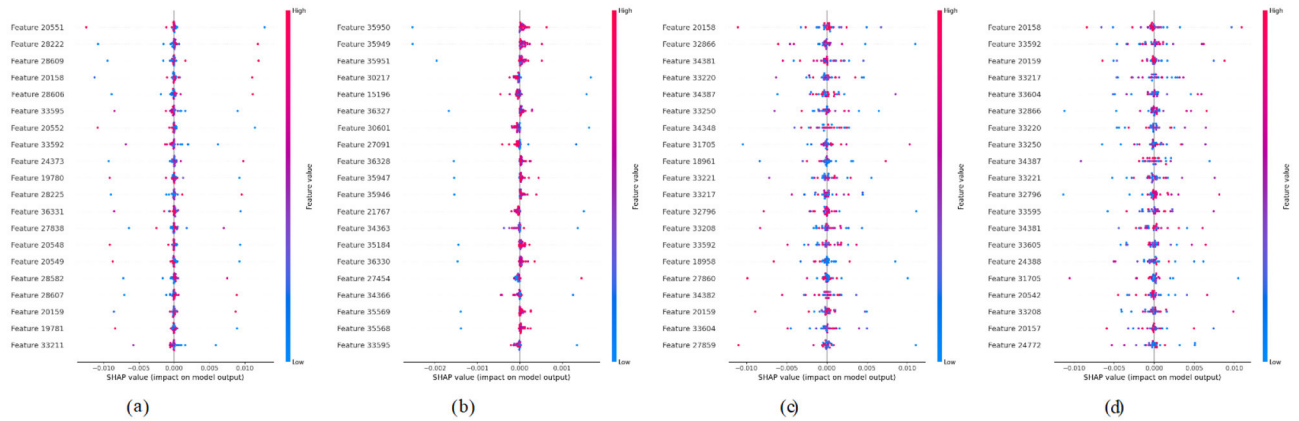


FIGURE 6
SHAP summary plots for Classes 0 through 3. **(a)** Class 0. **(b)** Class 1. **(c)** Class 2. **(d)** Class 3.

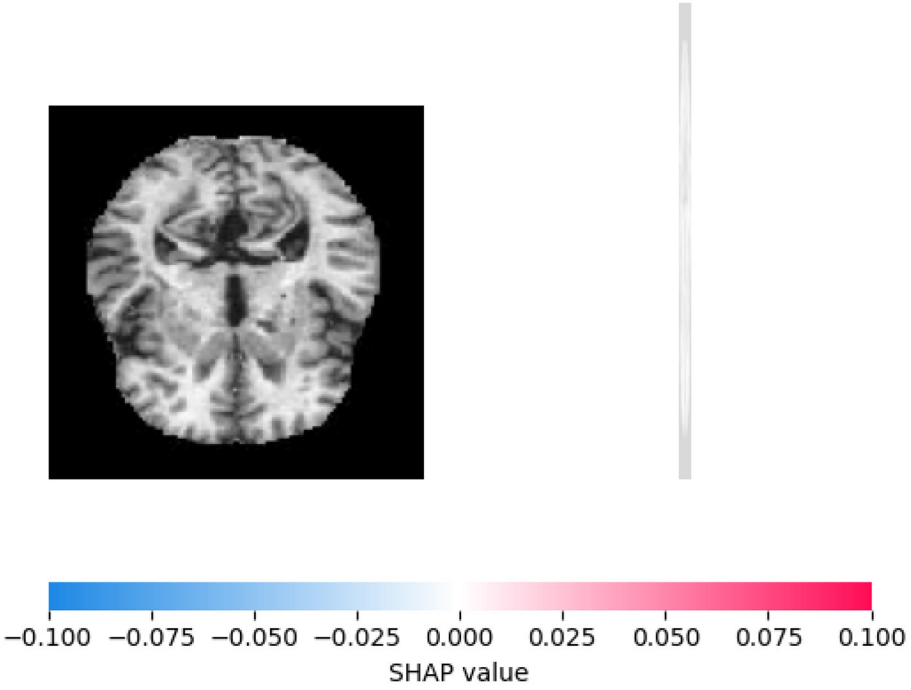


FIGURE 7
Visualizing SHAP explanation for sample 0, class 0.

33250 and 24772 predominantly display negative SHAP values, indicating a suppressive effect on predictions. This plot highlights key features that significantly shape model behavior based on their value ranges.

Figure 7 displays a cross-sectional brain image alongside a SHAP value color scale. The grayscale brain scan highlights structural features, while the adjacent gradient from blue (–0.1, negative contribution) to red (+0.1, positive contribution) represents each region's influence on model predictions. This integration aids in interpreting how specific brain areas affect analytical outcomes, linking neuroimaging data to model behavior.

4.1.2 SHAP summary plots of third dataset

Figure 8a presents a SHAP summary plot that illustrates the impact of various features, ranging from "Feature 21277" to "Feature 12959," on the model's predictions. The visualization indicates that certain features, such as "Feature 21280" and "Feature 29056," significantly influence the model's output, as evidenced by their extensive spread along the SHAP value axis. In contrast, features like "Feature 21337" and "Feature 24373" demonstrate minimal impact, as their SHAP values cluster closer to zero. Figure 8b presents a SHAP summary plot visualizing the influence of various features, specifically labeled from "Feature 15520" to "Feature 21276," on the model's output. In this plot, features such as

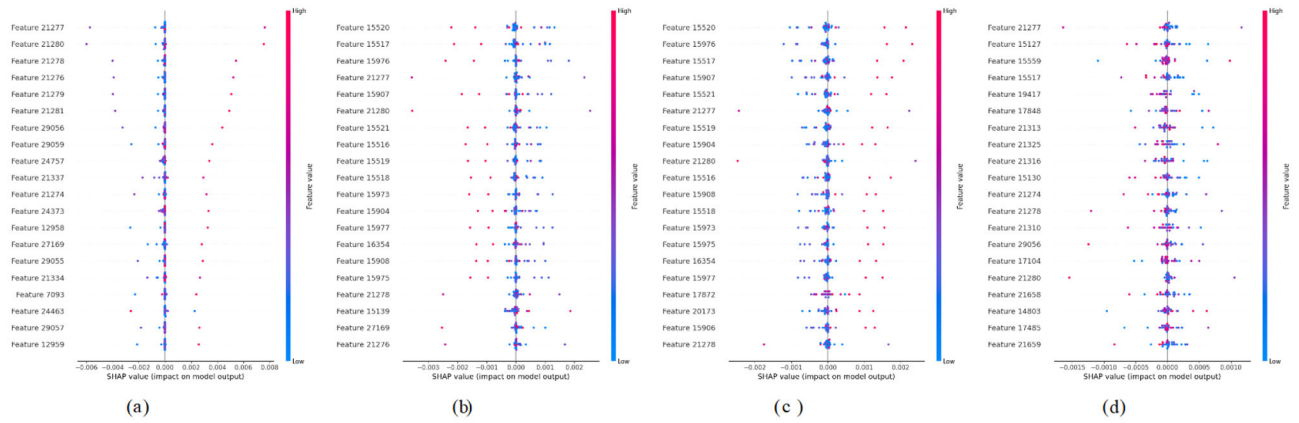


FIGURE 8
SHAP summary plots for Classes 0 to 3. **(a)** Class 0. **(b)** Class 1. **(c)** Class 2. **(d)** Class 3.

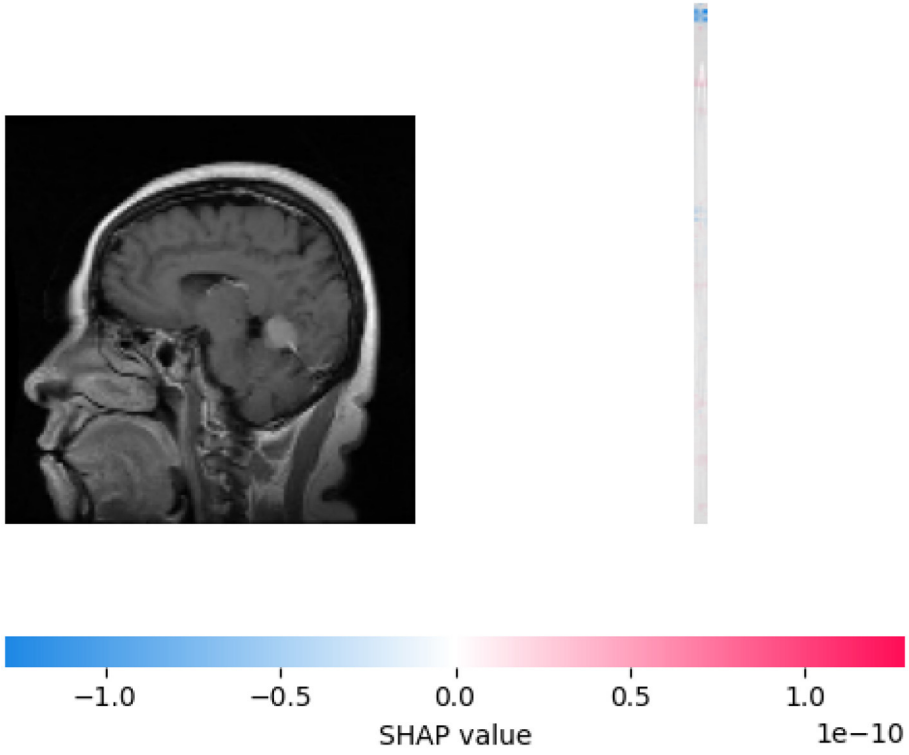


FIGURE 9
Visualizing SHAP explanation for sample 0, class 0.

"Feature 15976" and "Feature 15908" exhibit a significant influence, as indicated by their wider dispersion on the SHAP value axis. This means that these features contribute more substantially to the predicted outcomes when compared to others. Conversely, features like "Feature 15520" and "Feature 15139" cluster closer to zero, demonstrating minimal impact on the model's predictions. Figure 8c presents a SHAP summary plot that illustrates the influence of various features, specifically from "Feature 15520" to "Feature 21278," on the model's predictions. Certain features, such as "Feature 15520" and "Feature 15976," exhibit a more pronounced effect on the model's predictions, as evidenced by their greater dispersion along the SHAP value axis. This suggests that these features are critical in influencing the model's output. Conversely, features like "Feature 15518" and "Feature 15904" reveal a minimal impact, clustering closely to zero. This suggests that their contributions to the model's predictions are negligible compared to those of other features. Figure 8d presents a SHAP summary plot that represents the impact of various features on the model's predictions, focusing on features ranging from "Feature 21277" to "Feature 21659." For instance, features such as "Feature

21280" and "Feature 29056" significantly impact the predictions, as indicated by their wider distribution of SHAP values that extend toward both positive and negative extremes. Conversely, features like "Feature 17104" and "Feature 15130" exhibit minimal influence, clustering closer to the zero mark, which suggests that their effect on the model output is negligible.

Figure 9 combines a sagittal brain MRI image (left) with a SHAP value bar plot (right) to illustrate model interpretability in neuroimaging. The MRI highlights anatomical brain structures, while the SHAP plot uses a blue-to-red gradient to show each region's contribution to model predictions, with blue indicating a negative and red indicating a positive influence. SHAP values range from –1 to 1, capturing features' subtle and significant impacts. This integrated visualization aids in understanding how specific brain regions affect model outcomes, bridging neuroimaging with explainable AI.

4.2 Discussion

The proposed hybrid CNN-VGG16 framework addresses three key challenges in MRI-based neuroimaging diagnostics: limited labeled data, variability across datasets, and lack of interpretability in deep learning models. First, the use of transfer learning significantly mitigates the issue of data scarcity. By leveraging the pre-trained VGG16 architecture, the model benefits from rich feature representations learned from large-scale natural image datasets. This allows for effective feature extraction even with relatively small medical imaging datasets. The high classification accuracy achieved on the brain tumor dataset (94%) and the third dataset (93%) demonstrates the model's ability to generalize across similar pathological domains. Second, the sequential fine-tuning strategy across structurally distinct datasets of brain tumors and Alzheimer's and a third validation set demonstrates the framework's adaptability to different neuroimaging modalities. The model maintains a competitive performance of 81% on the augmented Alzheimer dataset despite its structural differences from the training domain. This highlights the framework's robustness and transferability, addressing the domain shift problem that often limits the practical deployment of deep learning models in medical diagnostics. Third, integrating SHAP-based Explainable AI resolves the critical issue of interpretability. By generating pixel-level explanations, the framework provides insight into which brain regions influence the model's predictions. This capability enhances clinical trust and offers potential support for diagnostic reasoning by aligning model attention with known anatomical and pathological patterns. The proposed approach combines performance and transparency, offering a concrete step toward clinically viable AI systems. It outperforms traditional single-dataset training and black-box models by effectively resolving challenges related to data diversity, cross-domain generalization, and explainability.

5 Conclusion

This paper demonstrated the effectiveness of transfer learning combined with XAI for classifying MRI images. SHAP values provide much insight into the decision-making path of the model, and the hybrid CNN-VGG16 model generalizes well over different datasets with high accuracy. In conclusion, this approach and its generalizations can be applied to other medical imaging tasks, possessing high performance and interpretability. This research has demonstrated the effectiveness of a hybrid CNN-VGG16 model, utilizing transfer learning in conjunction with XAI techniques, for MRI image classification. The high accuracy of the model across multiple datasets demonstrates that it is robust and easily adaptable in distinguishing between different neurological diseases, including brain tumors and Alzheimer's disease. While the model shows strong performance, it has certain limitations. The reliance on a limited number of public datasets may restrict its generalizability to real-world clinical scenarios. Additionally, the SHAP-based interpretability comes with a high computational cost, which may challenge real-time deployment. Future work will expand dataset diversity, incorporate 3D volumetric data, optimize model architecture for clinical deployment, and explore alternative interpretability methods. This research lays a solid foundation for developing high-performing, interpretable AI tools to support medical decision-making and improve patient outcomes. This work also lays the groundwork for future research to refine the model further and apply it to other medical imaging applications, ultimately leading to enhanced patient outcomes.

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

SA: Validation, Project administration, Conceptualization, Data curation, Writing – review & editing, Methodology, Investigation, Writing – original draft, Visualization, Formal analysis. SO: Writing – original draft, Project administration, Methodology, Resources, Investigation, Validation, Writing – review & editing, Funding acquisition. TN: Writing – original draft, Project administration, Resources, Validation, Writing – review & editing, Funding acquisition. MA: Formal analysis, Software, Methodology, Writing – review & editing, Investigation, Writing – original draft. JB: Writing – review & editing, Writing – original draft, Investigation, Project administration, Methodology. AAlm: Writing – review & editing, Investigation, Supervision, Methodology, Writing – original draft, Data curation. AAlH: Methodology, Writing – original draft, Visualization, Writing – review & editing.

Funding

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

Acknowledgments

The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large Research Project under grant number RGP.2/275/46 and the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through the project number "NBU-FFR-2025-2443-01."

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. Abid MA, Munir K. A systematic review on deep learning implementation in brain tumor segmentation, classification and prediction. *Multimed Tools Appl*. (2025) 1–40. doi: 10.1007/s11042-025-20706-4

2. Prajapati YN, Sonker SK, Agrawal PP, Jain J, Kumar M, Kumar V. Brain tumor detection and classification using deep learning on MRI images. In: *2025 2nd International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*. Ghaziabad: IEEE (2025). p. 131–5. doi: 10.1109/CICTN64563.2025.10932392

3. Dorfner FJ, Patel JB, Kalpathy-Cramer J, Gerstner ER, Bridge CP. A review of deep learning for brain tumor analysis in MRI. *NPJ Precis Oncol*. (2025) 9:2. doi: 10.1038/s41698-024-00789-2

4. Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, et al. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro Oncol*. (2021) 23:1231–51. doi: 10.1093/neuonc/noab106

5. Bhattacharjee R, Thakran S. Conventional and advanced magnetic resonance imaging methods. In: Saxena S, Paul S, editors. *High-Performance Medical Image Processing*. Burlington, ON: Apple Academic Press (2022). p. 131–44. doi: 10.1201/9781003190011-6

6. Serai SD. Basics of magnetic resonance imaging and quantitative parameters T1, T2, T2*, T1rho and diffusion-weighted imaging. *Pediatr Radiol*. (2022) 52:217–27. doi: 10.1007/s00247-021-05042-7

7. Silfina RO, Indrati R, Utami L. The role T1-weighted fluid attenuated inversion recovery (FLAIR) post contrast enhancement to improve image quality on MRI brain. *J Phys Conf Ser*. (2021) 1943:012052. doi: 10.1088/1742-6596/1943/1/012052

8. Song D, Fan G, Chang M. Research progress on glioma microenvironment and invasiveness utilizing advanced multi-parametric quantitative MRI. *Cancers*. (2024) 17:74. doi: 10.3390/cancers17010074

9. Ahn SJ, Taoka T, Moon WJ, Naganawa S. Contrast-enhanced fluid-attenuated inversion recovery in neuroimaging: a narrative review on clinical applications and technical advances. *J Magn Reson Imaging*. (2022) 56:341–53. doi: 10.1002/jmri.28117

10. Tawfeeq LA, Hussein SS, Altyar SS. Leveraging transfer learning in deep learning models for enhanced early detection of Alzheimer's disease from MRI scans. *J Inf Hiding Multimed Signal Process*. (2025). Available online at: https://bit.kuas.edu.tw/2025/vol16/N1/02.JIHMSP-241005.pdf

11. Bibi N, Courtney J, McGuinness K. Enhancing brain disease diagnosis with XAI: a review of recent studies. *ACM Trans Comput Healthc*. (2025) 6:1–35. doi: 10.1145/3709152

12. Nagarajan I, Lakshmi Priya G. A comprehensive review on early detection of Alzheimer's disease using various deep learning techniques. *Front Comput Sci*. (2025) 6:1404494. doi: 10.3389/fcomp.2024.1404494

13. Kaur I, Sachdeva R. Prediction models for early detection of Alzheimer: recent trends and future prospects. *Arch Comput Methods Eng*. (2025) 1–28. doi: 10.1007/s11831-025-10246-3

14. Viswan V, Shaffi N, Mahmud M, Subramanian K, Hajamohideen F. Explainable artificial intelligence in Alzheimer's disease classification: a systematic review. *Cognit Comput*. (2024) 16:1–44. doi: 10.1007/s12559-023-10192-x

15. Tounsi M, Aram E, Azar AT, Al-Khayyat A, Ibraheem IK. A comprehensive review on biomedical image classification using deep learning models. *Eng Technol Appl Sci Res*. (2025) 15:19538–45. doi: 10.48084/etasr.8728

16. Asmita MP. From black box AI to XAI in neuro-oncology: a survey on MRI-based tumor detection Discover. *Artif Intell*. (2025) 5:1–21. doi: 10.1007/s44163-025-00247-3

17. Tuncer T, Dogan S, Subasi A. FiboNeXt: investigations for Alzheimer's disease detection using MRI. *Biomed Signal Process Control*. (2025) 103:107422. doi: 10.1016/j.bspc.2024.107422

18. Lasagni L, Ciccarone A, Guerrini R, Lenge M, D'incerti L. Focal cortical dysplasia type II detection using cross modality transfer learning and grad-CAM in 3D-CNNs for MRI analysis. *arXiv*. (2025) [Preprint]. arXiv:2504.07775. doi: 10.48550/arXiv.2504.07775

19. Bhaskaran SB, Datta R. Explainability of brain tumor classification model based on inceptionv3 using XAI tools. *J Flow Vis Image Process*. (2025) 32. doi: 10.1615/JFlowVisImageProc.2024054026

20. Tonni SI, Sheakh MA, Tahosin MS, Hasan MZ, Shuva TF, Bhuiyan T, et al. A hybrid transfer learning framework for brain tumor diagnosis. *Adv Intell Syst*. (2025) 7:2400495. doi: 10.1002/aisy.202400495

21. Nahiduzzaman M, Abdulrazak LF, Kibria HB, Khandakar A, Ayari MA, Ahamed MF, et al. A hybrid explainable model based on advanced machine learning and deep learning models for classifying brain tumors using MRI images. *Sci Rep*. (2025) 15:1649. doi: 10.1038/s41598-025-85874-7

22. Vanaja T, Shanmugavadivel K, Subramanian M, Kanimozhiselvi C. Advancing Alzheimer's detection: integrative approaches in MRI analysis with traditional and deep learning models. *Neural Comput Appl*. (2025) 1–20. doi: 10.1007/s00521-025-10993-1

23. Joshi AL, Veetil IK, Premjith B, Sowmya V, Gopalakrishnan E, Vinayakumar R. Performance analysis of big transfer models on biomedical image classification. In: *Analytics Modeling in Reliability and Machine Learning and Its Applications*. Cham: Springer (2025). p. 141–60. doi: 10.1007/978-3-031-72636-1_7

24. Bin Shabbir Mugdha S, Uddin M. NeuroSight: a deep-learning integrated efficient approach to brain tumor detection. *Eng Rep*. (2025) 7:e13100. doi: 10.1002/eng2.13100

25. Khedgaonkar RS, Badhe SS, Bhoyar D, Mohod S. Design of an efficient model for brain MRI image classification using graph neural networks. In: *AIP Conference Proceedings, Vol. 3255*. Melville, NY: AIP Publishing (2025). doi: 10.1063/5.0254825

26. Ilani MA, Shi D, Banad YM. T1-weighted MRI-based brain tumor classification using hybrid deep learning models. *Sci Rep*. (2025) 15:7010. doi: 10.1038/s41598-025-92020-w

27. Rasool N, Bhat JI, Aoun NB, Alharthi A, Wani NA, Chopra V, et al. ResMHA-Net: enhancing glioma segmentation and survival prediction using a novel deep learning framework. *Comput Mater Contin*. (2024) 81:885–909. doi: 10.32604/cmc.2024.055900

28. Gasmi K, Ben Aoun N, Alsalem K, Ltaifa IB, Alrashdi I, Ammar LB, et al. Enhanced brain tumor diagnosis using combined deep learning models and weight selection technique. *Front Neuroinform*. (2024) 18:1444650. doi: 10.3389/fninf.2024.1444650

29. Ayeni J. Convolutional neural network (CNN): the architecture and applications. *Appl J Phys Sci*. (2022) 4:42–50. doi: 10.31248/AJPS2022.085

30. Borgalli MRA, Surve S. Deep learning for facial emotion recognition using custom CNN architecture. *J Phys Conf Ser*. (2022) 2236:012004. doi: 10.1088/1742-6596/2236/1/012004

31. 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), Vol. 1*. Bengaluru: IEEE (2021). p. 96–9. doi: 10.1109/CENTCON52345.2021.9687944

32. Xu W, Fu YL, Zhu D. ResNet and its application to medical image processing: Research progress and challenges. *Comput Methods Programs Biomed*. (2023) 240:107660. doi: 10.1016/j.cmpb.2023.107660

33. Belaid ON, Loudini M. Classification of brain tumor by combination of pre-trained vgg16 CNN. *J Inf Technol Manag*. (2020) 12:13–25. doi: 10.22059/jitm.2020.75788

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 03 July 2025
DOI 10.3389/fnins.2025.1618514

Check for updates

OPEN ACCESS

EDITED BY
Deepti Deshwal,
Maharaja Surajmal Institute of Technology, India

REVIEWED BY
Arun Kumar Sunaniya,
National Institute of Technology, Silchar, India
Anjana Subba,
National Institute of Technology, Silchar, India
Jane Rubel Angelina Jeyaraj,
Kalasalingam University, India

*CORRESPONDENCE
Wenna Chen
✉ chenwenna0408@163.com
Ganqin Du
✉ dgq99@163.com

RECEIVED 26 April 2025
ACCEPTED 12 June 2025
PUBLISHED 03 July 2025

CITATION
Chen W, Tan X, Zhang J, Du G, Fu Q and Jiang H (2025) MLG: a mixed local and global model for brain tumor classification. *Front. Neurosci.* 19:1618514. doi: 10.3389/fnins.2025.1618514

COPYRIGHT
© 2025 Chen, Tan, Zhang, Du, Fu and Jiang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

MLG: a mixed local and global model for brain tumor classification

Wenna Chen[1]*, Xinghua Tan[2], Jincan Zhang[2], Ganqin Du[1]*, Qizhi Fu[1] and Hongwei Jiang[1]

[1]The First Affiliated Hospital, and College of Clinical Medicine of Henan University of Science and Technology, Luoyang, China, [2]College of Information Engineering, Henan University of Science and Technology, Luoyang, China

Introduction: Brain tumors seriously endanger human health. Therefore, accurately identifying the types of brain tumors and adopting corresponding treatment methods is of vital importance, which is of great significance for saving patients' lives. The use of computer-aided systems (CAD) for the differentiation of brain tumors has proved to be a reliable scheme.

Methods: In this study, a highly accurate Mixed Local and Global (MLG) model for brain tumor classification is proposed. Compared to prior approaches, the MLG model achieves effective integration of local and global features by employing a gated attention mechanism. The MLG model employs Convolutional Neural Networks (CNNs) to extract local features from images and utilizes the Transformer to capture global characteristics. This comprehensive scheme renders the MLG model highly proficient in the task of brain tumor classification. Specifically, the MLG model is primarily composed of the REMA Block and the Biformer Block, which are fused through a gated attention mechanism. The REMA Block serves to extract local features, effectively preventing information loss and enhancing feature expressiveness. Conversely, the Biformer Block is responsible for extracting global features, adaptively focusing on relevant sets of key tokens based on query positions, thereby minimizing attention to irrelevant information and further boosting model performance. The integration of features extracted by the REMA Block and the Biformer Block through the gated attention mechanism further enhances the representation ability of the features.

Results: To validate the performance of the MLG model, two publicly available datasets, namely the Chen and Kaggle datasets, were utilized for testing. Experimental results revealed that the MLG model achieved accuracies of 99.02% and 97.24% on the Chen and Kaggle datasets, respectively, surpassing other state-of-the-art models. This result fully demonstrates the effectiveness and superiority of the MLG model in the task of brain tumor classification.

KEYWORDS

classification of brain tumor, CNN, transformer, feature fusion, gated attention mechanism

1 Introduction

Brain diseases most commonly manifest as brain tumors, which represent a severe health threat to the human body and necessitate early diagnosis and treatment (Lyu et al., 2024; Akter et al., 2024; Liu et al., 2023). The classification of brain tumors constitutes a significant area of research in medical imaging and artificial intelligence. Classification of brain tumors using Magnetic Resonance Imaging (MRI) is the main technique (Li and Zhou, 2025). This process is critical for accurate diagnosis, treatment planning, and prognosis assessment. Recently, Computer-Aided Detection and Diagnosis (CAD) systems have played a pivotal role in assisting medical professionals with the detection and classification of brain tumors. Traditional manual methods of brain tumor classification rely heavily on experienced specialists and are often time-consuming, labor-intensive, and inefficient (Sharma et al., 2024; Zhou et al., 2024). To address this issue, extensive research has been conducted into automatic classification techniques that can classify brain tumors from MRI, employing CAD technology for tumor classification from MRI, which exhibits high reliability due to its high accuracy.

Traditional machine learning often relies on manually designed features, which places high demands on the user's domain knowledge and experience. The selection and construction of features are complex and time-consuming, having a crucial impact on model performance. When faced with complex, high-dimensional, or nonlinear problems, the generalization ability of traditional machine learning algorithms may be limited (Kaur and Mahajan, 2025). More crucially, when confronted with new, unseen data, their predictive performance may decline, affecting their practical utility (Mehnatkesh et al., 2023; Pandiselvi and Maheswaran, 2019). In contrast, deep learning possesses stronger data representation capabilities, able to automatically learn high-level abstract representations of data, significantly enhancing the performance and effectiveness of machine learning. Deep learning models are not only highly complex but also capable of handling more complex tasks and larger datasets. Consequently, deep learning has found widespread application in the field of medical imaging, providing powerful support for disease diagnosis and treatment (Kshatri and Singh, 2023; Mazurowski et al., 2023; Mukadam and Patil, 2024; Yu et al., 2022).

Convolutional Neural Networks (CNNs), as a type of deep learning algorithm, have demonstrated remarkable prowess in the field of image processing, thanks to their unique advantages. The CNNs not only accept input images, but also adeptly assign varying degrees of importance to different elements or objects within those images through learnable weights and biases, enabling effective differentiation among them. Compared to other classification algorithms, the CNNs significantly reduce the need for preprocessing, greatly enhancing ease of use. In earlier image processing, filters were typically manually designed. However, CNNs can automatically learn these filters or features during training. Consequently, CNNs have seen widespread application in fields such as medical image analysis. Cao et al. (2024) introduced a Multi-branch Spectral Channel Attention Network (MbsCANet) for breast cancer classification. By extracting features in the frequency domain and applying attention mechanisms to the backbone network, MbsCANet achieves more precise feature extraction and classification, thereby not only improving classification accuracy but also providing robust support for early diagnosis and treatment of breast cancer. Regarding retinal disease classification, Peng et al. (2024) proposed a multi-scale-denoising residual convolutional network (MS-DRCN) model. This model integrates the strengths of Deep Residual Network (ResNet) along with multiscale processing and feature fusion techniques. Aimed at enhancing the accuracy and robustness of Optical Coherence Tomography (OCT) image classification, MS-DRCN offers an effective tool for precise diagnosis of retinal diseases. Moreover, SkinLesNet, a deep learning model specifically designed for skin lesion classification, is built upon a CNN architecture that has undergone meticulous design and optimization (Azeem et al., 2024). Through a series of CNNs, it progressively extracts image features, enabling in-depth understanding and analysis of lesion images. This structure enables the model to precisely capture subtle differences and key features within the images, significantly boosting classification accuracy and reliability. As a result, it provides crucial assistance in the early detection and treatment of skin lesions.

The Transformer, an attention mechanism originating from the field of natural language processing, has demonstrated remarkable performance in computer vision. Its advantages over CNNs are particularly evident in handling long-distance dependencies and global contextual information in images (Liu et al., 2021b; Yan et al., 2023; Huang S. K. et al., 2024). Bofan Song et al. (Song et al., 2024) utilized Vision Transformer (ViT) and Swin Transformer (SwinT) for the classification of oral cancer images. In the literature (Huang L. et al., 2024), Swin-residual transformer (SRT), was proposed for thyroid ultrasound image classification. The SRT model introduces residual blocks and triplet loss into the SwinT structure, aiming to improve sensitivity to both global and local features of thyroid nodules and better identify subtle feature differences. Additionally, Chincholi and Koestler (2024) designed a model combining ViT and Detection Transformer architectures for glaucoma detection. As the application of Transformers in disease detection continues to grow, researchers have begun exploring the integration of CNNs with Transformers to simultaneously extract local and global features. For instance, Fang et al. (2024) employed CNNs to extract local features while utilizing ViT for global feature extraction, designing a deep integrated feature fusion module for feature aggregation. Yan et al. (2023) developed the Transformer based High Resolution Network (TransHRNet) for brain tumor segmentation. TransHRNet initially used CNNs as an encoder for image preprocessing, followed by feeding the extracted features from the CNNs into an Effective Transformer (EffTrans) module, and finally generating segmentation results through a CNNs decoder. Notably, EffTrans incorporates Group Linear Transformations (GLTs) with an expansion-reduction strategy and spatial-reduction attention (SRA) layers, significantly reducing the computational burden and memory consumption of the Transformer.

The classification of brain tumors poses a highly challenging task in computer vision. These tumors vary significantly in size, shape, and location within the brain, and their categorization depends not only on the characteristics of the lesion itself but also on the surrounding tissue environment (ThamilSelvi et al., 2025; Verma and Yadav, 2025). Furthermore, the diversity and

spatial distribution of brain tumors underscore the importance of utilizing both local and global features. In response to these challenges, the Mixed Local and Global (MLG) model is introduced. The uniqueness of the MLG model lies in its utilization of two advanced feature extraction methods. On one hand, Residual Efficient Multi-scale Attention (REMA) block is designed to extract local fine-grained features. On the other hand, the Bi-Level transformer (Biformer) block is used to capture the global context features. The REMA module integrates two layers of convolution and an Efficient Multi-scale Attention (EMA) component (Ouyang et al., 2023), which are interconnected through residual connections. This classical residual connection design ensures that gradients can propagate more effectively throughout the network during training, thereby mitigating gradient vanishing issues (He et al., 2016; Shafiq and Gu, 2022). Channel attention and spatial attention mechanisms have proven to be highly effective in generating more discriminative feature representations (Hu et al., 2018; Woo et al., 2018; Yu et al., 2023). In this block, EMA enhances both spatial and channel-wise features and achieves the ability to capture feature information across different scales by constructing parallel subnetwork structures operating at multiple resolutions. The core of Biformer is its Bi-Level Routing Attention (BRA), which facilitates dynamic and query-based content-aware sparse attention allocation while circumventing the high computational cost of full-space attention. Biformer realizes this pattern by introducing the Bi-Level Routing Attention mechanism, where it first prunes irrelevant key-value pairs at a coarse-grained region level, and subsequently conducts fine-grained token-to-token attention computations only within the selected candidate regions (Zhu et al., 2023). The integration of features from REMA and Biformer via gated attention mechanisms further refines these features, enhancing model performance. To validate the efficacy of the MLG model, two publicly available brain tumor datasets were utilized for experimental evaluation. Experimental results demonstrated that the proposed model outperforms other existing advanced models in terms of performance. In summary, the main contributions of this paper are as follows:

- Development of a brain tumor classification model that integrates both local and global features.
- The innovative application of the REMA module to extract local features and the use of Biformer for capturing global features, with both being effectively fused through a gated attention mechanism.
- Validation of the proposed model on two open datasets, achieving superior results compared to the current state-of-the-art performance.

2 Related work

The application of deep learning techniques in medical image analysis is becoming increasingly popular, particularly in the study of brain tumor classification, where it has demonstrated significant value. In recent years, research efforts on brain tumor classification tasks have continued to deepen, and these studies can be broadly categorized into two camps: one is the CNN-based approach, and the other is the emerging strategy based on the Transformer architecture.

2.1 CNN in brain tumor classification

The CNN has been widely used in brain tumor classification tasks. In the task of brain tumor classification, CNNs have been widely employed. Kang et al. (2021) adopted a transfer learning-based framework using a pre-trained deep CNN to extract deep features from MRI data. By fusing features obtained from different levels of the network and integrating them with multiple machine learning classifiers, this method achieved significant results. Alanazi et al. (2022) proposed a 22-layer CNN model, which was initially trained on a binary brain tumor dataset. Subsequently, with the help of transfer learning technique, the model weight was utilized for multi-class data, resulting in promising outcomes. Saurav et al. (2023) designed an Attention-Guided Convolutional Neural Network (AG-CNN) specifically tailored for brain tumor classification tasks. The network incorporates an internal channel attention module, which aids in focusing on processing image regions relevant to tumors, thereby facilitating effective feature extraction and classification. Alturki et al. (2023) proposed an optimization scheme for brain tumor classification performance. The CNNs were utilized to extract deep features from raw brain tumor MRI data and two classification algorithms including logistic regression (LR) and stochastic gradient descent (SGD) were incorporated into a voting ensemble classifier. By inputting these deep features into the ensemble classifier, the model achieved accurate classification of brain tumors. Hossain et al. (2023) conducted a study implementing transfer learning to investigate the performance of various models, including VGG16, InceptionV3, and ResNet50, inceptionResNetv2, Xception, for brain tumor classification. Ultimately, three best performing models were chosen to be used to construct an ensemble model, which was named IVX16. Sachdeva et al. (2024) evaluated multiple pre-trained models such as ResNet50, DenseNet121, EfficientNetB0, and EfficientNetV2L, et al., by incorporating Dropout layers, global average pooling layers, and tuning hyperparameters to enhance model performance. The results show that EfficientNetB0 model achieved a higher classification accuracy.

2.2 Transformer in brain tumor classification

Transformer has also been applied in brain tumor classification tasks. Ferdous et al. (2023) proposed a Linear Complexity Data-Efficient Image Transformer (LCDEiT) framework based on a teacher-student mechanism specifically designed for tumor classification from brain MRI images. In the teacher model component, gated pooling techniques were employed to optimize the feature extraction efficiency of CNNs. The pre-trained teacher model was able to extract crucial knowledge pertinent to the tumor classification task. On the other hand, the student model introduced an image transformer equipped with an external attention mechanism, which leveraged the knowledge acquired

from the teacher model for tumor classification in brain MRI. In paper, Asiri et al. (2024) proposed an innovative and robust method based on the SwinT architecture, aiming to improve the accuracy of brain tumor image classification. This method integrated complex preprocessing procedure, sophisticated feature extraction techniques, and a thorough classification system, enabling the SwinT model to effectively analyze and discriminate various types of brain tumors. Wang et al. (2024) employed a pre-trained ViT as the backbone for their brain tumor classification model, named as RanMerFormer. Additionally, to enhance the computational efficiency of the ViT backbone, a Token Merging Algorithm (TMA) was used. Instead of using a traditional linear classification head, Random Vector Functional Link (RVFL) networks were utilized. Poornam and Angelina (2024) proposed the ViT with Attention and Linear Transformation module (VITALT) for brain tumor detection and classification. VITALT primarily consists of a ViT, a Split bidirectional feature pyramid network (S-BiFPN), and a linear transformation module (LTM). ViT was used to capture global and local features, while S-BiFPN fusions the features extracted by ViT. The LTM enhanced the model's linear expressive ability. In paper (Şahin et al., 2024), the Bayesian Multi-Objective (BMO) optimization method was employed to optimize the hyperparameters of the ViT network in order to improve its performance in brain tumor classification tasks. Gade et al. (2024) proposed the Lite Swin Transformer (OLiST) model for brain tumor detection. This model combined the Lite Swin Transformer's ability to capture global features with the advantage of CNNs in extracting local features. By fusing the features extracted by both, the model leveraged the strengths of both approaches.

In summary, the use of CNNs and Transformers have been used in brain tumor classification tasks with excellent performance. CNNs have the advantage of extracting local features of images, while Transformers have the advantage of exploiting global features of images. Therefore, this paper innovatively introduces a hybrid model, MLG, which effectively integrates the respective strengths of CNNs and Transformers, thus significantly enhancing the performance of brain tumor classification tasks.

3 Materials and methods

In this section, the datasets used and the proposed model are described in detail.

3.1 Datasets and preprocessing

In this study, two widely used public datasets, namely the Chen dataset and the Kaggle dataset, were adopted. The Chen dataset, provided by Cheng et al. (2015), primarily focuses on three types of brain tumors: gliomas, meningiomas, and pituitary tumors. Comprising a total of 3,064 images, this dataset offers a rich resource for our in-depth research and analysis. On the other hand, the Kaggle dataset is a meticulously compiled and shared public dataset by Bhuvaji et al. (2020). This dataset encompasses four categories of images: glioma tumors, meningioma tumors, pituitary tumors, and normal brain tissues, totaling 3,264 images. For efficient model training and testing, the two datasets were randomly divided into a training set and a test set. Specifically, 80% of the data was allocated to the training set for model training and optimization, while the remaining 20% was designated as the testing set for evaluating the model's performance. Detailed statistics on the number of images in each dataset are presented in Table 1.

A simple and efficient data preprocessing method is used in the preprocessing phase of the dataset. In the experimental process, to preserve the integrity of image content and stability of features, all images were uniformly resized to $224 \times 224 \times 3$ pixels. This resizing not only helps maintain the spatial structure and information integrity of the images but also significantly reduces computational burden during network training, thereby enhancing training efficiency. Additionally, normalization was performed, which is a standard preprocessing step in deep learning. This aims to mitigate differences in brightness, contrast, and other attributes among images, enabling the model to focus more acutely on learning the inherent features of the images. For medical images, acquiring a large volume of such data can be challenging (Dhar et al., 2023). Given that deep neural networks typically require substantial amounts of data for training, and considering the relatively limited scale of the datasets utilized in this study, data augmentation strategies were employed to alleviate overfitting concerns. Specifically, random rotation and random horizontal flipping techniques were utilized, both of which effectively enhance dataset diversity without introducing additional noise, thereby improving the model's generalization capability.

3.2 Mixed local and global model

In this section, details of the proposed model are provided. The architecture of the MLG model, which combines both local and global components, is depicted in Figure 1. Initially, brain tumor images undergo preprocessing before being fed into a convolutional layer with a kernel size of 5×5 and a stride of 1, designed to enlarge the receptive field. Subsequently, a max pooling layer is applied for downsampling and dimensionality reduction of the extracted features. And then, the features are further processed through five REMA and Biformer (RB) Mixing Blocks to refine the extraction of characteristics specific to brain tumor images. Finally, the resulting features are classified accordingly. The structure of the RB Mixing Block is illustrated in Figure 2.

Figure 2 presents the structure of the RB Mixing Block, primarily consisting of REMA and Biformer units. The REMA

TABLE 1 Details of the datasets.

Dataset name	Classes	Number of each class	Total image count
Chen	Glioma	1,426	3,064
	Meningioma	708	
	Pituitary tumor	930	
Kaggle	Glioma	826	3,264
	Meningioma	822	
	Pituitary tumor	827	
	No tumor	395	

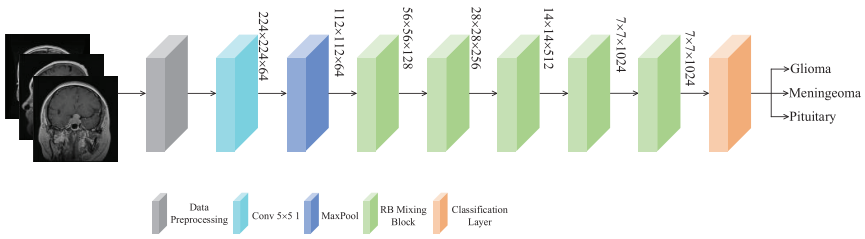


FIGURE 1
Proposed brain tumor classification system.

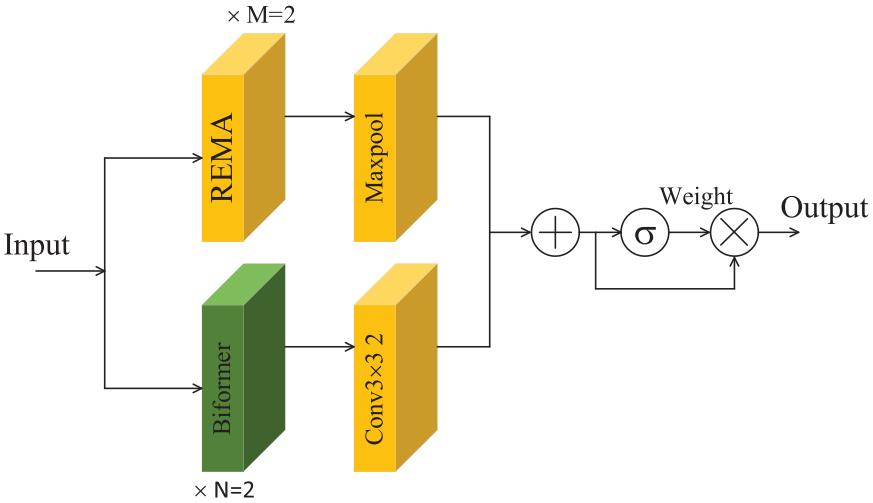


FIGURE 2
The RB Mixing Block structure.

unit is designed to extract local features from the images, while the Biformer unit focuses on extracting global features. After combining the features derived from these two modules, a gating mechanism adjusts the weights of the fused features to better suit the task of brain tumor classification, thereby enhancing the model's classification performance. Here, M denotes the number of REMA convolution modules used and N denotes the number of Biformer modules used, $M = N = 2$. REMA utilizes max pooling for downsampling, aiming to broaden the receptive field of the module. On the other hand, Biformer employs convolutions with a stride of 2 for downsampling, intending to derive higher-level feature representations. Subsequently, the features extracted by both REMA and Biformer are merged and subjected to processing by the gating mechanism. Then, the adjusted features are multiplied with the original ones to modulate their significance in influencing the model's overall performance, effectively filtering out a set of features that have a more substantial impact on the model's classification results. The output of the RB Mixing module can be expressed as:

$$\text{out}_{RB} = sigmoid(f_{\text{REMA}} + f_{\text{Biformer}}) \times (f_{\text{REMA}} + f_{\text{Biformer}}) \quad (1)$$

where, f_{REMA} and $f_{Biformer}$ represent the features extracted by the modules REMA and Biformer, respectively.

In order to present the structure and parameter characteristics of the REMA module and the Biformer module more clearly. We have detailed the number of parameters, input dimensions and output dimensions of these two modules in Table 2.

The structure and computational complexity of the REMA block and the Biformer block in the MLG model can be understood more specifically through Table 2.

3.3 REMA Block

The structure of the REMA block is depicted in Figure 3. This module consists of two convolutional layers and an EMA unit, interconnected via residual connections to facilitate information fusion and propagation. This design aims to enhance the model's representation learning capacity while alleviating the gradient

TABLE 2 Parameters and dimension information of the REMA block and the Biformer block.

Block	Input size	Output size	No. of parameters
REMA	112 × 112 × 64	112 × 112 × 64	74,160
Biformer	112 ×112 × 64	112 × 112 × 64	10,4576

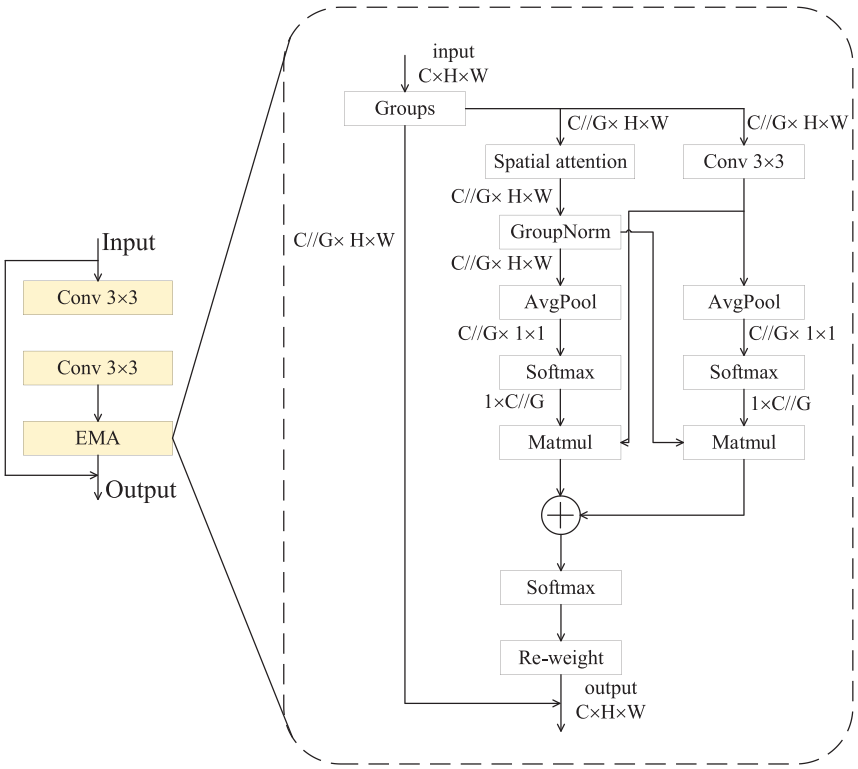


FIGURE 3
REMA structure.

vanishing problem often encountered in deep networks. By incorporating the EMA unit (Ouyang et al., 2023), the REMA block is better equipped to capture inherent data features, thereby boosting the model's performance. The core idea of the EMA module is to group the channel dimensions into multiple sub-features and ensure good distribution of spatial semantic features within each feature group. This method not only preserves information in each channel but also reduces computational overhead. Specifically, the EMA module recalibrates the channel weights of each parallel branch using global information encoding. Moreover, the output features from the two parallel branches are aggregated through cross-dimensional interaction methods, further enhancing the representational power of the features. Inside the EMA module, there are three parallel paths designed to extract attention weight descriptors for the grouped feature maps. Two of these paths belong to the 1×1 branch, while the third one is part of the 3×3 branch. Within the 1×1 branch, two one-dimension global average pooling operations along two spatial directions are employed to encode channel attention. In contrast, the 3×3 branch uses a single 3×3 convolutional kernel to capture multi-scale feature representations. The output of the REMA module can be mathematically represented as follows:

$$\text{out} = EMA(conv(conv(x))) + x \tag{2}$$

The structure of the Biformer Block is depicted in Figure 4. The core of the Biformer lies in its BRA, which consists of a deep convolution, two layers of Layer Normalization (LN), and a Multilayer Perceptron (MLP) interconnected through residual connections (Zhu et al., 2023).

The design principle of BRA revolves around dynamic, query-content based sparsity. Initially, irrelevant key-value pairs are filtered out at a coarse-grained regional level by constructing and pruning a directed graph representing region-level relationships. Subsequently, a fine-grained token-to-token attention mechanism is applied over the joint set of the remaining, or routed, regions to selectively focus on locally relevant information while bypassing globally unrelated data. In BRA process, given a two-dimensional input feature map X, it is partitioned into $S \times S$ non-overlapping regions, each containing a specific number of feature vectors. These region-based features undergo linear projections to generate query, key, and value tensors Q, K, V. An inter-region association matrix A^{γ} is then constructed by computing average query and key vectors across regions, with its elements indicating semantic relevance between pairs of regions. The critical step involves selecting the top k most related adjacent regions for each region based on this relevance measure, yielding a routing index matrix I^{γ} via row-wise top-k operations. Building upon this, the model applies fine-grained token-to-token attention. Specifically, for a query token originating from region i, it attends to all key-value pairs within the k routed regions indexed by $I^{\gamma}_{(i,1)}$ through $I^{\gamma}_{(i,k)}$. To efficiently execute this, despite these regions potentially being scattered throughout the feature map, the model first employs a gather operation to collect the key and value tensors from these regions, forming aggregated key and value sets K_g and V_g. Finally,

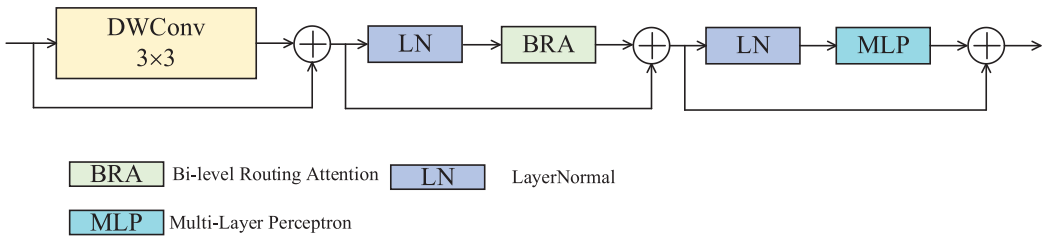


FIGURE 4
Biformer structure.

attention computation is performed using the gathered key and value tensors:

$$O = soft\max(\frac{(QK_g)^T}{\sqrt{C}})V_g + LCE(V) \quad (3)$$

here, $\sqrt{C}$ is usually a factor that scales the denominator in the formula for calculating the attention score in order to prevent the occurrence of over-concentration of weights and loss of gradients. *LCE (V)* represents local context enhancement, which is implemented by depth separable convolution to enhance local information.

3.4 Loss function

In classification tasks, the cross-entropy loss function is a commonly used loss function. Originating from the concepts of entropy and mutual information in information theory, it serves to quantify the discrepancy between two probability distributions. Specifically, when training neural networks, it is employed to measure the difference between the model's predicted probability distribution and the true distribution of the observed data. For classification tasks, assuming the true label is y and the model predicted probability is q, the cross-entropy loss function can be expressed as:

$$H(y, q) = -\sum_i y_i \log(q_i) \quad (4)$$

where, y_i represents the true label for the i-th category and q_i denotes the model predicted probability that the sample belongs to the i-th class.

4 Results

This section introduces the experimental setup, experimental results, and ablation experiments, collectively serving to comprehensively and rigorously substantiate the proposed model.

4.1 Experimental apparatus

A PyTorch implementation is performed for the model proposed by us, while experiments were carried out on a Windows 11 system equipped with a 12GB RTX 4070 GPU and an Intel i5-13400F processor. The Adam optimizer was utilized, with the initial learning rate set at 0.0001, the batch size fixed at 16, and the number of epochs specified as 50. In our experiments, early stopping was utilized to prevent overfitting. Detailed information about the parameters can be found in Table 3.

4.2 Evaluation metrics

In the experiments, the accuracy, recall, precision, and F1-score were employed as evaluation metrics, with their respective calculation methods presented in Formulas (5–8). The accuracy is one of the most commonly used evaluation metrics in classification problems, representing the proportion of correctly classified samples out of the total number of samples. The recall, focuses on the ability of the model to correctly identify positive samples, which refers to the ratio of true positives (correctly identified positive instances) to all actual positive instances in the dataset. The precision measures the proportion of instances predicted by the model as positive that are truly positive, that is, the ratio of true positives to all instances predicted as positive. The F1-score, being the harmonic mean of precision and recall, integrates the performance of both precision and recall, offering a more comprehensive assessment of the model's performance (Zulfiqar et al., 2023; Zebari et al., 2024). When both precision and recall are high, the F1-score will also be high, and conversely, when either of these values is low, so will the F1-score. This implies that a high F1-score indicates strong overall performance in terms of both accurately identifying true positives and minimizing false predictions.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

TABLE 3 Training Hyper-parameter values of proposed network.

Parameters	Value
Initial learning rate	0.0001
Batch size	16
Optimizer	Adam
Number of epoch	50
Learning rate decays	0.1

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{8}$$

4.3 The results of the experiment

Figure 5 illustrates the confusion matrices for the classification results of the model on the test sets of two publicly available datasets, where G, M, and P stand for glioma, meningioma, and pituitary adenoma, respectively, and N stands for normal state, indicating the absence of brain tumor. From the confusion matrices, detailed classification performance metrics for the model were calculated according to Formulas (5–8) and summarized in Table 4. From Table 4, it is evident that, on the test set of the Chen dataset, the average performance metrics for model MLG include a recall of 98.88%, precision of 98.94%, F1-score of 98.91%, and accuracy of 99.02%. On the Kaggle dataset test set, MLG corresponding metrics are 96.89% for recall, 97.21% for precision, 96.89% for F1-score, and 97.24% for accuracy. These indicators demonstrate that across both the Chen and Kaggle datasets, the MLG model exhibits outstanding classification performance, which further validates the effectiveness and generalization capabilities of the MLG model, enabling it to achieve satisfactory performance in brain tumor classification tasks on diverse datasets.

4.4 Ablation study

In Section 4.3, performance metrics for the classification results of the proposed model are presented. To further confirm the validity of the proposed model, an ablation study was performed. In this study, different combinations of modules are explored within the framework of the model. This process allows for a meticulous examination of each component's contribution to the overall performance, thereby providing deeper insights into the effectiveness and robustness of the proposed model architecture.

In the first part of the study, brain tumor classification was conducted separately using REMA and Biformer independently. Figure 6 presents the testing results of various models in the Chen dataset during the ablation experiment. The accuracies achieved by REMA and Biformer are 98.53 and 98.37%, respectively, both lower than the 99.02% accuracy obtained by MLG. Upon conducting a detailed analysis of the ablation experiment results, it becomes clear that the integration of the strengths of both the REMA and Biformer modules within the MLG model effectively boosts the accuracy rate in brain tumor classification.

In the second part of the study, the performance of the MLG model upon incorporating the gated attention mechanism was meticulously examined. The gated attention mechanism plays a pivotal role within the model, serving to regulate the flow of information by deciding which pieces of information should be emphasized and which should be disregarded. By means of gating, the attention mechanism assigns weights to information based on its importance, thereby enhancing the model performance by focusing on crucial features. Figure 7 shows the performance of the model with and without the gated attention mechanism. Where, GA stands for Gated Attention. It can be observed that when the model does not include the gated attention, its performance lags behind the version with the gated attention mechanism by 2.12%. The results strongly demonstrate the effectiveness of the gated attention in improving the performance of the model.

In the third segment of the investigation, the impact of data augmentation on the MLG model was thoroughly explored, particularly in scenarios involving small sample datasets. Data augmentation is a critical technique that can significantly enhance a model generalization capability while mitigating overfitting issues. In this work, two prevalent data augmentation strategies were employed: random rotation and random flipping. Figure 8 provides a detailed account of the model accuracy rates on both the training and test sets of the Chen dataset when data augmentation is applied. Ar stands for data augmentation. From the figure, it is evident

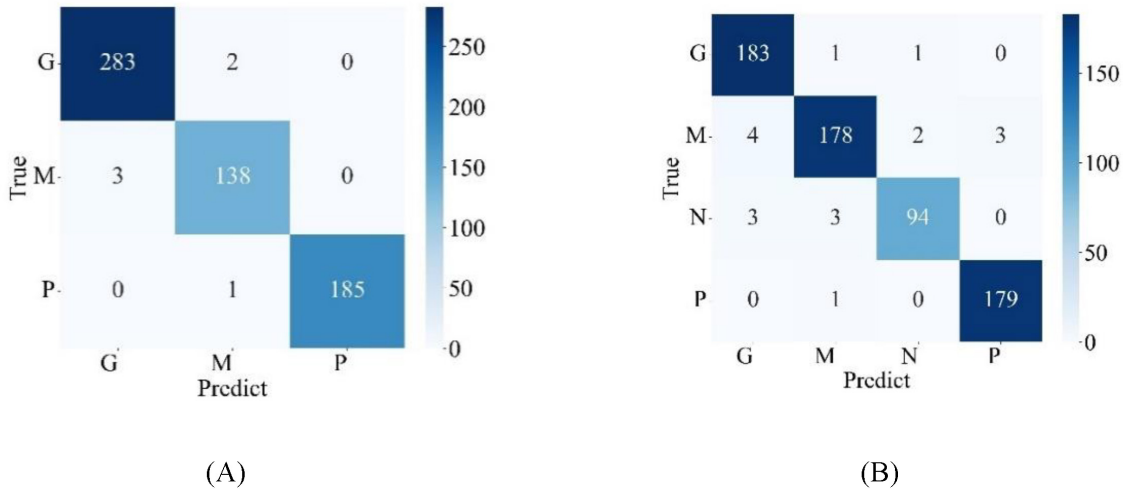


FIGURE 5
Confusion matrix for model classification results **(A)** Chen dataset **(B)** Kaggle dataset.

TABLE 4 Detailed values of metrics for the proposed model.

Dataset	Tumor type	Recall (%)	Precision (%)	F1-score (%)	Accuracy (%)
Chen	Glioma	99.30	98.95	99.12	99.02
	Meningioma	97.87	97.87	97.87	
	Pituitary	99.46	1.00	99.73	
	Average	**98.88**	**98.94**	**98.91**	
Kaggle	Glioma	98.92	96.32	97.60	97.24
	Meningioma	95.19	97.27	96.22	
	No Tumor	94.00	96.91	95.43	
	Pituitary	99.44	98.35	98.90	
	Average	**96.89**	**97.21**	**96.89**	

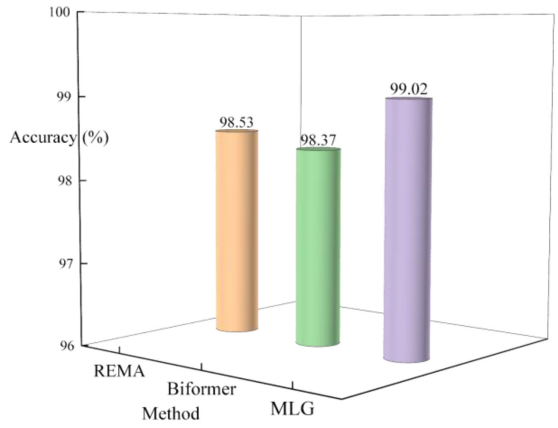


FIGURE 6
Classification results by REMA and Biformer.

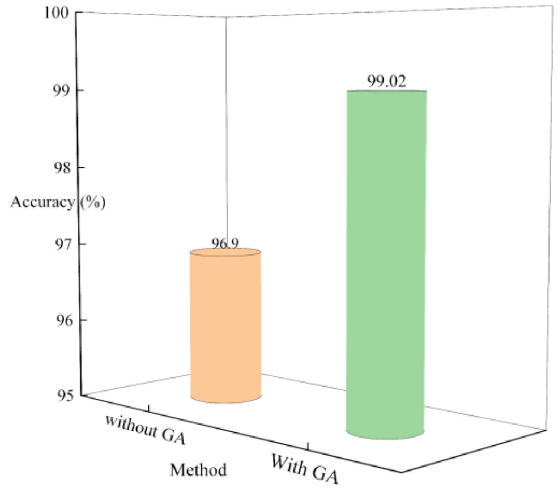


FIGURE 7
Effects of gating attention mechanism on MLG.

that with data augmentation, the training and test set accuracies reach 99.96 and 99.02%, respectively. In contrast, without data augmentation, while the accuracy on the training set reached 100%, the accuracy on the test set notably decreased to 96.73%. This

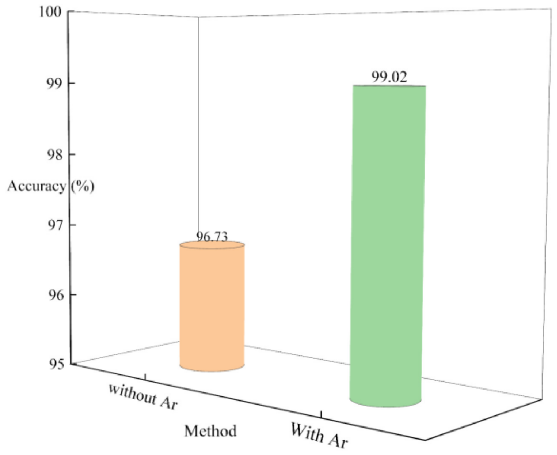


FIGURE 8
Impact of data augmentation on MLG.

comparative outcome vividly demonstrates that data augmentation has a pronounced effect on improving model performance.

5 Discussion

According to the data in Table 4, the MLG model achieves impressive accuracies of 99.02% on the Chen dataset and 97.24% on the Kaggle dataset, which attest to its effectiveness and satisfactory performance. Moreover, through ablation studies, the superiority of the MLG model was further substantiated, emphasizing the significant improvements gained by fusing the REMA and Biformer modules via the gated attention mechanism, rather than merely adding them together. Additionally, the application of data augmentation has led to noticeable performance enhancements, further bolstering the model generalization capabilities.

Beyond internal validation, the proposed model was also compared against other advanced methods utilizing the same datasets. Table 5 clearly outlines these comparative results. On the Chen dataset, the MLG model outperforms the current best-performing model, Multimodal-CNN Model (Maqsood et al., 2022), by 0.1% in accuracy. Similarly, on the Kaggle dataset, the MLG model surpasses the previously best-reported model IVX16 (Hossain et al., 2023) by an accuracy margin of 0.3%. When juxtaposed against methodologies outlined in literature sources paper (Alanazi et al., 2022)and paper (Saurav et al., 2023), the MLG model consistently demonstrates higher performance on both the Chen and Kaggle datasets. Precisely, on the Chen dataset, MLG accuracy exceeds that of paper (Alanazi et al., 2022) by 2.13% and that of paper (Saurav et al., 2023) by 1.79%. On the Kaggle dataset, MLG accuracy advantage over paper (Alanazi et al., 2022) is 1.49%, while over (Saurav et al., 2023) it is 1.53%. These comparative results serve as compelling evidence of the MLG model superior performance in the task of brain tumor classification, reinforcing its potential applicability in real-world scenarios.

The Receiver Operating Characteristic Curve (ROC Curve) is a widely used visualization tool in statistics, machine learning, medical diagnostics, and other fields that require categorical judgments for evaluating the performance of classification models.

TABLE 5 Compare with advanced methods on datasets Chen and Kaggle.

Method category	References	Method	Dataset	Accuracy (%)
CNN	Sachdeva et al. (2024)	Transfer learning	Kaggle	96.25
	Jun and Liyuan (2022)	Attention-Guided	Chen	98.61
	Maqsood et al. (2022)	Multimodal-CNN Model	Chen	98.92b
	Alanazi et al. (2022)	22-layer CNN	Chen	96.89
			Kaggle	95.75
	Saurav et al. (2023)	AG-CNN	Chen	97.23
			Kaggle	95.71
Transformer	Wang et al. (2024)	RanMerFormer	Chen	98.86
	Şahin et al. (2024)	BMO	Chen	98.09
	Hossain et al. (2023)	IVX16	Kaggle	96.94
	Anaya-Isaza et al. (2023)	Cross-Transformer	Chen	97.22
	Dosovitskiy et al. (2021)	Vision Transformer	Chen	97.39
			Kaggle	95.88
	Liu et al. (2021a)	Swin Transformer	Chen	98.69
			Kaggle	97.10
CNN+ transformer	Ferdous et al. (2023)	LCDEiT	Chen	98.11
	Chen et al. (2025)	EnSLDe	Chen	98.69
	Proposed model	MLG	Chen	99.02
			Kaggle	97.24

It graphically illustrates the trade-off relationship between the true positive rate (TPR) and false positive rate (FPR) of the model under different threshold conditions. The area under curve (AUC), indicates better model performance when its value is larger. Typically, the closer the curve is to the upper left corner (with higher TPR and lower FPR), the better the model performance. The ROC curves of the model on the two datasets are shown in Figure 9. It can be observed that the ROC curves closely adhere to the upper left corner. On the Chen dataset, the AUC values of the MLG model for glioma, meningioma, and pituitary tumors are 0.9996, 0.9993, and 1.00, respectively. Meanwhile, on the Kaggle dataset, the AUC values of the MLG model for glioma, meningioma, normal tissue, and pituitary tumors are 0.9991, 0.9965, 0.9989, and 0.9999, respectively.

6 Conclusion

Brain tumors, constituting a severe health issue, pose a significant threat to people's lives. Therefore, timely and accurate identification of brain tumor types, followed by appropriate treatment planning, is critical for patients. The advent of CAD technology has provided substantial support to doctors in diagnosing brain tumors. In this paper, a novel MLG brain tumor classification model is proposed, and the model skillfully integrates local features and global features, and provides a new solution for the classification of brain tumors. The core components of the MLG model are RMEA, Biformer and gated attention. The RMEA Block, through carefully designed convolutional structures, efficiently retains information across channels, emphasizing spatial and channel-wise features, thereby extracting richly informative local features. Conversely, the Biformer employs a unique BRA mechanism to dynamically and contextually select a subset of the most relevant key-value pairs for each query, optimizing the computational process. Meanwhile, BRA can capture remote dependencies across regions and even objects, providing powerful support for extracting global features. The MLG model uses a gated attention to selectively filter and fuse the local features extracted by the RMEA block with the global features extracted by the Biformer block. This significantly enhances the representation capability of the fused features, thereby improving the classification performance of the model. The integration of both local and global features enables the MLG model to exhibit outstanding

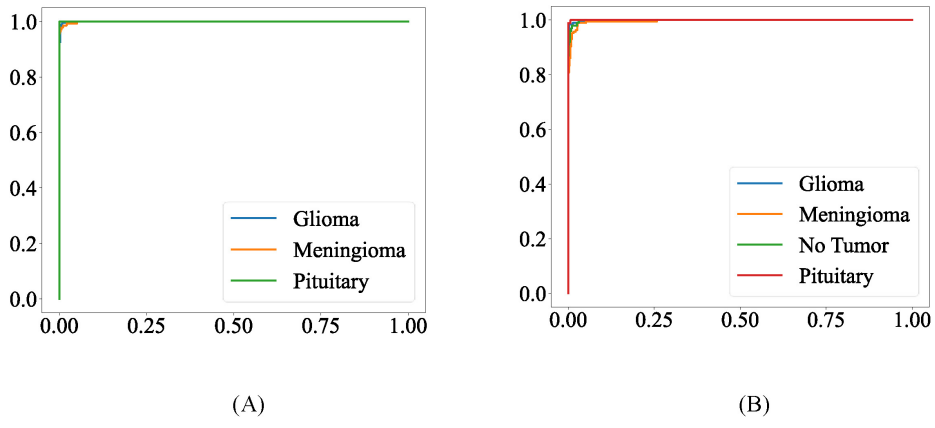


FIGURE 9
ROC curves for the proposed model on **(A)** Chen dataset **(B)** Kaggle dataset.

performance in brain tumor classification tasks. Experimental results on two public datasets demonstrate that the MLG model achieves satisfactory performance across multiple metrics, including accuracy, precision, recall, and F1-score. Compared with existing advanced methods, the MLG model exhibits marked advantages, fully validating its effectiveness in practical applications. In future work, it is planned to continue exploring other methods of feature fusion first to further improve the performance of the MLG model. Secondly, the introduction of more refined feature detection methods will be explored, or they will be combined with other advanced attention mechanisms to enhance the selection ability for key areas. In addition, efforts will also be made to obtain data on other brain diseases, expand the application scope of the model, and provide more auxiliary diagnostic tools for the medical field.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: the datasets used are free and open. Dataset Chen from figshare (https://figshare.com/articles/dataset/brain_tumor_dataset/1512427). Dataset Kaggle from Kaggle (https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri).

Author contributions

WC: Project administration, Conceptualization, Visualization, Writing – review & editing, Investigation. XT: Formal Analysis, Software, Writing – original draft. JZ: Conceptualization, Project administration, Writing – review & editing, Software. GD: Project administration, Supervision, Writing – review & editing. QF: Writing – review & editing, Validation. HJ: Writing – review & editing, Project administration, Validation.

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 Henan Province Young Backbone Teachers Training Program (No. 2023GGJS045), the Major Science and Technology Projects of Henan Province (No. 221100210500), the Foundation of Henan Educational Committee (No. 24A320004), the Medical and Health Research Project in Luoyang (No. 2001027A), and the Construction Project of Improving Medical Service Capacity of Provincial Medical Institutions in Henan Province (No. 2017-51).

Acknowledgments

The provision of these two public datasets by Kaggle and Chen is greatly appreciated by us.

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 Generative 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

Akter, A., Nosheen, N., Ahmed, S., Hossain, M., Yousuf, M., Almoyad, A., et al. (2024). Robust clinical applicable CNN and U-Net based algorithm for MRI classification and segmentation for brain tumor. *Expert Syst. Appl.* 238:122347. doi: 10.1016/j.eswa.2023.122347

Alanazi, M., Ali, M., Hussain, S., Zafar, A., Mohatram, M., Irfan, M., et al. (2022). Brain tumor/mass classification framework using magnetic-resonance-imaging-based isolated and developed transfer deep-learning model. *Sensors* 22:372. doi: 10.3390/s22010372

Alturki, N., Umer, M., Ishaq, A., Abuzinadah, N., Alnowaiser, K., Mohamed, A., et al. (2023). Combining CNN features with voting classifiers for optimizing performance of brain tumor classification. *Cancers* 15:1767. doi: 10.3390/cancers15061767

Anaya-Isaza, A., Mera-Jiménez, L., Verdugo-Alejo, L., and Sarasti, L. (2023). Optimizing MRI-based brain tumor classification and detection using AI: A comparative analysis of neural networks, transfer learning, data augmentation, and the cross-transformer network. *Eur. J. Radiol. Open* 10:100484. doi: 10.1016/j.ejro.2023.100484

Asiri, A., Shaf, A., Ali, T., Pasha, M., Khan, A., Irfan, M., et al. (2024). Advancing brain tumor detection: Harnessing the Swin Transformer's power for accurate classification and performance analysis. *PeerJ Comput. Sci.* 10:e1867. doi: 10.7717/peerj-cs.1867

Azeem, M., Kiani, K., Mansouri, T., and Topping, N. (2024). SkinLesNet: Classification of skin lesions and detection of melanoma cancer using a novel multi-layer deep convolutional neural network. *Cancers* 16:108. doi: 10.3390/cancers16010108

Bhuvaji, S., Kadam, A., Bhumkar, P., Dedge, S., and Kanchan, S. (2020). *Brain tumor classification (MRI)*. San Francisco, CA: Kaggle, doi: 10.34740/KAGGLE/DSV/1183165

Cao, L., Pan, K., Ren, Y., Lu, R., and Zhang, J. (2024). Multi-branch spectral channel attention network for breast cancer histopathology image classification. *Electronics* 13:459. doi: 10.3390/electronics13020459

Chen, W., Liu, J., Tan, X., Zhang, J., Du, G., Fu, Q., et al. (2025). EnSLDe: An enhanced short-range and long-range dependent system for brain tumor classification. *Front. Oncol.* 15:1512739. doi: 10.3389/fonc.2025.1512739

Cheng, J., Huang, W., Cao, S., Yang, Ru, Yang, W., Yun, Z., et al. (2015). Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PLoS One* 10:e0140381. doi: 10.1371/journal.pone.0140381

Chincholi, F., and Koestler, H. (2024). Transforming glaucoma diagnosis: Transformers at the forefront. *Front. Artif. Intell.* 7:1324109. doi: 10.3389/frai.2024.1324109

Dhar, T., Dey, N., Borra, S., and Sherratt, R. S. (2023). Challenges of deep learning in medical image analysis–improving explainability and trust. *IEEE Trans. Technol. Soc.* 4, 68–75. doi: 10.1109/TTS.2023.3234203

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* [Preprint]. doi: 10.48550/arXiv.2010.11929 arXiv:2010.11929.

Fang, M., Fu, M., Liao, B., Lei, X., and Wu, F.-X. (2024). Deep integrated fusion of local and global features for cervical cell classification. *Comput. Biol. Med.* 171:108153. doi: 10.1016/j.compbiomed.2024.108153

Ferdous, G. J., Sathi, K. A., Hossain, A., Hoque, M. M., and Dewan, M. A. A. (2023). LCDEiT: A linear complexity data-efficient image transformer for MRI brain tumor classification. *IEEE Access* 11, 20337–20350. doi: 10.1109/ACCESS.2023.3244228

Gade, V. S. R., Cherian, R. K., Rajarao, B., and Kumar, M. A. (2024). BMO based improved Lite Swin transformer for brain tumor detection using MRI images. *Biomed. Signal Process. Control* 92:91. doi: 10.1016/j.bspc.2024.106091

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE conference on computer vision and pattern recognition (CVPR)*, (Las Vegas, NV: IEEE), 770–778. doi: 10.1109/CVPR.2016.90

Hossain, S., Chakrabarty, A., Gadekallu, T. R., Alazab, M., and Piran, J. (2023). Vision transformers, ensemble model, and transfer learning leveraging explainable AI for brain tumor detection and classification. *IEEE J. Biomed. Health Inform.* 28, 1261–1272. doi: 10.1109/JBHI.2023.3266614

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (Salt Lake City, UT), 7132–7141.

Huang, S. K., Yu, Y.-T., Huang, C.-R., and Cheng, H.-C. (2024). Cross-scale fusion transformer for histopathological image classification'. *IEEE J. Biomed. Health Inform.* 28, 297–308. doi: 10.1109/JBHI.2023.3322387

Huang, L., Xu, Y., Wang, S., Sang, L., and Ma, H. (2024). SRT: Swin-residual transformer for benign and malignant nodules classification in thyroid ultrasound images'. *Med. Eng. Phys.* 124:104101. doi: 10.1016/j.medengphy.2024.104101

Jun, W., and Liyuan, Z. (2022). Brain tumor classification based on attention guided deep learning model'. *Int. J. Comput. Intell. Syst.* 15:35. doi: 10.1007/s44196-022-00090-9

Kang, J., Ullah, Z., and Gwak, J. (2021). MRI-based brain tumor classification using ensemble of deep features and machine learning classifiers'. *Sensors* 21:2222. doi: 10.3390/s21062222

Kaur, P., and Mahajan, P. (2025). Detection of brain tumors using a transfer learning-based optimized ResNet152 model in MR images.'. *Comput. Biol. Med.* 188:109790. doi: 10.1016/j.compbiomed.2025.109790

Kshatri, S. S., and Singh, D. (2023). Convolutional neural network in medical image analysis: A review'. *Arch. Comput. Methods Eng.* 30, 2793–2810. doi: 10.1007/s11831-023-09898-w

Li, Z., and Zhou, X. (2025). A global-local parallel dual-branch deep learning model with attention-enhanced feature fusion for brain tumor MRI classification'. *CMC Comput. Mater. Contin.* 83, 739–760. doi: 10.32604/cmc.2025.059807

Liu, H., Huo, G., Li, Q., Guan, X., and Tseng, M. (2023). Multiscale lightweight 3D segmentation algorithm with attention mechanism: Brain tumor image segmentation'. *Expert Syst. Appl.* 214:9166. doi: 10.1016/j.eswa.2022.119166

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021b). "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the 2021 IEEE/CVF international conference on computer vision (ICCV)*, (Montreal, QC: IEEE), 9992–10002. doi: 10.1109/ICCV48922.2021.00986

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021a). Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv* [Preprint]. doi: 10.48550/arXiv.2103.14030 arXiv:2103.14030.

Lyu, I. J., Han, K., Park, K.-A., and Oh, S. Y. (2024). Ocular motor cranial nerve palsies and increased risk of primary malignant brain tumors: South Korean national health insurance data'. *Cancers* 16:781. doi: 10.3390/cancers16040781

Maqsood, S., Damaševičius, R., and Maskeliūnas, R. (2022). Multi-modal brain tumor detection using deep neural network and multiclass SVM'. *Medicina* 58:1090. doi: 10.3390/medicina58081090

Mazurowski, M. A., Dong, H., Gu, H., Yang, J., Konz, N., and Zhang, Y. (2023). Segment anything model for medical image analysis: An experimental study'. *Med. Image Anal.* 89:102918. doi: 10.1016/j.media.2023.102918

Mehnatkesh, H., Jalali, S. M. J., Khosravi, A., and Nahavandi, S. (2023). An intelligent driven deep residual learning framework for brain tumor classification using MRI images'. *Expert Syst. Appl.* 213:119087. doi: 10.1016/j.eswa.2022.119087

Mukadam, S. B., and Patil, H. Y. (2024). Machine learning and computer vision based methods for cancer classification: A systematic review'. *Arch. Comput. Methods Eng.* 31, 3015–3050. doi: 10.1007/s11831-024-10065-y

Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., et al. (2023). "Efficient multi-scale attention module with cross-spatial learning," in *Proceedings of the ICASSP 2023–2023 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (Rhodes), 1–5. doi: 10.1109/ICASSP49357.2023.10096516

Pandiselvi, T., and Maheswaran, R. (2019). Efficient framework for identifying, locating, detecting and classifying MRI brain tumor in MRI images. *J. Med. Syst.* 43:189. doi: 10.1007/s10916-019-1253-1

Peng, J., Lu, J., Zhuo, J., and Li, P. (2024). Multi-scale-denoising residual convolutional network for retinal disease classification using OCT'. *Sensors* 24:150. doi: 10.3390/s24010150

Poornam, S., and Angelina, J. J. R. (2024). VITALT: A robust and efficient brain tumor detection system using vision transformer with attention and linear transformation'. *Neural Comput. Appl.* 36, 6403–6419. doi: 10.1007/s00521-023-09306-1

Sachdeva, J., Sharma, D., and Ahuja, C. K. (2024). Comparative analysis of different deep convolutional neural network architectures for classification of brain tumor on magnetic resonance images'. *Arch. Comput. Methods Eng.* 31, 1959–1978. doi: 10.1007/s11831-023-10041-y

Şahin, E., Özdemir, D., and Temurtaş, H. (2024). Multi-objective optimization of ViT architecture for efficient brain tumor classification. *Biomed. Signal Process. Control* 91:105938. doi: 10.1016/j.bspc.2023.105938

Saurav, S., Sharma, A., Saini, R., and Singh, S. (2023). An attention-guided convolutional neural network for automated classification of brain tumor from MRI. *Neural Comput. Appl.* 35, 2541–2560. doi: 10.1007/s00521-022-07742-z

Shafiq, M., and Gu, Z. (2022). Deep residual learning for image recognition: A survey. *Appl. Sci. Basel* 12:8972. doi: 10.3390/app12188972

Sharma, P., Nayak, D. R., Balabantaray, B. K., Tanveer, M., and Nayak, R. (2024). A survey on cancer detection via convolutional neural networks: Current challenges and future directions'. *Neural Netw.* 169, 637–659. doi: 10.1016/j.neunet.2023.11.006

Song, B., Kc, D. R., Yang, R. Y., Li, S., Zhang, C., and Liang, R. (2024). Classification of mobile-based oral cancer images using the vision transformer and the Swin transformer'. *Cancers* 16:987. doi: 10.3390/cancers16050987

ThamilSelvi, C., Vinoth Kumar, S., Asaad, R. R., Palanisamy, P., and Rajappan, L. K. (2025). An integrative framework for brain tumor segmentation and classification using neuraclassnet'. *Intell. Data Anal.* 29, 435–458. doi: 10.3233/IDA-240108

Verma, A., and Yadav, A. K. (2025). FusionNet: Dual input feature fusion network with ensemble based filter feature selection for enhanced brain tumor classification'. *Brain Res.* 1852:149507. doi: 10.1016/j.brainres.2025.149507

Wang, J., Lu, S.-Y., Wang, S.-H., and Zhang, Y.-D. (2024). RanMerFormer: Randomized vision transformer with token merging for brain tumor classification'. *Neurocomputing* 573:127216. doi: 10.1016/j.neucom.2023.127216

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). "CBAM: Convolutional block attention module," in *Computer vision – ECCV 2018*, Vol. 11211, eds V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss (Cham: Springer International Publishing), 3–19. doi: 10.1007/978-3-030-01234-2_1

Yan, Q., Liu, S., Xu, S., Dong, C., Li, Z., Shi, J., et al. (2023). 3D medical image segmentation using parallel transformers'. *Pattern Recogn.* 138:109432. doi: 10.1016/j.patcog.2023.109432

Yu, X., Wang, J., Hong, Q.-Q., Teku, R., Wang, S.-H., and Zhang, Y.-D. (2022). Transfer learning for medical images analyses: A survey'. *Neurocomputing* 489, 230–254. doi: 10.1016/j.neucom.2021.08.159

Yu, Y., Zhang, Y., Cheng, Z., Song, Z., and Tang, C. (2023). MCA: Multidimensional collaborative attention in deep convolutional neural networks for image recognition'. *Eng. Appl. Artif. Intell.* 126:107079. doi: 10.1016/j.engappai.2023.107079

Zebari, N. A., Mohammed, C., Zebari, D., Mohammed, M., Zeebaree, D., Marhoon, H., et al. (2024). A deep learning fusion model for accurate classification of brain tumours in magnetic resonance images'. *CAAI Trans. Intell. Technol.* 9:76. doi: 10.1049/cit2.12276

Zhou, L., Jiang, Y., Li, W., Hu, J., and Zheng, S. (2024). Shape-scale co-awareness network for 3d brain tumor segmentation'. *IEEE Trans. Med. Imaging* 43, 2495–2508. doi: 10.1109/TMI.2024.3368531

Zhu, L., Wang, X., Ke, Z., Zhang, W., and Lau, R. (2023). "BiFormer: Vision transformer with bi-level routing attention," in *Proceedings of the 2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, (Vancouver, BC: IEEE), 10323–10333. doi: 10.1109/CVPR52729.2023.00995

Zulfiqar, F., Bajwa, U. I., and Mehmood, Y. (2023). Multi-class classification of brain tumor types from MR images using EfficientNets'. *Biomed. Signal Process. Control* 84:104777. doi: 10.1016/j.bspc.2023.104777

frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 21 July 2025
DOI 10.3389/fnins.2025.1609547

Check for updates

OPEN ACCESS

EDITED BY
Pardeep Sangwan,
Maharaja Surajmal Institute of Technology, India

REVIEWED BY
Ju Gao,
Suzhou Guangji Hospital, China
Aviral Chharia,
Carnegie Mellon University, United States

*CORRESPONDENCE
Shaolong Wei
✉ weishaolong37@gmail.com
Hongcheng Yao
✉ yaohongcheng19@gmail.com

RECEIVED 10 April 2025
ACCEPTED 01 July 2025
PUBLISHED 21 July 2025

CITATION
Yuan X, Wei S, Sun Y, Gu L, He Y, Chen T, Yao H and Rao H (2025) Robust multi-task feature selection with counterfactual explanation for schizophrenia identification using functional brain networks. *Front. Neurosci.* 19:1609547. doi: 10.3389/fnins.2025.1609547

COPYRIGHT
© 2025 Yuan, Wei, Sun, Gu, He, Chen, Yao and Rao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Robust multi-task feature selection with counterfactual explanation for schizophrenia identification using functional brain networks

Xinyan Yuan[1], Shaolong Wei[2]*, Ying Sun[1], Lingling Gu[1], Yanyan He[1], Tiantian Chen[1], Hongcheng Yao[3]* and Haonan Rao[3]

[1]School of Electronics and Information, Jiangsu Vocational College of Business, Nantong, China, [2]School of Artificial Intelligence and Computer Science, Nantong University, Nantong, China, [3]School of Information Science and Technology, Nantong University, Nantong, China

Introduction: Functional brain networks measured by resting-state functional magnetic resonance imaging (rs-fMRI) have become a promising tool for understanding the neural mechanisms underlying schizophrenia (SZ). However, the high dimensionality of these networks and small sample sizes pose significant challenges for effective classification and model generalization.

Methods: We propose a robust multi-task feature selection method combined with counterfactual explanations to improve the accuracy and interpretability of SZ identification. rs-fMRI data are preprocessed to construct a functional connectivity matrix, and features are extracted by sorting the upper triangular elements. A multi-task feature selection framework based on the Gray Wolf Optimizer (GWO) is developed to identify abnormal functional connectivity (FC) features in SZ patients. A counterfactual explanation model is applied to reduce perturbations in abnormal FC features, returning the model prediction to normal and enhancing clinical interpretability.

Results: Our method was tested on five real-world SZ datasets. The results demonstrate that the proposed method significantly outperforms existing methods in terms of classification accuracy while offering new insights into the analysis of SZ through improved feature selection and explanation.

Discussion: The integration of multi-task feature selection and counterfactual explanation improves both the accuracy and interpretability of SZ identification. This approach provides valuable clinical insights by revealing the key functional connectivity features associated with SZ, which could assist in the development of more effective diagnostic tools.

KEYWORDS

schizophrenia, functional connectivity, rs-fMRI, feature selection, counterfactual explanation

1 Introduction

Schizophrenia (SZ) is a chronic, often disabling mental disorder that affects one percent of the world's population (Insel, 2010; McCutcheon et al., 2020). Patients' clinical symptoms manifest in perception, thinking, and emotion, such as hallucinations, delusions, incoordinated excitement, and anxiety (Song et al., 2023; Rantala et al., 2022). Although the pathogenesis of SZ is still unclear, it is increasingly recognized that analyzing the brain network of SZ can help improve differential diagnosis and understand the pathological mechanism (Zhang et al., 2021). Recent studies have shown that functional

brain networks measured by resting-state functional magnetic resonance imaging (rs-fMRI) have become a promising tool to reveal the underlying neural mechanisms of SZ (Zhu et al., 2024; Chyzhyk et al., 2015). SZ causes widespread changes in functional brain networks, including changes in global brain topology, abnormal connectivity in local regions, and the formation of specific abnormal subgraphs (Huang et al., 2025).

However, although functional brain networks provide rich pathological information, these data often have high-dimensional characteristics, making analysis and modeling face great challenges (Mhiri and Rekik, 2020). Therefore, feature selection (FS) becomes an indispensable step, which can remove irrelevant or redundant features and retain only the most diagnostically valuable information (Naheed et al., 2020). In addition, functional brain network data usually face the problem of small samples. Due to the high cost of data acquisition, the long experimental cycle, and the difficulty in recruiting subjects, the number of samples is often much lower than the feature dimension, making model training susceptible to overfitting, thereby reducing generalization ability (Turner et al., 2018; Ding et al., 2024). In this context, robust and effective FS is vital. In fact, FS plays a key role in identifying meaningful biomarkers, such as functional connectivity between brain regions, which can characterize abnormalities in brain function associated with brain diseases such as SZ, thus providing insight into understanding the neural basis of brain diseases, as well as diagnosis and prediction (Xing et al., 2022).

For functional brain network data, the traditional FS method often exhibits poor robustness across datasets, primarily due to the high dimensionality of the feature space and the scarcity of training samples, and it is difficult to identify connection features with consistency and biological interpretability (Wang et al., 2015; Lv et al., 2015; Hu et al., 2021). At present, most existing FS methods have combined advanced technologies such as machine learning or deep learning to improve performance, such as using graph neural networks to model FC structures, or improving feature selection efficiency through embedded FS strategies, but these methods still have obvious limitations. On the one hand, many models still lack consistent evaluation across data sets, making it difficult to identify robust disease-related connection features (Chan et al., 2024); on the other hand, most existing methods are black-box in form and lack interpretability, especially in clinical applications. It is difficult to provide actionable explanations or intervention recommendations (Verma et al., 2023). In addition, although some studies have introduced multimodal or high-order connection features in SZ diagnosis, it is still difficult to achieve a good balance between model generalization and explanatory power (Sunil et al., 2024).

To address the above challenges and fill this gap, we proposed a novel and robust multi-task feature selection method for SZ diagnosis, and explained the changes in brain functional connectivity (FC) caused by the disease through a counterfactual explanation model. The schematic diagram of our proposed method is shown in Figure 1. Specifically, we first preprocessed the rs-fMRI data, constructed the FC matrix, and then extracted the upper triangular elements as feature vectors and sorted them. Subsequently, we developed a robust multi-task feature selection framework based on the Gray Wolf Optimizer (GWO), and selected the abnormal FC features of SZ patients by adopting feature stratification and weight-based task generation. Finally, we used the counterfactual explanation model to generate a set of counterfactual examples for SZ patients, that is, by fine-tuning the abnormal FC features of SZ patients to make their state close to normal, thus providing theoretical guidance for the analysis and diagnosis of SZ. We verified the effectiveness of our method on five real SZ datasets, and the results showed that our method not only improved the interpretability of the model, but also provided a new perspective for the analysis of SZ. The main contributions of this paper are as follows:

- We propose a Robust Multi-Task Feature Selection with Counterfactual Explanation for Schizophrenia Identification to assist SZ analysis and diagnosis.
- We construct a multi-task feature selection framework based on GWO and combine it with the counterfactual explanation model to fine-tune the abnormal FC features of SZ patients to make their status closer to that of healthy individuals, thereby improving the accuracy of SZ classification and the interpretability of the model.
- We evaluate the performance of the proposed method using five real SZ datasets. The results show that the proposed method outperforms existing methods.

2 Related work

2.1 Gray wolf optimizer

Gray Wolf Optimizer (GWO) (Mirjalili et al., 2014) is an intelligent optimization algorithm that simulates the hunting behavior of gray wolf groups. In the context of multitasking, GWO provides efficient global search capabilities and information-sharing mechanisms between individuals, which can improve optimization performance in a multi-task environment.

Gray wolf packs are generally divided into four levels: (i) α is the leader of the wolf pack, representing the current optimal solution, (ii) β is the second-level wolf, assisting α in decision-making, representing the second-best solution, (iii) δ is the third-level wolf, assisting β, representing the third-best solution, and (iv) θ is an ordinary wolf that obeys other high-level wolves and represents the remaining candidate solutions. When searching for prey, gray wolves will gradually approach the prey and surround it:

$$D = |C \cdot X_p - X| \tag{1}$$

$$X(t+1) = X_p - A \cdot D \tag{2}$$

where X_p is the location of the prey or the current optimal solution, X is the location of the individual wolf, t is the number of iterations, and A and C are coefficient vectors, which are calculated as follows:

$$A = 2d \cdot r_1 - d, \; C = 2r_2 \tag{3}$$

where d is the convergence factor that decreases linearly with the number of iterations, from 2 to 0, and r_1 and r_2 are random numbers between $[0, 1]$. GWO uses three optimal solutions (α, β, δ) to jointly guide the search:

$$X(t+1) = \frac{1}{3} \sum_{i=\alpha,\beta,\delta} (X_i - A_i \cdot D_i) \tag{4}$$

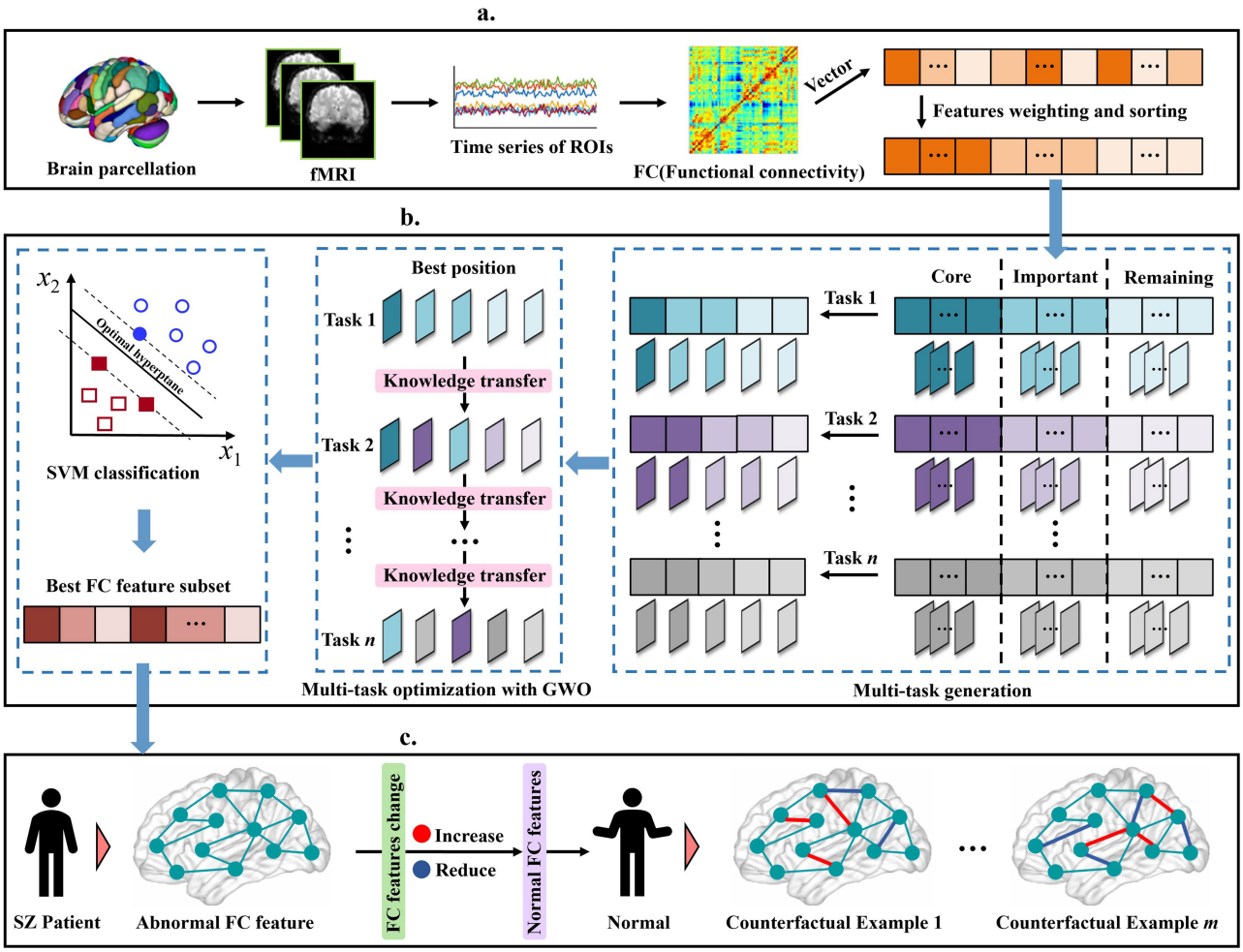


FIGURE 1
Illustration of our proposed schizophrenia analysis method, including **(a)** data pre-processing, **(b)** Robust Multi-Task Feature Selection, **(c)** Diversity counterfactual explanation.

where $D_i = |C_i \cdot X_i - X|, i \in \{\alpha, \beta, \delta\}$. When $|A|$ becomes smaller (approaches 0), the search range is reduced, and the wolf pack gradually converges to the optimal solution. When $|A| > 1$, the wolf pack stays away from the prey and performs a global search to avoid falling into the local optimum.

2.2 Counterfactual explanation

Counterfactual explanations are a method for making machine learning models more transparent by showing how to change attributes to obtain different results (Spreitzer et al., 2022). Cheng et al. (2020) introduced counterfactuals with a classic example: A person submitted a loan request but was rejected by the bank. If his credit score had been 700 instead of 600, his loan application would have been approved.

Counterfactual explanations are currently widely used in different fields, including medical diagnosis, decision reasoning, and artificial intelligence. Richens et al. (2020) have improved the application of machine learning in the field of medical diagnosis, especially in identifying rare diseases, by establishing a counterfactual causal diagnosis model. Prado-Romero et al. (2023) use counterfactual explanations to provide a way to understand model decisions by providing specific changes in input features to explain the model's decision-making process. In addition, counterfactual explanations also have many applications in brain networks. For example, in the study of Abrate and Bonchi (2021), they proposed an explanation method for a black-box graph classifier for brain network classification. By analyzing counterfactual graphs, brain region connection patterns associated with specific brain region diseases can be identified. Matsui et al. (2022) proposed a new generative deep neural network (DNN) called Counterfactual Activation Generator to provide counterfactual explanations for DNN-based brain activation classifiers.

Counterfactual explanation has emerged as an important branch in the field of machine learning interpretability; however, it has not yet been applied to FC analysis. In this work, we introduce a counterfactual perspective: if the abnormal FC between brain regions in SZ patients is adjusted toward the normal range, their

predicted state may shift closer to that of healthy individuals. Such counterfactual reasoning is particularly valuable in the medical domain, as it can assist clinicians in evaluating the potential impact of different treatment strategies, especially in the context of brain diseases.

3 Materials and methods

3.1 Schizophrenia dataset

In this study, five public datasets are used, including the Center for Biomedical Research (COBRE) dataset (120 subjects), the Huaxi dataset (311 subjects), the Nottingham dataset (68 subjects), the Taiwan dataset (131 subjects) and the Xiangya dataset (143 subjects). All subjects met the following conditions: (i) no other Diagnostic and Statistical Manual of Mental Disorders (DSMIV) disease exists, (ii) no history of drug abuse, (iii) no clinically significant head trauma. The specific information of the subjects is presented in Table 1.

3.2 Data pre-processing

The rs-fMRI data of the five datasets are collected by different types of scanners, including COBRE and Xiangya by 3-T Siemens Tim-Trio scanner with an eight or 12-channel head coil, Huaxi by 3-T General Electric MRI scanner, and Nottingham by 3-T Philips Achieva MRI scanner. The rs-fMRI data are preprocessed using the program standard procedures of SPM 8 and the Data Processing Assistant for Resting-State fMRI (DPARSF). The following steps are performed: (i) removing the first 10 volumes, (ii) slice timing correction, (iii) head motion correction, (iv) regress out the nuisance covariates, (v) normalized to standardized space, (vi) voxel-wise bandpass filtering, (vii) normalization of anatomical images to MNI template space, and (viii) smoothing with a 4 mm Full Width at Half Maximum (FWHM) Gaussian kernel. After processing, we defined the nodes of the brain network according to the Automatic Anatomical Labeling (AAL) template, and calculated the pairwise similarities between the noded1s of the time series as the connecting edges of the brain network.

Next, let $A_i^F \in \mathbb{R}^{N\times N}$ be the connectivity matrix of the functional brain network, N be the number of regions of the brain network, $i = 1, 2, ..., p$, and p be the number of subjects. We take the upper triangular elements of the matrix as features and represent them as vectors $S_i = (s_i^1, ..., s_i^j, ..., s_i^q) \in \mathbb{R}^{1\times q}$, $q = \frac{N(N-1)}{2}$, s_i^j represents the j-th feature of the i-th subject, and $Y_i \in \mathbb{R}$ is the label of the i-th subject. It is worth noting that in this paper, we divided the brain network into 90 regions of interest (ROI), that is, $N = 90$, so each subject contains a vector of dimension $1 \times 4{,}005$, which reflects the functional connectivity strength pattern between the 90 brain regions of the subject.

3.3 Robust multi-task feature selection

3.3.1 Multi-task generation

To identify the most critical FC features for brain disease diagnosis, we use the infinite feature selection (IFS) (Roffo et al., 2020) method to calculate the importance of each feature and rank the features accordingly. Specifically, the weight of each feature is calculated based on the linear weighting of the following three aspects (i.e., Fisher criterion h_j, mutual information m_j, and standard deviation σ_j). The first is the Fisher criterion:

$$h_j = \frac{|\mu_{j,1} - \mu_{j,2}|^2}{\sigma_{j,1}^2 + \sigma_{j,2}^2} \tag{5}$$

where $\mu_{j,g}$ and $\sigma_{j,g}$ represent the mean and standard deviation of the j-th feature in the g-th class, respectively. In our experiments, both are binary classifications, so $g \in \{0, 1\}$.

The second is the normalized mutual information m_j between feature s^j and class label Y:

$$m_j = \sum_{y\in Y}\sum_{z\in s^j} u(z, y) log(\frac{u(z, y)}{u(z)u(y)}) \tag{6}$$

where Y is the set of class labels and $u(\cdot)$ represents the joint distribution probability.

TABLE 1 Characteristics of subjects in the five datasets in this study.

Datasets	Class	Gender (M/F)	*P*-value of gender	Age (years)	*P*-value of age
COBRE	NC	46/21	0.1927	34.82+11.28	0.3987
	SZ	42/11		36.75+13.68	
Huaxi	NC	79/71	0.6748	27.80+12.50	1.000
	SZ	80/81		27.80+12.50	
Nottingham	NC	26/10	0.2277	33.38+8.98	0.9855
	SZ	27/5		33.34+9.05	
Taiwan	NC	25/37	0.2329	29.87+8.62	0.2847
	SZ	35/34		31.59+9.60	
Xiangya	NC	35/25	0.9333	27.17+6.64	0.1025
	SZ	49/34		23.37+7.83	

NC, normal control; SZ, schizophrenia.

The third is the standard deviation σ_j, which reflects the dispersion of feature s^j in the sample.

The final weight of each feature s_i is calculated as follows:

$$s_i = \alpha_1 \cdot h_j + \alpha_2 \cdot m_j + \alpha_3 \cdot \sigma_j. \tag{7}$$

where $\alpha_1+\alpha_2+\alpha_3 = 1$, this weighting approach allows us to flexibly adjust the contribution of each indicator in the selection of features, thus selecting the most informative features for the diagnosis of schizophrenia (SZ).

Based on the preliminary evaluation of FC feature importance based on the above three factors, we further constructed a feature weight curve and optimized the FS process by introducing a knee point detection algorithm, following the knee point detection method proposed by Chen et al. (2021). This approach provides an automated criterion for determining the optimal feature subset size. Specifically, after obtaining the weight of each feature, we first construct a straight line connecting the starting point and the end point of the weight curve, and then calculate the vertical distance from each point on the curve to the straight line. The knee point (x_{knee}, y_{knee}) is the point that maximizes the distance:

$$(x_{knee}, y_{knee}) = \arg\max_j \left(\frac{|y_j - (ax_j + b)|}{\sqrt{a^2 + 1}}\right) \tag{8}$$

where a and b are the slope and intercept of the straight line determined by the starting point and the end point, (x_j, y_j) is the coordinate of the j-th feature point on the curve, $j = 2, 3, ..., q-1$. The identified knee points divide the feature weight curve into multiple intervals, and the features in each interval are given different priorities according to their weights.

Based on the location of the knee points, as shown in Figure 1b, we divide the features into three categories:

(i) Core features: located before the first knee point. These features are usually highly correlated with the predicted target variable and have low redundancy, and contribute the most to the model's predictive ability.

(ii) Important features: located between the two knee points. Although these features are not as important as the core features, may still contain useful information for specific scenarios. When combined with other features, they can enhance overall model performance, especially in complex cases where feature interactions are significant.

(iii) Remaining features: located after the second knee point. These features contribute less to the prediction task, contain redundant information, or have low correlation with the target variable.

After the above steps, we further use this category information to guide the task generation process. To ensure that the feature extraction process not only reflects its relative importance but also maintains appropriate diversity, we adopt a probabilistic extraction method based on feature weights. Specifically, we determine the initial selection probability of each feature based on the feature weight.

$$P_j = \frac{\omega_j}{\sum_{j=1}^{q} \omega_j} \tag{9}$$

where ω_j is the weight of the j-th feature. The larger ω_j is, the higher its initial extraction probability is, and thus it is given priority in FS. To ensure that all features have a certain chance of being selected and to avoid the extraction probability of low-weight features becoming too small, we adjust the initial probability:

$$P'_j = \frac{P_j}{\max(P_j)} \tag{10}$$

The above formula ensures that the maximum extraction probability of a feature is 1, and the extraction probabilities of all other features are adjusted proportionally, avoiding excessive neglect of low-weight features while still maintaining the priority of high-weight features during extraction.

During the task generation process, a random number λ between 0 and 1 is first randomly generated, which is used to determine which features will be selected for the current task. For each feature s^j, if $\lambda \leq P'_j$, the feature will be selected for the current task. As shown in Figure 1b, after n rounds of independent extraction, n different task sets are generated, each of which contains a set of selected feature subsets. This mechanism ensures that high-weight features are selected first and fully retain the potential contribution of low-weight features, thereby effectively improving the diversity and flexibility of the task generation process.

3.3.2 Multi-task optimization with GWO

In multi-task optimization, we propose to combine the knowledge transfer mechanism with the GWO-based multi-task optimization method to enhance information sharing between different tasks, thereby improving the efficiency and effect of overall optimization. Specifically, we directly integrate the knowledge transfer mechanism in the initialization phase of GWO to make full use of the optimization experience of existing tasks.

To achieve effective knowledge transfer, in the multi-task optimization process, we first need to quantify the importance of each feature in the previous task. In other words, we need to calculate the cumulative number of times Q_{KT} that feature s^j is selected in all previous tasks:

$$Q_{KT}(s^j) = \sum_{t=1}^{n} Q_{KT}^t(s^j) \tag{11}$$

where n represents the total number of tasks, $Q_{KT}^t(s^j)$ represents whether the feature is selected in the t-th task (if selected, it is 1, otherwise it is 0). Then, calculate the probability $P(s^j)$ of feature s^j being selected in the initial population of the new task:

$$P(s^j) = \frac{Q_{KT}(s^j)}{\sum_{j=1}^{q} Q_{KT}(s^j)} \tag{12}$$

The above formula converts the historical performance of the feature into a probability value, which will be directly applied to initialize the wolf pack:

$$G_{wo} = \begin{cases} 1, & \lambda \leq P(s^j) \\ 0, & \lambda > P(s^j) \end{cases} \tag{13}$$

where the random number $\lambda \in [0, 1]$, the feature s^j is selected only when it is less than or equal to $P(s^j)$. For ease

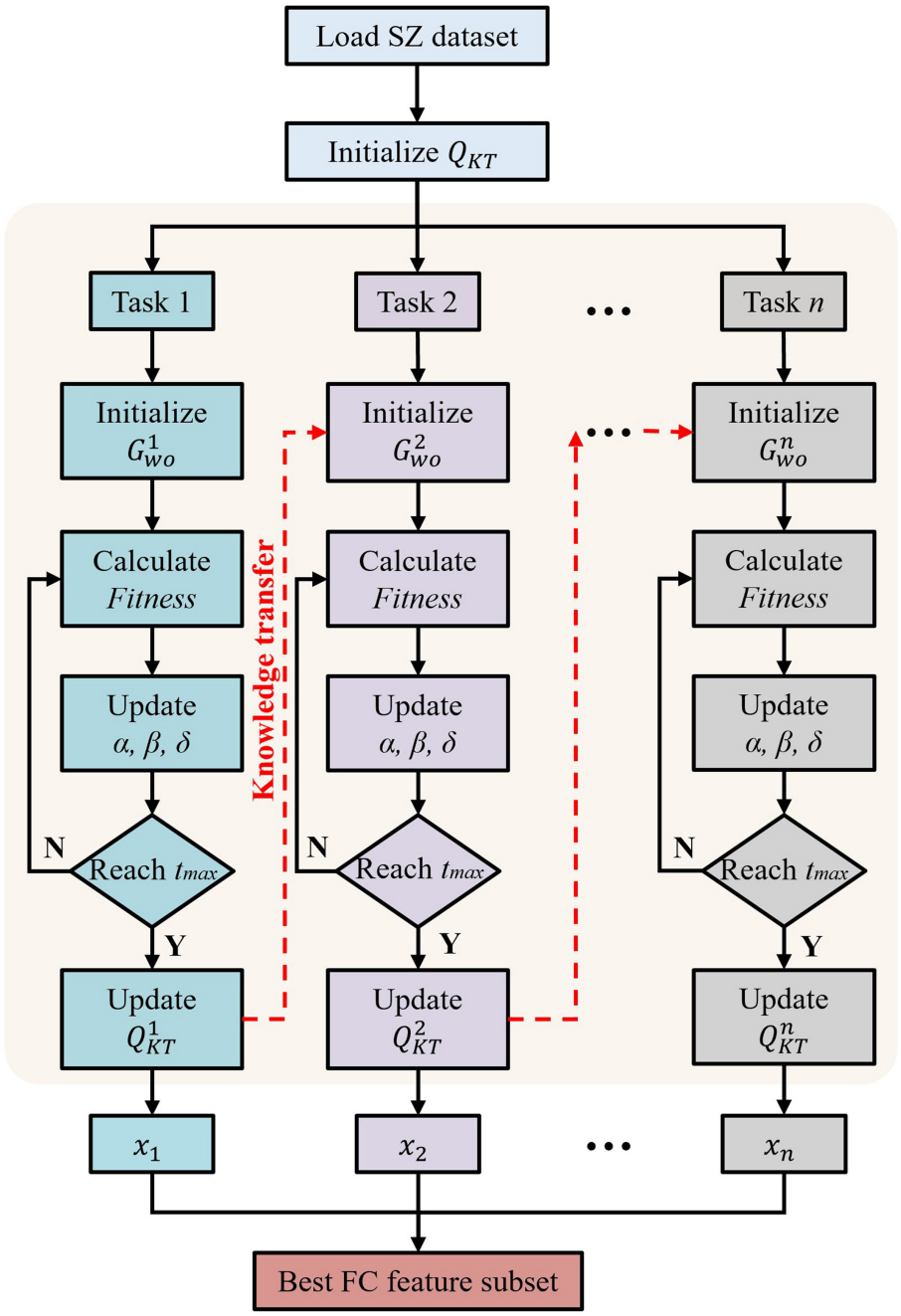


FIGURE 2
Flowchart of proposed multi-task optimization with GWO.

of understanding, we show the specific process of the proposed multi-task optimization method in Figure 2. First, the global environment is set. Subsequently, the algorithm enters a loop and processes n tasks in turn. For each task, the wolf pack is initialized independently, using the global knowledge of the previously processed tasks to provide information for the initial state of the search for the new task. The position of the wolf is iteratively updated to optimize the FS problem. After optimization, the best solution is used to update the global knowledge base. This cycle is repeated for each task, ensuring the continuous flow of information and the improvement of the solution. Finally, n feature subsets $(x_1, x_2, ..., x_n)$ are obtained from the n tasks.

In addition, to minimize the number of selected features while maintaining a high classification accuracy, we designed a fitness function in multi-task optimization and introduced a penalty term to constrain the number of features:

$$Fitness = \rho \times ACC - (1 - \rho) \times \frac{q_{sf}}{q} \tag{14}$$

where ρ is a weight coefficient, which ranges between $[0, 1]$ and is used to balance the classification accuracy ACC and the number of selected features q_{sf}.

After the above operations, we represent the selected feature matrix as $S' \in \mathbb{R}^{p\times k}$, where $k \ll q$. Based on the selected feature matrix S', we can train a suitable machine learning model [i.e., $f(\cdot)$] to predict schizophrenia. In our experiment, since the support vector machine (SVM) is strongly adaptable to small sample data sets, we used SVM as the classification model.

3.3.3 Diversity counterfactual explanation

To enhance the interpretability of our method, we further introduce a counterfactual explanation model (Mothilal et al., 2020) to generate sample-level explanations. The input of this model includes a trained SVM model [i.e., $f(\cdot)$] and the feature vector $c_i \in \mathbb{R}^{1\times k}$ of the i-th subject. Our goal is to generate a set of counterfactual examples $\{x_i^1, x_i^2, ..., x_i^L\}$ for subject i such that its decision outcome $x_i^l \in \mathbb{R}^{1\times k}$ is different from the prediction of the original feature vector c_i.

The counterfactual explanation model consists of three parts: loss function $loss(\cdot)$, distance function $dist(\cdot)$, and diversity metric $diversity(\cdot)$. Specifically, the first part pushes counterfactual x_i^l toward different predictions, the second part makes counterfactual examples closer to the original input, and the third part is used to increase the diversity of counterfactual explanations. In the first part, we use a hinge loss function that helps generate counterfactuals with less variation by reducing the preference for extreme values. The hinge loss is expressed as follows:

$$loss_{hinge} = max(0, 1 - z \cdot logit(f(x)) \tag{15}$$

where z is 1 when $\hat{Y} = 1$ and -1 when $\hat{Y} = 0$, and $logit(f(x))$ is the unscaled output of the SVM model. It is worth noting that in our experiments, 1 corresponds to normal subjects and 0 corresponds to patients, so in the verification of converting patients into normal subjects, $\hat{Y}$ is usually set to 1. For the choice of distance function in the second part, we follow Wachter et al. (2017) proposal and divide the distance of each feature by the median absolute deviation (MAD) of the feature values in the training set:

$$dist(x, c) = \frac{1}{L}\sum_{\alpha=1}^{L}\frac{|x^\alpha - c^\alpha|}{MAD_\alpha} \tag{16}$$

where MAD_α is the median absolute deviation of the α-th feature, L is the total number of counterfactual examples to generate, x represents the counterfactual example and c represents the original feature vector. For the third part, we use a determinant-based point procedure to measure the diversity of counterfactual examples, computed by the determinant value of its kernel matrix K:

$$diversity = det(K) \tag{17}$$

where $K_{u,v} = \frac{1}{1+dist(x^u, x^v)}$, x^v and x^u represent two counterfactual examples. In the experiments, to avoid uncertain determinants, we add small random perturbations on the diagonal elements to calculate the determinant.

Finally, we can obtain counterfactual examples by optimizing the following loss:

$$\begin{aligned} X(c_i) = {} & \frac{\gamma_1}{L}\sum_{l=1}^{L} dist(x_i^l, c_i) \\ & -\gamma_2 diversity(x_i^1, x_i^2, ..., x_i^L) \\ & +\arg\min_{x_i^1, x_i^2, ..., x_i^L} \frac{1}{L}\sum_{l=1}^{L} loss_{hinge}(f(x_i^l), \hat{Y}) \end{aligned} \tag{18}$$

where $X(c_i)$ is the final counterfactual explanation model, γ_1 and γ_2 are hyperparameters for balancing the three parts of the loss function. The above formula reveals the minimum change required for the input data to achieve the idealized result. By adjusting the FC values between abnormal brain regions of SZ patients, their state may be closer to normal. This method not only provides an intuitive explanation scheme, but also provides SZ patients and doctors with the guidance needed to treat the disease.

4 Experiments and results

4.1 Experimental setting

In this work, we use a support vector machine (SVM) classifier to perform the classification task on five SZ datasets. During the experiments, we evaluate the performance of different methods based on diagnostic accuracy (ACC $= \frac{TP+TN}{TP+TN+FP+FN}$), sensitivity (SEN $= \frac{TP}{TP+FN}$) and specificity (SPE $= \frac{TN}{TN+FP}$). FP, TP, FN, and TN represent false-positive, true-positive, false-negative, and true-negative classification results. To ensure fairness, all compared FS methods use SVM classifiers. The parameters of our method are set as $\alpha_1 = \alpha_2 = 0.4$, $\alpha_3 = 0.2$, $t_{max} = 100$, $\rho = 0.9$, $n = 8$, $L = 10$, $\gamma_1 = 0.5$ and $\gamma_2 = 1$. It is worth noting that we use a five-fold cross-validation strategy in all experiments.

4.2 Statistical analysis of FC features

In this set of experiments, we perform statistical analysis on the functional connectivity (FC) remaining after feature selection by our method to demonstrate the effectiveness of our method. For intuitiveness, we first show the FC features retained after feature selection by our method in Figure 3. As can be seen from Figure 3, there are 16 shared FCs in the five datasets, and these shared FCs are selected as features in different datasets, indicating that they are crucial in identifying SZ. In addition, these shared FCs are mainly distributed in key brain regions such as the prefrontal cortex (PFC), cingulate gyrus (CC), and hippocampus (HIP), which is consistent with the findings of existing studies on SZ in brain network abnormalities (Orellana and Slachevsky, 2013; Wei et al., 2021; Frankle et al., 2022; Haznedar et al., 2004).

We select the five most statistically significant FC values between SZ and NC based on the statistical significance of each dataset, and the results are shown in Figure 4. From Figure 4, we find that the FC values between SZ and NC show different distribution patterns in the five datasets. Specifically, in some datasets, the FC values of SZ patients are significantly higher than those of NC, while in other datasets, the FC values of SZ patients are

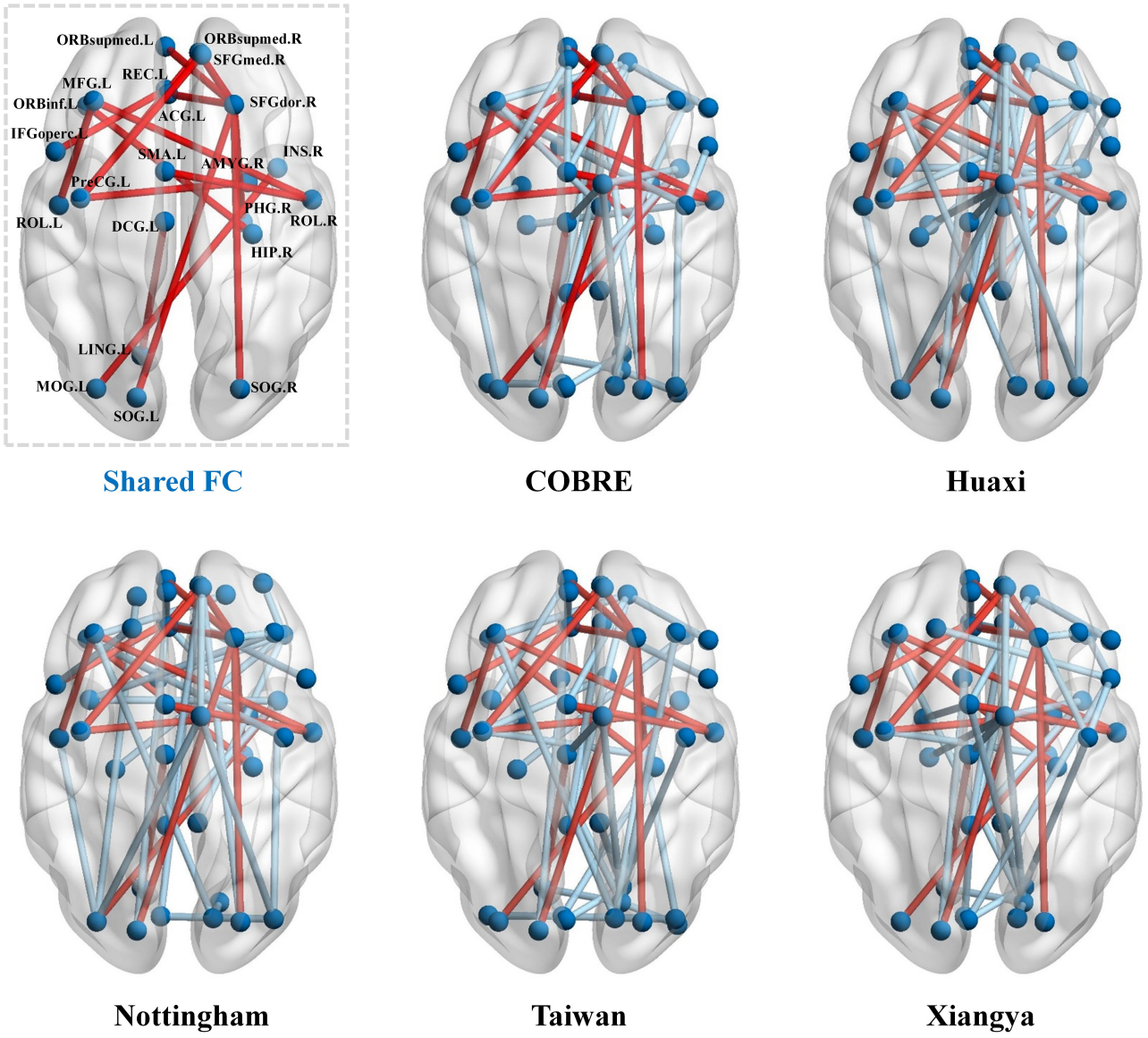


FIGURE 3
Functional connectivity (FC) retained after feature selection by our method in the five datasets. The red lines indicate the common functional connectivity among the five datasets.

significantly lower than those of NC. This suggests that there may be some heterogeneity in the functional connectivity patterns of SZ patients in different datasets. However, although the distribution of FC values in different datasets is different, some specific FCs show significant differences in multiple datasets, indicating that these FCs may play a key role in the neural mechanism of SZ.

Overall, the above results show that our method effectively extracts stable and biologically meaningful FC features, which helps to improve the accuracy and interpretability of SZ classification.

4.3 Comparison methods

We compare our proposed method with seven methods, including (i) RAW: classification without feature selection, as a baseline to illustrate the effect of applying feature selection techniques. (ii) LASSO: Lasso regression model based on L1 regularization (Cui et al., 2021). (iii) MFCSO: Multitasking Feature Selection via Competitive Swarm Optimizer (Li L. et al., 2023). (iv) MOEA\D: Multi-Objective Evolutionary Algorithm based on Decomposition (Wang et al., 2021). (v) SPEA: Strength Pareto Evolutionary Algorithm (Jiang and Yang, 2017). (vi) PSO-MET: Evolutionary Multitasking-Based Feature Selection via Particle Swarm Optimization (PSO) (Chen et al., 2020). (vii) MTPSO: Multitasking feature selection via PSO (Chen et al., 2021).

For all the above methods, the hyperparameters were set according to the values recommended in their respective original papers. Additionally, the number of iterations for all methods was set to 100, ensuring a consistent and fair comparison across all approaches.

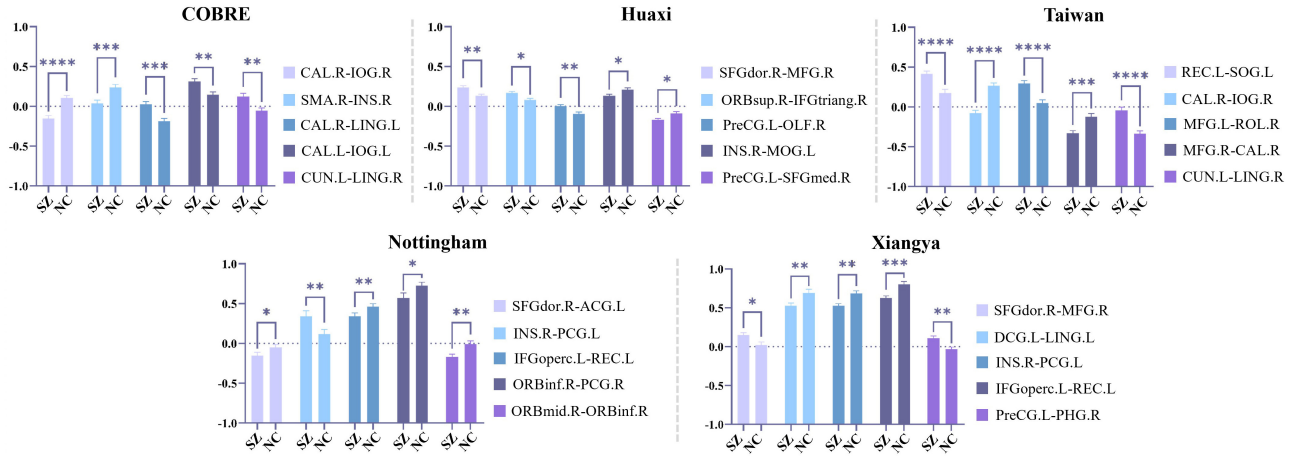


FIGURE 4
Statistical analysis is performed on the functional connections (FC) retained by feature selection using our method, and the five most statistically significant FC values between SZ and NC in the five datasets are shown here. Among them, * indicates $0.01 < p < 0.05$, ** indicates $0.001 < p < 0.01$, *** indicates $0.0001 < p < 0.001$, and **** indicates $p < 0.0001$.

MFCSO uses three filter methods for multi task feature selection, with each task optimized as an independent task without direct correlation between them. Therefore, the feature selection process may lack consistency. When dealing with specific datasets, especially on the schizophrenia (SZ) dataset, MFCSO may not be able to ensure consistency of selected features across different tasks, which may result in unstable performance on different datasets. Due to the lack of inter task correlation, feature selection results may be affected by randomness, making it difficult to effectively capture stable features related to schizophrenia.

Multi-objective evolutionary algorithms, such as MOEA\D and SPEA, are designed to address multiple objectives in feature selection. These algorithms provide a better balance between accuracy and feature diversity by considering multiple criteria in the optimization process. However, they are computationally intensive and can be prone to converging to local optima, especially in high-dimensional spaces. Furthermore, they often struggle with the trade-off between model complexity and accuracy, which can result in overfitting in small-sample scenarios, limiting their generalization ability.

PSO-MET and MTPSO are both particle swarm optimization-based methods that aim to improve feature selection by leveraging the concept of multitasking. While these methods are effective at identifying relevant features in some cases, they tend to be overly sensitive to initial conditions and parameter settings, leading to performance fluctuations. The lack of consistency across tasks and datasets reduces their reliability, particularly in real-world clinical settings where the data may be noisy or heterogeneous.

In comparison, our proposed method integrates robust multi-task feature selection with counterfactual explanation, offering several advantages over the methods discussed above. By using the Gray Wolf Optimizer (GWO) for feature selection, we ensure that our method not only handles high-dimensional data efficiently but also maintains stability across different datasets. The multi-task learning framework in our method allows for the sharing of knowledge across tasks, which improves generalization and reduces the risk of overfitting, particularly in small-sample situations.

4.4 Parameter analysis

In this section, we investigate the impact of varying the number of tasks on the performance of our multi-task optimization framework, as shown in the Figure 5. We observe that increasing the number of tasks generally leads to improvements in classification accuracy, especially for datasets such as Taiwan and Xiangya. These datasets achieve their highest classification accuracy at around six–nine tasks, where the accuracy reaches 0.87 and 0.89, respectively. This indicates that knowledge sharing between tasks is particularly effective in enhancing model performance when the task number is moderate. However, beyond a certain point, specifically around 10–12 tasks, the performance begins to plateau, with only marginal improvements in classification accuracy. The graph clearly shows that the datasets, such as Xiangya and Nottingham, while still improving with increasing task numbers, experience diminishing returns as the number of tasks exceeds 10. This suggests that while task number does play a role in boosting performance, there is an optimal task count that provides the best trade-off between performance enhancement and computational cost.

A deeper analysis reveals that the knowledge sharing between tasks is highly beneficial for improving classification performance. As the number of tasks increases, the model can leverage a broader range of features, which enhances its ability to generalize. However, once the number of tasks exceeds a threshold, redundancy starts to creep into the shared knowledge. This results in the transmission of features that do not contribute significantly to the performance improvement, thereby leading to a less efficient model. The redundancy of features becomes particularly evident when the number of tasks increases beyond 10, where the

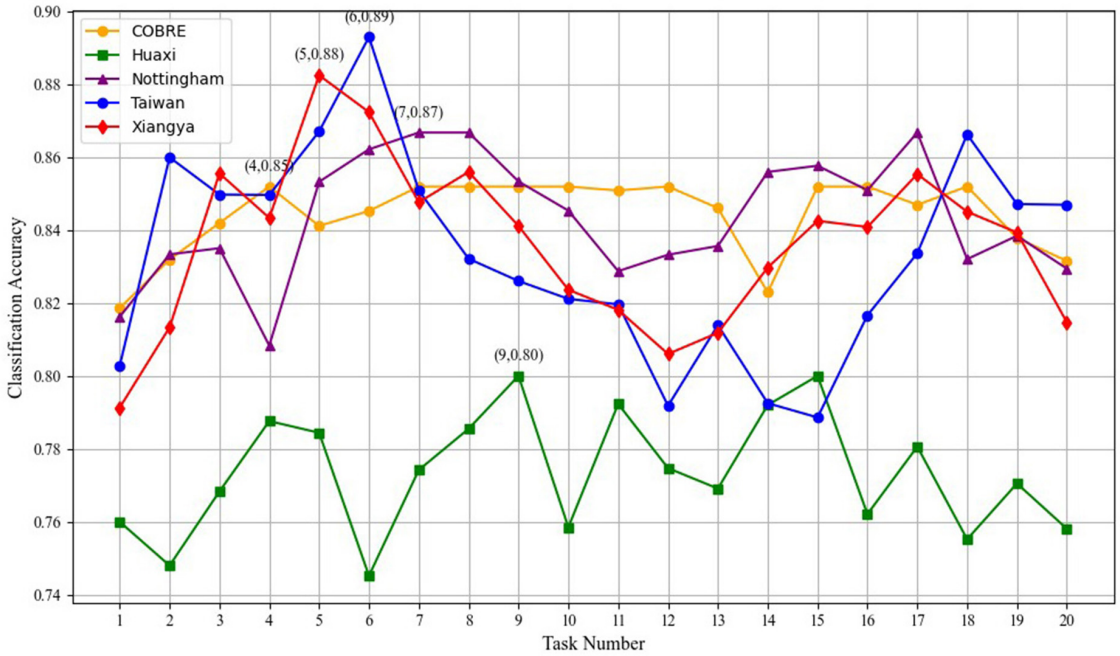


FIGURE 5
Impact of varying task numbers on model performance.

TABLE 2 Classification performance comparison with existing methods.

Datasets	Metric	RAW	LASSO	MFCSO	MOEA\D	SPEA	PSO-MET	MTPSO	Our method
COBRE	ACC (%)	63.41	75.00	68.10	73.40	69.44	78.38	81.19	**85.19**
	SEN (%)	58.33	66.67	60.00	76.47	78.57	68.75	**83.93**	80.00
	SPE (%)	73.68	79.17	76.19	69.77	63.64	85.71	79.45	**91.67**
Huaxi	ACC (%)	61.29	69.89	72.31	76.60	77.66	75.53	76.74	**80.00**
	SEN (%)	55.56	70.83	64.57	80.39	80.85	69.39	78.55	**82.86**
	SPE (%)	71.43	68.89	74.81	72.09	74.47	**82.22**	74.60	76.67
Nottingham	ACC (%)	65.00	66.12	72.22	72.34	75.53	80.95	82.71	**86.67**
	SEN (%)	66.67	68.97	66.67	74.51	76.60	80.00	82.13	**85.71**
	SPE (%)	63.64	62.50	77.78	69.77	74.47	81.82	83.08	**87.50**
Taiwan	ACC (%)	70.21	79.49	77.32	77.50	80.00	85.00	81.55	**89.29**
	SEN (%)	74.47	77.27	73.68	73.68	**88.24**	80.95	78.26	87.50
	SPE (%)	65.96	82.35	80.95	80.95	73.91	89.47	84.52	**91.67**
Xiangya	ACC (%)	66.90	69.23	79.41	76.74	70.77	72.31	82.79	**88.24**
	SEN (%)	51.35	58.82	72.22	72.00	74.29	68.57	**83.58**	80.00
	SPE (%)	67.39	69.73	87.50	83.33	66.67	76.67	81.79	**94.74**

Bold values represent the optimal values.

performance gains start to level off, and the computational overhead grows significantly.

Thus, while task quantity is crucial for leveraging task interdependencies and improving model accuracy, an excessive number of tasks may lead to inefficiency due to the sharing of redundant or less informative features. Therefore, it is essential to strike a balance between the number of tasks and the computational cost to ensure the model remains both effective and efficient.

4.5 Classification performance

In this set of experiments, we compare our proposed method with seven methods and show the results in Table 2. It is not difficult to see that our method shows excellent stability and consistency on the five datasets. Specifically, in the five datasets, the ACC of our method reaches 85.19% (COBRE), 80.00% (Huaxi), 86.67% (Nottingham), 89.29% (Taiwan), and 88.24% (Xiangya), while the ACC of most methods does not exceed 85%. Secondly, our method

performs outstandingly in both SEN and SPE, with SPE reaching 94.74% on the Xiangya dataset and SEN reaching 82.86% on the Huaxi dataset, indicating that our method has strong stability in the ability to distinguish between positive and negative samples. PSO-MET and MTPSO perform well in terms of SEN. For example, in the COBRE dataset, the SEN of MTPSO is 83.93%, which is higher than other methods, indicating that it has a strong ability to identify positive samples. In addition, we find that the methods based on multi-task optimization and evolutionary algorithms (i.e., PSO-MET and MTPSO) perform better overall. For example, in the Xiangya dataset, the ACC of MTPSO reaches 82.79%, which is significantly higher than other methods. This can be attributed to the fact that multi-task methods utilize shared knowledge across tasks, thereby improving the overall learning process. In general, the methods based on multi-task optimization and evolutionary algorithms have higher accuracy in SZ identification, while our method shows even better performance.

In addition, for the statistical significance of model performance, we select the three best-performing comparison methods (SPEA, PSO-MET, and MTPSO) in the experiment, and perform paired t-tests on the ACC indicators of each method on multiple datasets. The results are shown in Table 3. As can be seen from Table 3, our proposed method shows statistically significant differences with the three comparison methods on all datasets ($p < 0.05$). Specifically, the comparison with the SPEA method shows extremely significant differences on the COBRE, Nottingham, and Xiangya datasets ($p < 0.005$), and the comparison with PSO-MET has p values less than 0.025 on all datasets, indicating that the differences are highly statistically significant. At the same time, compared with the MTPSO method, although the p values in some datasets (such as Huaxi and COBRE) are relatively high, they do not exceed the significance level ($p < 0.05$), which still shows the stable advantages of our method on various datasets. These results further verify the universality and effectiveness of our method on multiple datasets from a statistical perspective.

4.6 Counterfactual explanations

In this set of experiments, we demonstrate how to generate a set of intuitive and diverse counterfactual (CF) examples for patients through the counterfactual explanation model. We provide counterfactual explanations by fine-tuning the abnormal FC value changes of patients, that is, adjusting the FC values between specific regions to make the patient's state closer to that of normal people. We generate two different counterfactual examples for SZ patients and present them in the form of brain maps and heat maps, as shown in Figure 6. It is not difficult to see that we can make the patient's state close to normal by only slightly adjusting the FC values between the corresponding regions. Specifically, in the Huaxi dataset, CF1 increases the FC values between ORBinf.R–HIP.L, SMA.R–SFGmed.R, SFGmed.L–ORBsupmed.L, and SMA.R–PHG.L from –0.2994, 0.0043, 0.2313, and 0.6822 to 0.1712, 0.8632, 0.2981, and 1.2072, and decreases the FC values between MFG.L–ROL.R and SFGdor.R–MFG.R from 0.1875 and 0.4143 to –0.6375 and –0.4230. In the Xiangya dataset, CF1 decreases the FC values between MFG.L–ROL.R, SFGdor.R–SOG.R, SFGdor.R–ACG.L, ORBsup.R–IFGtriang.R, and CUN.L–LING.R from 0.2149, 0.0883, –0.0146, –0.3282, and –0.0603 to –0.5490, 0.0619, –0.4669, –0.4412, and –0.8791, and increases the FC values between ORBsup.R–PCG.L and INS.R–PCG.L from –0.1435 and 0.4575 to 0.6884 and 1.2428, respectively. We find that the changes in functional connectivity (FC) after counterfactual interpretation remain stable within 1, without large-scale fluctuations, which further illustrates the robustness of our method. In addition, the role of FC changes in SZ patients has been observed in a large number of studies, such as Lynall et al. (2010), Fornito and Bullmore (2015), and Li et al. (2017).

TABLE 3 The t-test p-value results of our method and the three best performing comparison methods (SPEA, PSO-MET and MTPSO) on ACC.

Datasets	SPEA/our	PSO-MET/our	MTPSO/our
COBRE	0.0015	0.0220	0.0490
Huaxi	0.0439	0.0133	0.0269
Nottingham	0.0037	0.0019	0.0249
Taiwan	0.0143	0.0195	0.0174
Xiangya	0.0016	0.0029	0.0428

5 Discussion

In this paper, we propose a multi-task feature selection method for SZ diagnosis, and combine it with the counterfactual explanation model to fine-tune the abnormal FC features of SZ patients to make their state closer to that of healthy individuals, thereby improving the accuracy of SZ classification and the interpretability of the model. To demonstrate the effectiveness of our method, we conduct empirical studies on five SZ datasets. Our results show that across the five datasets, 16 FC features are selected simultaneously. These shared FC features are mainly distributed in key brain regions such as the prefrontal cortex (PFC), cingulate gyrus (CC) and hippocampus (HIP), which are widely considered to be closely related to the pathological mechanism of SZ in previous studies. For example, the study by Minzenberg et al. (2009) shows that PFC dysfunction is closely related to executive function deficits in SZ patients. Whitfield-Gabrieli et al. (2009) find that SZ patients have significant abnormalities in FC in the default mode network (including CC), which is associated with cognitive dysfunction. Gangadin et al. (2021) and Li X.-W. et al. (2023) find that SZ patients have significant abnormalities in FC between HIP and other brain regions in the resting state. These results not only verify that the abnormal FC features screened out by our method under multiple datasets are consistent and stable, but also further confirm its potential value in the diagnosis and interpretation of SZ from a neurobiological perspective.

Although previous studies reveal a variety of brain FC abnormalities associated with SZ, there is still a lack of an interpretable diagnostic tool in the diagnosis of SZ. Our study proposes an innovative method that integrates multi-task feature selection and counterfactual explanation. To generate accurate counterfactual examples, we construct a

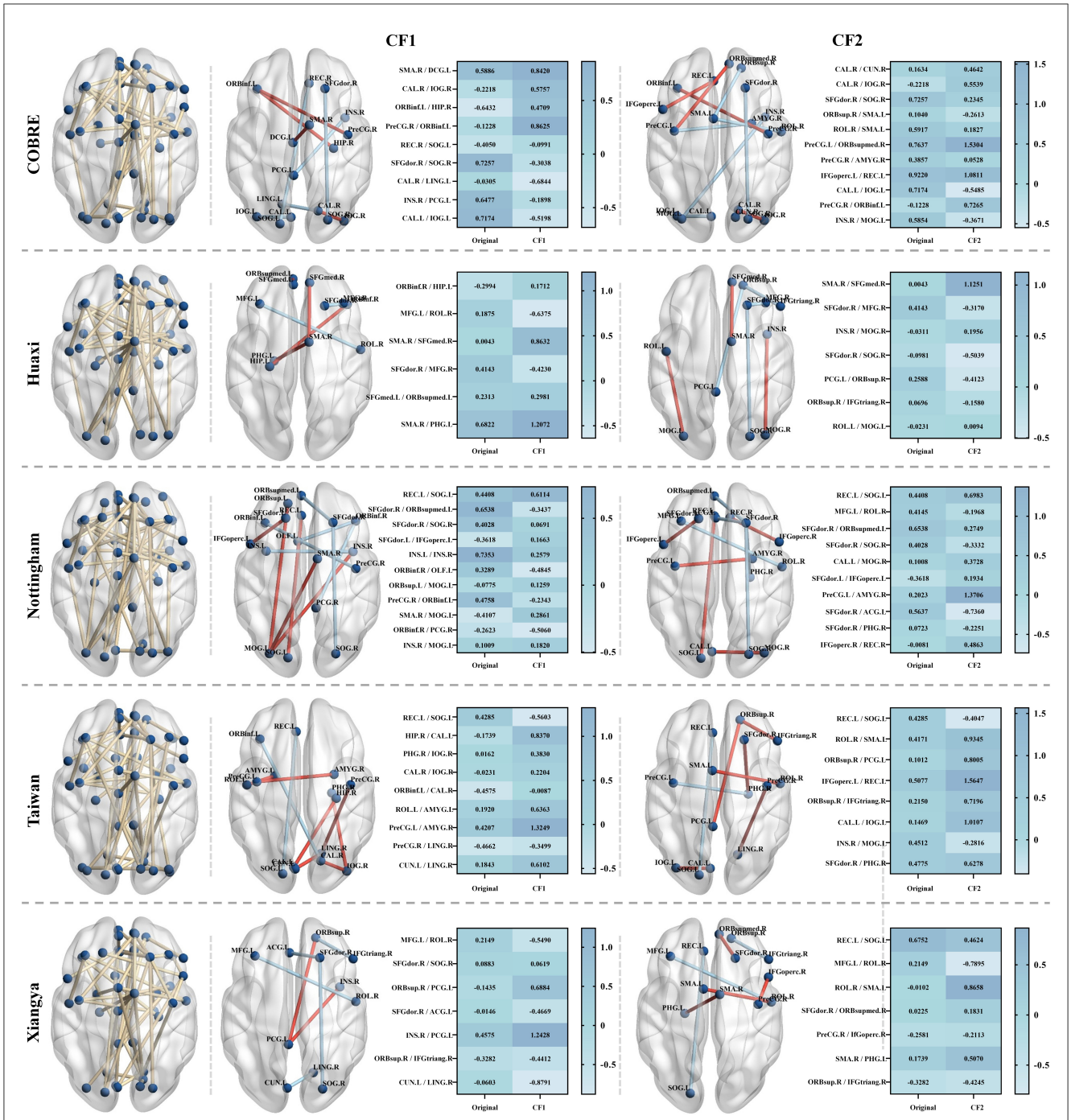


FIGURE 6
Counterfactual (CF) examples generated for randomly selected SZ patients in the test set in the five datasets. The yellow lines in the left brain map indicate the remaining functional connectivity (FC) after feature selection by our method, and the blue nodes indicate the corresponding brain regions. The brain maps in CF1 and CF2 show two counterfactual examples generated for the abnormal FC of the patients. The red lines indicate an increase in the FC value between the corresponding regions, and the blue lines indicate a decrease in the FC value. The heat map on the right shows the original FC value between the corresponding brain regions of the patients and the FC value after the counterfactual explanation.

counterfactual explanation model through three parts: loss function *loss*(·), distance function *dist*(·), and diversity index *diversity*(·). Specifically, *loss*(·) pushes counterfactual examples toward different predictions, *dist*(·) brings the counterfactual example closer to the original input, and *diversity*(·) increases the diversity of counterfactual explanations. We capture the brain regions where patients show abnormal FC features and slightly adjust the FC values between abnormal brain regions to make them closer to the normal state. This analysis method not only improves the interpretability of the classification model, but also provides an intuitive individual-level explanatory perspective for understanding brain FC abnormalities in SZ

patients, which helps to identify potential intervention targets and promotes the application of precision medicine in the diagnosis of SZ.

However, the current study still has several limitations. First, we only use the AAL model to define brain regions. In the future, we use different templates to evaluate the effectiveness of our proposed method. Second, we have not yet established cooperation with clinical medical institutions and lack counterfactual change explanations reviewed by clinicians. We plan to introduce clinical validation to further demonstrate the practicality and effectiveness of the method. Finally, this study focuses on the SZ dataset and further verifies the generalization ability and application potential of the method on other brain disease datasets such as Alzheimer's disease and autism.

6 Conclusion

In this paper, we propose a robust feature selection method based on multi-task optimization for SZ identification, and explain the changes in brain functional connectivity caused by the disease through a counterfactual explanation model. Compared with traditional methods, our proposed method not only improves the recognition performance, but also provides an intuitive explanation for the prediction of SZ, and verifies the effectiveness of the method on five SZ datasets.

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.

Ethics statement

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

XY: Investigation, Methodology, Writing – original draft, Writing – review & editing. SW: Data curation, Investigation, Writing – original draft. YS: Methodology, Writing – review & editing. LG: Validation, Writing – review & editing. YH: Formal analysis, Writing – review & editing. TC: Formal analysis, Writing – review & editing. HY: Resources, Software, Writing – review & editing. HR: Validation, Writing – review & editing.

Funding

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

Acknowledgments

We sincerely appreciate the researchers and institutions that provided the publicly available datasets used in this study, including COBRE, Huaxi, Nottingham, Taiwan and Xiangya. These datasets have greatly contributed to the advancement of schizophrenia research. Additionally, we acknowledge the efforts of all participants and staff involved in data collection and preprocessing.

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

Abrate, C., and Bonchi, F. (2021). "Counterfactual graphs for explainable classification of brain networks," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (New York, NY: ACM), 2495–2504. doi: 10.1145/3447548.3467154

Chan, Y. H., Girish, D., Gupta, S., Xia, J., Kasi, C., He, Y., et al. (2024). Discovering robust biomarkers of psychiatric disorders from resting-state functional MRI via graph neural networks: a systematic review. *arXiv* [Preprint]. arXiv:2405.00577. doi: 10.48550/arXiv.2405.00577

Chen, K., Xue, B., Zhang, M., and Zhou, F. (2020). An evolutionary multitasking-based feature selection method for high-dimensional classification. *IEEE Trans. Cybern.* 52, 7172–7186. doi: 10.1109/TCYB.2020.3042243

Chen, K., Xue, B., Zhang, M., and Zhou, F. (2021). Evolutionary multitasking for feature selection in high-dimensional classification via particle swarm optimization. *IEEE Trans. Evol. Comput.* 26, 446–460. doi: 10.1109/TEVC.2021.3100056

Cheng, F., Ming, Y., and Qu, H. (2020). Dece: decision explorer with counterfactual explanations for machine learning models. *IEEE Trans. Vis. Comput. Graph.* 27, 1438–1447. doi: 10.1109/TVCG.2020.3030342

Chyzhyk, D., Savio, A., and Graña, M. (2015). Computer aided diagnosis of schizophrenia on resting state fMRI data by ensembles of elm. *Neural Netw*. 68, 23–33. doi: 10.1016/j.neunet.2015.04.002

Cui, L., Bai, L., Wang, Y., Yu, P. S., and Hancock, E. R. (2021). Fused lasso for feature selection using structural information. *Pattern Recognit*. 119:108058. doi: 10.1016/j.patcog.2021.108058

Ding, W., Zhou, T., Huang, J., Jiang, S., Hou, T., Lin, C.-T., et al. (2024). FMDNN: a fuzzy-guided multi-granular deep neural network for histopathological image classification. *IEEE Trans. Fuzzy Syst*. 32, 4709–4723. doi: 10.1109/TFUZZ.2024.3410929

Fornito, A., and Bullmore, E. T. (2015). Reconciling abnormalities of brain network structure and function in schizophrenia. *Curr. Opin. Neurobiol*. 30, 44–50. doi: 10.1016/j.conb.2014.08.006

Frankle, W. G., Himes, M., Mason, N. S., Mathis, C. A., and Narendran, R. (2022). Prefrontal and striatal dopamine release are inversely correlated in schizophrenia. *Biol. Psychiatry* 92, 791–799. doi: 10.1016/j.biopsych.2022.05.009

Gangadin, S. S., Cahn, W., Scheewe, T. W., Pol, H. E. H., and Bossong, M. G. (2021). Reduced resting state functional connectivity in the hippocampus-midbrain-striatum network of schizophrenia patients. *J. Psychiatr. Res*. 138, 83–88. doi: 10.1016/j.jpsychires.2021.03.041

Haznedar, M. M., Buchsbaum, M. S., Hazlett, E. A., Shihabuddin, L., New, A., Siever, L. J., et al. (2004). Cingulate gyrus volume and metabolism in the schizophrenia spectrum. *Schizophr. Res*. 71, 249–262. doi: 10.1016/j.schres.2004.02.025

Hu, R., Peng, Z., Zhu, X., Gan, J., Zhu, Y., Ma, J., et al. (2021). Multi-band brain network analysis for functional neuroimaging biomarker identification. *IEEE Trans. Med. Imaging* 40, 3843–3855. doi: 10.1109/TMI.2021.3099641

Huang, J., Wang, M., Ju, H., Ding, W., and Zhang, D. (2025). Agbn-transformer: anatomy-guided brain network transformer for schizophrenia diagnosis. *Biomed. Signal Process. Control* 102:107226. doi: 10.1016/j.bspc.2024.107226

Insel, T. R. (2010). Rethinking schizophrenia. *Nature* 468, 187–193. doi: 10.1038/nature09552

Jiang, S., and Yang, S. (2017). A strength pareto evolutionary algorithm based on reference direction for multiobjective and many-objective optimization. *IEEE Trans. Evol. Comput*. 21, 329–346. doi: 10.1109/TEVC.2016.2592479

Li, L., Xuan, M., Lin, Q., Jiang, M., Ming, Z., Tan, K. C., et al. (2023). An evolutionary multitasking algorithm with multiple filtering for high-dimensional feature selection. *IEEE Trans. Evol. Comput*. 27, 802–816. doi: 10.1109/TEVC.2023.3254155

Li, T., Wang, Q., Zhang, J., Rolls, E. T., Yang, W., Palaniyappan, L., et al. (2017). Brain-wide analysis of functional connectivity in first-episode and chronic stages of schizophrenia. *Schizophr. Bull*. 43, 436–448. doi: 10.1093/schbul/sbw099

Li, X.-W., Liu, H., Deng, Y.-Y., Li, Z.-Y., Jiang, Y.-H., Li, D.-Y., et al. (2023). Aberrant intra-and internetwork functional connectivity patterns of the anterior and posterior hippocampal networks in schizophrenia. *CNS Neurosci. Ther*. 29, 2223–2235. doi: 10.1111/cns.14171

Lv, J., Jiang, X., Li, X., Zhu, D., Chen, H., Zhang, T., et al. (2015). Sparse representation of whole-brain fMRI signals for identification of functional networks. *Med. Image Anal*. 20, 112–134. doi: 10.1016/j.media.2014.10.011

Lynall, M.-E., Bassett, D. S., Kerwin, R., McKenna, P. J., Kitzbichler, M., Muller, U., et al. (2010). Functional connectivity and brain networks in schizophrenia. *J. Neurosci*. 30, 9477–9487. doi: 10.1523/JNEUROSCI.0333-10.2010

Matsui, T., Taki, M., Pham, T. Q., Chikazoe, J., and Jimura, K. (2022). Counterfactual explanation of brain activity classifiers using image-to-image transfer by generative adversarial network. *Front. Neuroinform*. 15:802938. doi: 10.3389/fninf.2021.802938

McCutcheon, R. A., Marques, T. R., and Howes, O. D. (2020). Schizophrenia—an overview. *JAMA Psychiatry* 77, 201–210. doi: 10.1001/jamapsychiatry.2019.3360

Mhiri, I., and Rekik, I. (2020). Joint functional brain network atlas estimation and feature selection for neurological disorder diagnosis with application to autism. *Med. Image Anal*. 60:101596. doi: 10.1016/j.media.2019.101596

Minzenberg, M. J., Laird, A. R., Thelen, S., Carter, C. S., and Glahn, D. C. (2009). Meta-analysis of 41 functional neuroimaging studies of executive function in schizophrenia. *Arch. Gen. Psychiatry* 66, 811–822. doi: 10.1001/archgenpsychiatry.2009.91

Mirjalili, S., Mirjalili, S. M., and Lewis, A. (2014). Grey wolf optimizer. *Adv. Eng. Softw*. 69, 46–61. doi: 10.1016/j.advengsoft.2013.12.007

Mothilal, R. K., Sharma, A., and Tan, C. (2020). "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 conference on Fairness, Accountability, and Transparency* (New York, NY: ACM), 607–617. doi: 10.1145/3351095.3372850

Naheed, N., Shaheen, M., Khan, S. A., Alawairdhi, M., and Khan, M. A. (2020). Importance of features selection, attributes selection, challenges and future directions for medical imaging data: a review. *Comput. Model. Eng. Sci*. 125, 314–344. doi: 10.32604/cmes.2020.011380

Orellana, G., and Slachevsky, A. (2013). Executive functioning in schizophrenia. *Front. Psychiatry* 4:35. doi: 10.3389/fpsyt.2013.00035

Prado-Romero, M. A., Prenkaj, B., Stilo, G., and Giannotti, F. (2023). A survey on graph counterfactual explanations: definitions, methods, evaluation, and research challenges. *ACM Comput. Surv*. 56, 1–37. doi: 10.1145/3618105

Rantala, M. J., Luoto, S., Borráz-León, J. I., and Krams, I. (2022). Schizophrenia: the new etiological synthesis. *Neurosci. Biobehav. Rev*. 142:104894. doi: 10.1016/j.neubiorev.2022.104894

Richens, J. G., Lee, C. M., and Johri, S. (2020). Improving the accuracy of medical diagnosis with causal machine learning. *Nat. Commun*. 11:3923. doi: 10.1038/s41467-020-17419-7

Roffo, G., Melzi, S., Castellani, U., Vinciarelli, A., and Cristani, M. (2020). Infinite feature selection: a graph-based feature filtering approach. *IEEE Trans. Pattern Anal. Mach. Intell*. 43, 4396–4410. doi: 10.1109/TPAMI.2020.3002843

Song, X., Wu, K., and Chai, L. (2023). Brain network analysis of schizophrenia patients based on hypergraph signal processing. *IEEE Trans. Image Process*. 32, 4964–4976. doi: 10.1109/TIP.2023.3307975

Spreitzer, N., Haned, H., and van der Linden, I. (2022). "Evaluating the practicality of counterfactual explanations," in *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS* 2022.

Sunil, G., Gowtham, S., Bose, A., Harish, S., and Srinivasa, G. (2024). Graph neural network and machine learning analysis of functional neuroimaging for understanding schizophrenia. *BMC Neurosci*. 25:2. doi: 10.1186/s12868-023-00841-0

Turner, B. O., Paul, E. J., Miller, M. B., and Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Commun. Biol*. 1:62. doi: 10.1038/s42003-018-0073-z

Verma, S., Goel, T., Tanveer, M., Ding, W., Sharma, R., Murugan, R., et al. (2023). Machine learning techniques for the schizophrenia diagnosis: a comprehensive review and future research directions. *J. Ambient Intell. Humaniz. Comput*. 14, 4795–4807. doi: 10.1007/s12652-023-04536-6

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harv. JL Tech*. 31:841. doi: 10.2139/ssrn.3063289

Wang, P., Xue, B., Liang, J., and Zhang, M. (2021). Multiobjective differential evolution for feature selection in classification. *IEEE Trans. Cybern*. 53, 4579–4593. doi: 10.1109/TCYB.2021.3128540

Wang, Y., Li, Z., Wang, Y., Wang, X., Zheng, J., Duan, X., et al. (2015). A novel approach for stable selection of informative redundant features from high dimensional fMRI data. *arXiv* [Preprint]. arXiv:1506.08301. doi: 10.48550/arXiv.1506.08301

Wei, G.-X., Ge, L., Chen, L.-Z., Cao, B., and Zhang, X. (2021). Structural abnormalities of cingulate cortex in patients with first-episode drug-naïve schizophrenia comorbid with depressive symptoms. *Hum. Brain Mapp*. 42, 1617–1625. doi: 10.1002/hbm.25315

Whitfield-Gabrieli, S., Thermenos, H. W., Milanovic, S., Tsuang, M. T., Faraone, S. V., McCarley, R. W., et al. (2009). Hyperactivity and hyperconnectivity of the default network in schizophrenia and in first-degree relatives of persons with schizophrenia. *Proc. Nat. Acad. Sci*. 106, 1279–1284. doi: 10.1073/pnas.0809141106

Xing, Y., Kochunov, P., van Erp, T. G., Ma, T., Calhoun, V. D., Du, Y., et al. (2022). A novel neighborhood rough set-based feature selection method and its application to biomarker identification of schizophrenia. *IEEE J. Biomed. Health Inform*. 27, 215–226. doi: 10.1109/JBHI.2022.3212479

Zhang, X., Braun, U., Harneit, A., Zang, Z., Geiger, L. S., Betzel, R. F., et al. (2021). Generative network models of altered structural brain connectivity in schizophrenia. *Neuroimage* 225:117510. doi: 10.1016/j.neuroimage.2020.117510

Zhu, C., Tan, Y., Yang, S., Miao, J., Zhu, J., Huang, H., et al. (2024). Temporal dynamic synchronous functional brain network for schizophrenia classification and lateralization analysis. *IEEE Trans. Med. Imaging* 43, 4307–4318. doi: 10.1109/TMI.2024.3419041

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 20 August 2025
DOI 10.3389/fmed.2025.1659653

Check for updates

OPEN ACCESS

EDITED BY
Deepti Deshwal,
Maharaja Surajmal Institute of Technology, India

REVIEWED BY
Neeraj Kumar Pandey,
Graphic Era University, India
Weiwei Jiang,
Beijing University of Posts and Telecommunications (BUPT), China
Bo Kyu Choi,
Yonsei University College of Medicine, Republic of Korea

*CORRESPONDENCE
Gökalp Tulum
gokalptulum@topkapi.edu.tr
Jawad Rasheed
jawad.rasheed@izu.edu.tr

RECEIVED 04 July 2025
ACCEPTED 08 August 2025
PUBLISHED 20 August 2025

CITATION
Cüce F, Tulum G, Isik MI, Jalili M, Girgin G, Karadaş Ö, Baş N, Özcan B, Savaşci Ü, Şakir S, Karadaş AÖ, Teomete E, Osman O and Rasheed J (2025) A novel MRI-based deep learning–radiomics framework for evaluating cerebrospinal fluid signal in central nervous system infection.
Front. Med. 12:1659653.
doi: 10.3389/fmed.2025.1659653

COPYRIGHT
© 2025 Cüce, Tulum, Isik, Jalili, Girgin, Karadaş, Baş, Özcan, Savaşci, Şakir, Karadaş, Teomete, Osman and Rasheed. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A novel MRI-based deep learning–radiomics framework for evaluating cerebrospinal fluid signal in central nervous system infection

Ferhat Cüce[1], Gökalp Tulum[2]*, Muhammed Ikbal Isik[1], Marziye Jalili[3], Güven Girgin[4], Ömer Karadaş[5], Niray Baş[5], Berza Özcan[6], Ümit Savaşci[6], Sena Şakir[6], Akçay Övünç Karadaş[7], Eda Teomete[8], Onur Osman[2] and Jawad Rasheed[9,10,11,12]*

[1]Department of Radiology, Health Science University, Gulhane Training, and Research Hospital, Ankara, Türkiye, [2]Department of Electrical and Electronics Engineering, Topkapi University, Istanbul, Türkiye, [3]Department of Artificial Intelligence, Üsküdar University, Istanbul, Türkiye, [4]Department of Neurology, Muğla Training, and Research Hospital, Muğla, Türkiye, [5]Department of Neurology, Health Science University, Gulhane Training, and Research Hospital, Ankara, Türkiye, [6]Department of Infection Disease, Health Science University, Gulhane Training, and Research Hospital, Ankara, Türkiye, [7]Private Clinic, Ankara, Türkiye, [8]Department of Classics, University of Michigan, Ann Arbor, MI, United States, [9]Department of Computer Engineering, Istanbul Sabahattin Zaim University, Istanbul, Türkiye, [10]Department of Software Engineering, Istanbul Nisantasi University, Istanbul, Türkiye, [11]Applied Science Research Center, Applied Science Private University, Amman, Jordan, [12]Research Institute, Istanbul Medipol University, Istanbul, Türkiye

Introduction: Accurate and timely diagnosis of central nervous system infections (CNSIs) is critical, yet current gold-standard techniques like lumbar puncture (LP) remain invasive and prone to delay. This study proposes a novel noninvasive framework integrating handcrafted radiomic features and deep learning (DL) to identify cerebrospinal fluid (CSF) alterations on magnetic resonance imaging (MRI) in patients with acute CNSI.

Methods: Fifty-two patients diagnosed with acute CNSI who underwent LP and brain MRI within 48 h of hospital admission were retrospectively analyzed alongside 52 control subjects with normal neurological findings. CSF-related signals were segmented from the ventricular system and sub-lentiform nucleus parenchyma, including perivascular spaces (PVSs), using semi-automated methods on axial T2-weighted images. Two hybrid models (DenseASPP-RadFusion and MobileASPP-RadFusion), fusing radiomics and DL features, were developed and benchmarked against base DL architectures (DenseNet-201 and MobileNet-V3Large) via 5-fold nested cross-validation. Radiomics features were extracted from both original and Laplacian of Gaussian–filtered MRI data.

Results: In the sub-lentiform nucleus parenchyma, the hybrid DenseASPP-RadFusion model achieved superior classification performance (accuracy: 78.57 ± 4.76%, precision: 84.09 ± 3.31%, F1-score: 76.12 ± 6.86%), outperforming its corresponding base models. Performance was notably lower in ventricular system analyses across all models. Radiomics features derived from fine-scale filtered images exhibited the highest discriminatory power. A strict, clinically motivated patient-wise classification strategy confirmed the sub-lentiform nucleus region as the most reliable anatomical target for distinguishing infected from non-infected CSF.

Discussion: This study introduces a robust and interpretable MRI-based deep learning–radiomics pipeline for CNSI classification, with promising diagnostic

potential. The proposed framework may offer a noninvasive alternative to LP in selected cases, particularly by leveraging CSF signal alterations in PVS-adjacent parenchymal regions. These findings establish a foundation for future multicenter validation and integration into clinical workflows.

KEYWORDS

central nervous system infection, cerebrospinal fluid, brain MRI, Radiomics, deep learning, lumbar puncture, perivascular spaces

1 Introduction

Central nervous system infections (CNSIs) are neurological emergencies that demand prompt and accurate diagnosis to reduce morbidity and mortality. The gold standard for confirming CNSI involves isolating the microbial agent or detecting its antigen in cerebrospinal fluid (CSF), typically via culture or polymerase chain reaction (PCR) analysis following lumbar puncture (LP) (1, 2). However, in clinical practice, the turnaround time for these methods is often inadequate for urgent decision-making. As such, CSF pleocytosis observed on microscopy is frequently used as a proxy to initiate empirical therapy with antibiotics, antivirals, or antifungals (3). Yet, reactive or false-positive pleocytosis may occur—particularly following initial LPs or in immunocompromised patients—raising concerns about overtreatment and diagnostic uncertainty (1).

Furthermore, LP is an invasive procedure with contraindications, including the presence of intracranial mass lesions, bleeding diathesis, spinal malformations, or local infections at the puncture site (2). These factors highlight the need for reliable, noninvasive, and rapid diagnostic tools to support or replace traditional CSF sampling in specific clinical contexts.

MRI plays a vital complementary role in the evaluation of CNSI. Certain imaging patterns—such as asymmetric involvement of the temporal lobe, insula, and cingulum in herpes encephalitis; leptomeningeal enhancement in meningitis; or abscess formation and tuberculous granulomas—may suggest an infectious etiology (4). Nonetheless, normal MRI findings do not exclude infection, and the sensitivity of MRI for viral and bacterial meningitis ranges between 67.4 and 83.3% (5–7). Therefore, neuroimaging alone is insufficient, and there is an urgent demand for advanced image analysis tools that can extract diagnostic information beyond the visual capabilities of radiologists.

Radiomics addresses this gap by converting conventional medical images into high-dimensional quantitative data, capturing subtle image patterns such as intensity, texture, shape, and spatial relationships (8–10). These handcrafted features have shown promise in multiple domains, but their performance can be enhanced when fused with deep learning (DL)–derived features. DL models can automatically learn abstract, hierarchical representations from imaging data, offering complementary insights into disease phenotypes.

Recent studies have demonstrated the efficacy of DL–radiomics fusion models specifically within neurology, such as multimodal neuroimaging feature learning for Alzheimer's disease diagnosis (11), deep radiomic analysis of MRI data for Alzheimer's disease classification (12), and fusion of MRI and cognitive assessments for mild cognitive impairment diagnostics (13). Similarly, these approaches have shown promise in distinguishing multiple sclerosis lesions (14) and differentiating Parkinson's disease patients from healthy individuals using radiomic features from MRI (15) and PET imaging (16). Additionally, deep learning radiomic frameworks have been effectively used for predicting hemorrhage progression in intracerebral hemorrhage (17), forecasting outcomes after acute ischemic stroke (18), and diagnosing temporal lobe epilepsy through FDG-PET imaging (19).

Despite the growing interest in end-to-end deep learning pipelines, current evidence suggests that combining DL with handcrafted radiomics yields more interpretable and robust results especially in datasets with limited sample sizes (20–22). Consequently, standardization initiatives now recommend best practices for preprocessing, feature selection, and model validation to improve reproducibility across institutions (23).

In this study, we propose a hybrid DL–radiomics framework for classifying infected versus non-infected CSF regions in patients with suspected CNSI. We focus on two anatomical targets: the ventricular system and the sub-lentiform nucleus parenchyma, including the perivascular spaces (PVSs), which are implicated in glymphatic CSF circulation. We hypothesize that the fusion of radiomic descriptors and DL-based spatial features can enable noninvasive discrimination of CSF infection patterns, thereby supporting earlier diagnosis and potentially reducing the reliance on lumbar puncture.

2 Methods and materials

2.1 Patient

The local ethics committee approved this retrospective study, and written consent was waived.

This retrospective study included patients diagnosed with CNSI who underwent brain MRI as part of their routine clinical work-up between 2017 and 2024. Fifty-two patients in the infection group were diagnosed with acute bacterial, viral and aseptic meningitis based on a combination of clinical presentation (e.g., fever, headache, neck stiffness), CSF analysis, and microbiological testing. Importantly, none of the included patients met the diagnostic criteria for encephalitis or meningoencephalitis, and there were no findings suggestive of parenchymal involvement (such as diffusion restriction, edema, or signal abnormalities and contrast enhancement in the brain parenchyma) on MRI. Mild to moderate leptomeningeal enhancement was observed in the majority of cases on post-contrast T1-weighted images, which was consistent with active meningeal inflammation. No significant ventriculitis, abscess formation, or hydrocephalus was detected. Clinically, patients presented primarily with headache and fever, and none exhibited focal neurological deficits, altered mental status, or seizures at the time of imaging. This strict inclusion criterion

ensured a clinically and radiologically homogeneous infection cohort, thereby allowing a focused evaluation of CSF-related signal features in isolated meningitis and minimizing potential confounding from parenchymal disease.

All patients diagnosed with CNSI underwent an LP on the day of admission and had brain MRIs performed within the first 48 h after being admitted to the hospital. We excluded patients who did not undergo LP, had no brain MRI, had MRIs taken more than 48 h after treatment commenced.

The control group consisted of 52 patients with chronic headaches with normal neurological examinations and normal brain MRI reports. A total of 104 patients, including both the patient and control groups, were included in the analysis.

2.2 Imaging parameters

All brain MRIs were performed on a Philips 3 T imaging system with a dedicated head coil. All studies included axial plane fat-saturated fast spin eco T2-weighted sequence with time repetition (TR): 2,600–5,600 millisecond (ms), time echo (TE): 70–90 ms, echo train length (ETL):10–12. The slice thickness was 5 millimeters (mm). To accurately evaluate subtle cerebrospinal fluid (CSF)-specific signal alterations and to minimize inadvertent segmentation errors arising from CSF flow artifacts, pre-contrast T2-weighted images were exclusively utilized in this study. T2-weighted imaging was selected for its inherent sensitivity and superior contrast resolution regarding fluid characteristics, enabling precise and artifact-aware segmentation of CSF regions. On the other hand, sequences such as T1-weighted, post-contrast T1-weighted, FLAIR, and diffusion-weighted images (DWI) were deliberately excluded. T1-weighted and post-contrast sequences primarily emphasize anatomical structures and contrast-enhanced parenchymal or meningeal lesions, providing limited utility in isolated CSF analysis without parenchymal involvement. Likewise, FLAIR imaging suppresses CSF signals, inherently limiting its applicability for dedicated CSF signal assessment. DWI is particularly sensitive to acute parenchymal lesions, but since our study specifically excluded patients with parenchymal abnormalities, its inclusion was not considered beneficial. Since no 3D modeling was employed in our study, the slice thickness of 5 mm did not constitute a significant limitation for our analysis. This selective approach ensured methodological consistency and enhanced reliability in analyzing isolated CSF-related radiomic and deep learning features.

2.3 Semi-automated segmentation procedure

Upon consensus, two independent radiologists determined the slices in the axial planes of T2-weighted images. Subsequently, MRI images were stored in the DICOM file format and imported to the ManSeg (v.2.7d) software (24). Initially, the radiologists focused on segmenting the CSF signal in both the upper and posterior sections of the lateral ventricles' lumen, avoiding areas with visible flow artifacts. Next, to reduce the risk of missing any subtle, instantaneous changes in the normal CSF flow signal, they separately segmented the parenchyma of the sub-lentiform nucleus, which includes the perivascular spaces (PVSs) supplied by the lenticulostriate arteries. Sub-lentiform nucleus parenchyma with the PVSs would effectively represent the features of the CSF, including its contents. For each patient, the lateral ventricles' lumen and the parenchyma of the sub-lentiform nucleus were segmented bilaterally. For the segmentation of suspicious regions, the radiologists roughly delineated the boundaries of the regions of interest independently, and the segmentation process was then performed automatically using the active contour algorithm (25). Final consensus segmentation masks were obtained after resolving discrepancies through joint review. Inter-observer agreement was assessed retrospectively on a randomly selected subset of 10 patients. Mean Dice similarity coefficients were 0.92 ± 0.03 for ventricular regions and 0.91 ± 0.04 for sub-lentiform parenchyma. Figure 1 depicts samples of infected CSF and normal CSF on T2-weighted images, respectively.

2.4 Feature extraction

Radiomics features were extracted from the segmented regions on both the native T2-weighted MRI images and three Laplacian-of-Gaussian (LoG)–filtered counterparts generated with kernel sizes of $3 \times 3 \times 1$ (fine), $5 \times 5 \times 2$ (medium), and $7 \times 7 \times 3$ (coarse). While 2D morphological features were derived solely from the original T2 images, both first-order and second-order statistical features, including those from gray level co-occurrence matrix (GLCM), gray level size zone matrix (GLSZM), gray level run length matrix (GLRLM), neighboring gray-tone difference matrix (NGTDM), and gray level dependence matrix (GLDM) were extracted from all image sources. A comprehensive list of the extracted features is presented in Table 1, comprising a total of 378 features.

2.5 Classification methodology

First, the region of interest (ROI) images and their corresponding radiomics features were imported. For the deep-learning analysis, each segmented region was centrally cropped into a 32×32 pixel patch, which was then resized to 224×224 pixels using bicubic interpolation. A patient-based 5-fold cross-validation (CV) approach was employed, ensuring that each patient's ROI images and associated radiomics data remained grouped during the splitting process. One-fold was allocated as the test set, while the remaining folds were used for training and validation. Feature selection was conducted solely on the radiomics features derived from the training and validation sets. From a total of 378 radiomics features, the top 50 most discriminative features were selected using a filter-based approach. Subsequently, the training and validation sets were split into an internal 3-fold cross-validation (CV) to divide them into training and validation subsets further. Data augmentation techniques, including rotation, zooming, translation, and flipping, were applied to enhance the diversity of the training data.

In our preliminary analyses, we evaluated several advanced architectures, including Swin Transformer, Vision Transformer (ViT), and attention-based networks. However, these approaches yielded poor performance and instability due to the relatively limited size of our dataset. Therefore, DenseNet-201 and MobileNet-V3Large were selected as robust baseline architectures, given their known ability to

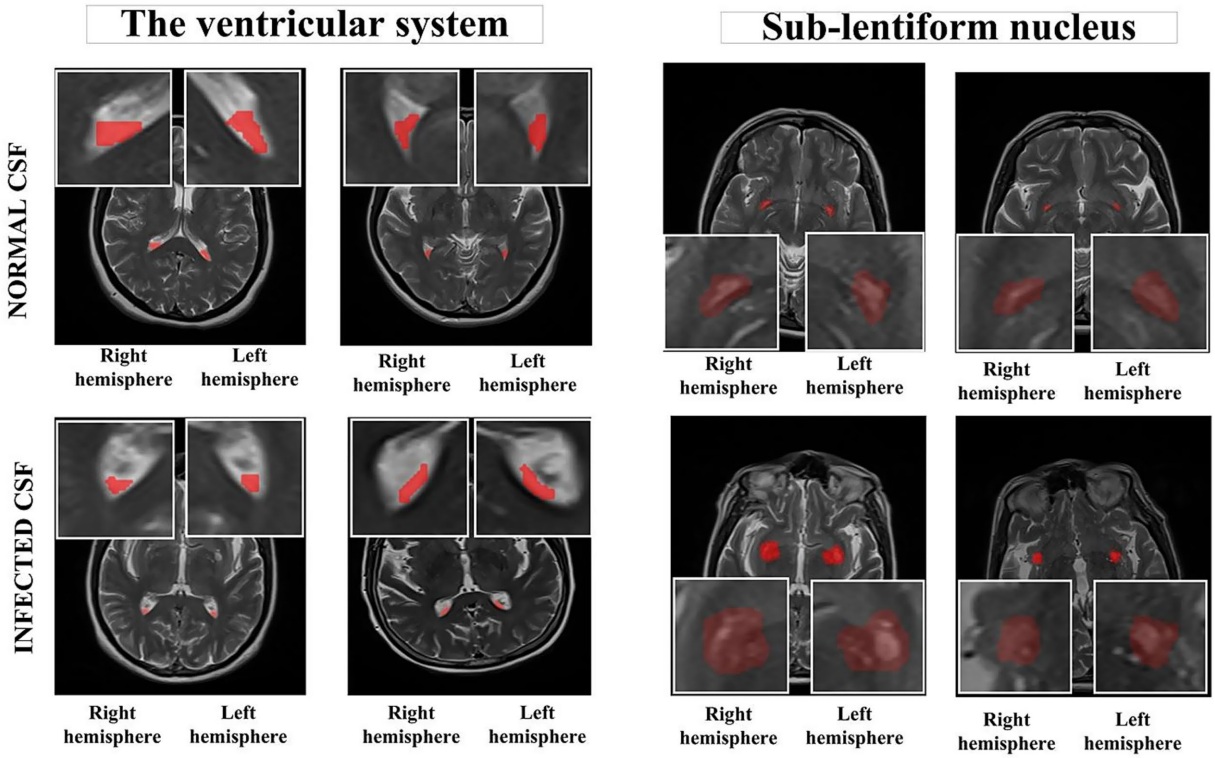


FIGURE 1
Segmented anatomical regions from the ventricular system and the sub-lentiform nucleus in both right and left hemispheres. The top row illustrates two representative cases from the control group with normal CSF, while the bottom row presents two cases from the CNSI group with infected CSF. Red-highlighted regions indicate the manually segmented areas used for radiomics feature extraction. The bounding boxes were generated as standardized input patches for deep learning models. All images are derived from T2-weighted MRI sequences. CNSI, Central Nervous System Infection; CSF, Cerebrospinal fluid.

generalize well on smaller datasets and their compatibility with our hybrid feature fusion strategy.

Model training was conducted in two phases. For the first outer fold, both the customized models DenseASPP-RadFusion and MobileASPP-RadFusion and the base models DenseNet-201 (26) and MobileNet-V3Large (27) were initialized from scratch. For the remaining folds, the weights from the previous fold were loaded to continue training. During the initial training phase, the learning rate was set to 1e-4 with a reduction factor of 0.5 and a minimum learning rate of 1e-7. Training proceeded for up to 200 epochs, with early stopping implemented after 10 epochs. During the fine-tuning phase, the learning rate was reduced to 1e-5, and the first 70% of the layers were frozen. Training was conducted for 20 epochs, with early stopping triggered after five epochs. These hyperparameters were empirically determined based on iterative experimentation within the internal training-validation splits to minimize overfitting. No hyperparameter tuning was performed on the external test sets. Throughout the process, training and validation loss, as well as accuracy metrics, were monitored. At the end of each fold, model weights and performance metrics were saved. During the testing phase, the feature selection obtained from the outer fold was applied to the test set, and model performance was evaluated using standard metrics, including accuracy, precision, recall, and F1-score. Finally, the results from all five folds were reported as mean ± standard deviation for each performance metric. Figure 2 depicts the flowchart of the classification process.

In the baseline architecture, models such as DenseNet-201 and MobileNet-V3Large were employed as feature extractors. These base models processed the input MRI images to generate feature maps, which were subsequently passed through a global average pooling layer, followed by a fully connected layer with 256 neurons and a dropout layer (rate = 0.3), leading directly to the classification output. In contrast, the proposed fusion models were designed to integrate both deep image features and handcrafted radiomics features. In the image branch of the proposed models, the backbone feature map was processed through five parallel paths. Four of these paths constituted the Atrous Spatial Pyramid Pooling (ASPP) module, employing 3×3 convolutions with dilation rates of 1, 6, 12, and 18, each followed by batch normalization and ReLU activation, producing four parallel $7 \times 7 \times 512$ feature maps. The fifth path was designed to inject global contextual information by applying global average pooling to the backbone feature map (resulting in $1 \times 1 \times 1920$), followed by a 1×1 convolution with 512 filters, and then bilinear upsampling to reach a size of $7 \times 7 \times 512$. All five outputs were concatenated to form a unified representation of size $7 \times 7 \times 2{,}560$ and then compressed via a 1×1 convolution with 512 filters.

In parallel to the image pathway, radiomics features were processed through a separate branch. A total of 378 radiomics features were extracted and reduced to 50 using filter-based feature selection.

TABLE 1 The description and the total number of radiomics features.

Image Type	Feature Class	Number of features	Total number of features
Original image	1. First order statistics	17	102
	2. 2D shape features	9	
	3. Gray level co-occurrence matrix (GLCM) features	24	
	4. Gray level size zone matrix (GLSZM) features	16	
	5. Gray level run length matrix (GLRLM) features	16	
	6. Neighboring gray tone difference matrix (NGTDM) features	5	
	7. Gray level dependence matrix (GLDM) features	14	
Log filter (FINE, MEDIUM, COARSE PATTERNS)	1. First order statistics	51	276
	2. Gray level co-occurrence matrix (GLCM) features	72	
	3. Gray level size zone matrix (GLSZM) features	48	
	4. Gray level run length matrix (GLRLM) features	48	
	5. Neighboring gray tone difference matrix (NGTDM) features	15	
	6. Gray level dependence matrix (GLDM) features	42	

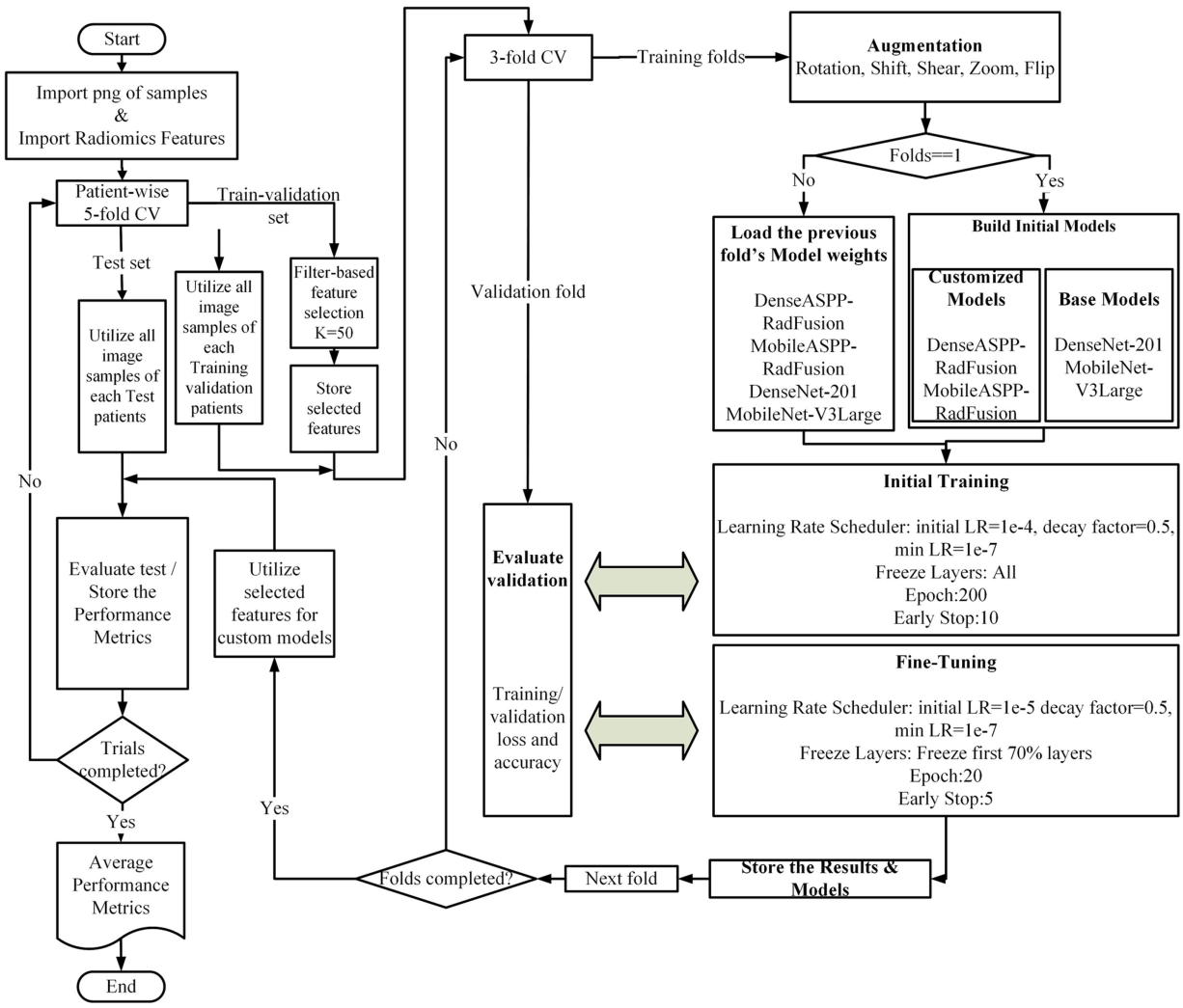


FIGURE 2
Overview of the proposed classification framework.

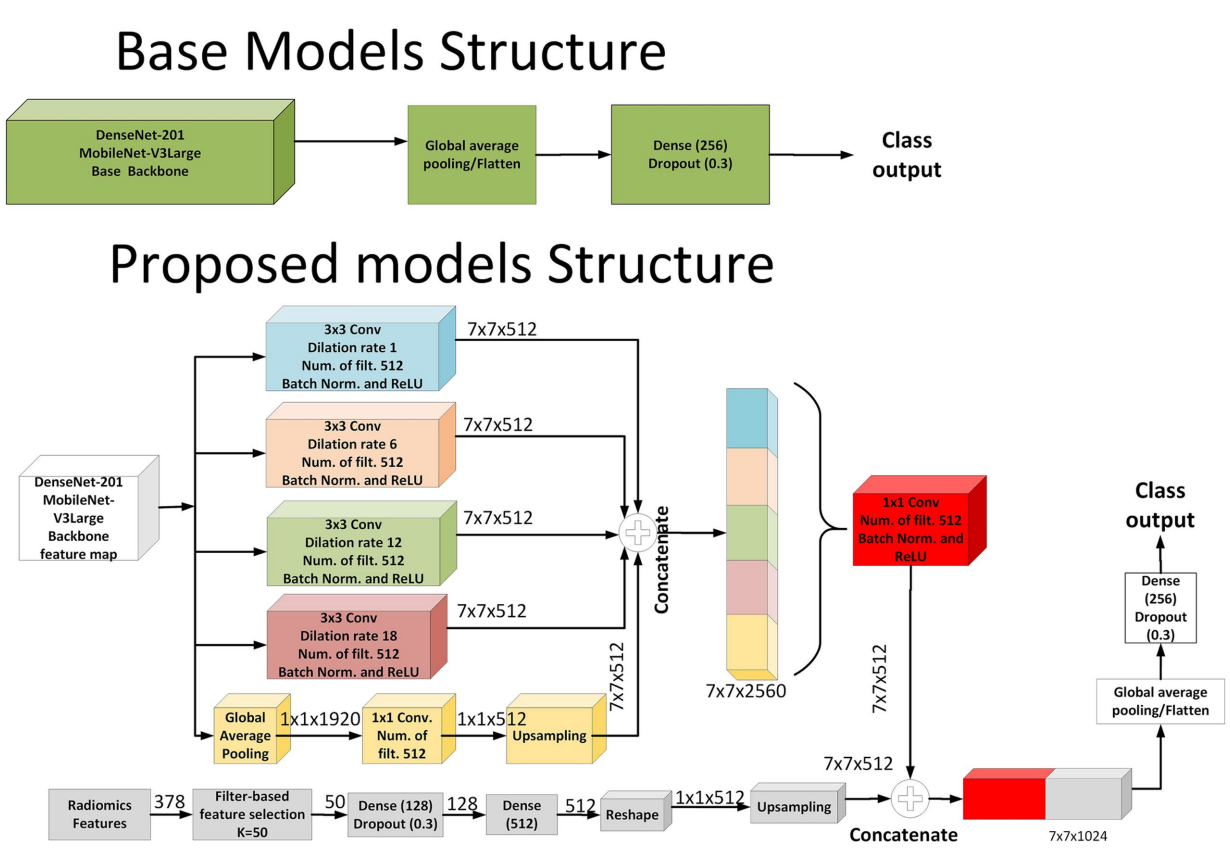


FIGURE 3
Schematic representation of the baseline and proposed model architectures.

These selected features passed through two fully connected layers [Dense (128) and Dense (512)] with dropout, reshaped into a $1 \times 1 \times 512$ tensor and then upsampled to $7 \times 7 \times 512$ to match the spatial resolution of the image features. Finally, the outputs from both the image and radiomics branches were concatenated along the channel axis, forming a $7 \times 7 \times 1{,}024$ fused representation. This combined feature map was subjected to global average pooling, followed by a Dense (256) layer with dropout, and terminated with a softmax classification layer. This architecture effectively captured both spatial and contextual information from MRI data, enriched by complementary radiomics descriptors. As illustrated in Figure 3, the proposed model architecture integrates both ASPP-enhanced image features and spatially fused radiomics features. The implementation code for the proposed MRI-based deep learning–radiomics framework is publicly available at: https://github.com/DrGokalpTulum/MRI-Based-Deep-Learning-Radiomics-Framework-for-Evaluating-Cerebrospinal-Fluid-Signal-.git.

3 Results

In the CNSI group, 55.7% ($n = 29$) of the patients were male, 44.3% ($n = 23$) were female, and the mean age was 43.5 ± 22.5 years. In the control group, 33.9% ($n = 18$) of the patients were male, 66.1% ($n = 34$) were female, and the mean age was 46.7 ± 11 years.

The CSF analysis was performed on the patient's admission to the health institution. The macroscopic appearance of the CSF, the amount of CSF glucose and protein, pleocytosis in microscopy, and the presence of microorganisms in the Gram stain were evaluated. High CSF protein, low glucose, leukocyte count of 100 or more cells/mm3, and neutrophil predominance are evaluated as bacterial meningitis; normal CSF glucose, borderline high protein levels, and lymphocytes being the predominant cell in the cell count were evaluated as viral meningitis; normal CSF findings were accepted as aseptic meningitis.

According to early biochemical and microscopy results, bacterial meningitis was observed in 37 patients, viral meningitis in 14 patients, and CSF findings of 1 patient were evaluated as aseptic meningitis. While no culture medium growth was detected in the CSF of 24 patients, Streptococcus was detected in 5 patients, *E. coli* in 3 patients, Brucella in 2 patients, Acinetobacter in 1 patient, Neisseria in 1 patient, and Proteus in 1 patient, according to CSF culture results. Varicella Zoster Virus PCR positivity was detected in the CSF of two patients. Based on clinical and laboratory results in the patient group, antimicrobial treatment for CNSI was empirically started. After the diagnosis of the agent was confirmed by culture, PCR, and serology, treatment revision was performed with de-escalation in three patients.

During the 5-fold outer cross-validation, a total of 378 radiomics features were subjected to feature selection, and the top 50 features were retained in each fold. Across all folds, a total of 92 unique features were selected. Among these, 20 features were consistently selected in all five folds, indicating strong discriminative capacity. These high-frequency features primarily originated from the Laplacian of Gaussian (LoG) filtered MRI with fine kernels (2 mm). In particular,

TABLE 2 Performance metrics (mean ± std) of all models.

Evaluation Area	Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Performance metrics for the sub-lentiform nucleus parenchyma	DenseASPP-RadFusion	78.57 ± 4.76	84.09 ± 3.31	70.00 ± 10.93	76.12 ± 6.86
	MobileASPP-RadFusion	74.40 ± 2.28	73.66 ± 4.41	77.05 ± 14.82	74.42 ± 6.66
	DenseNet-201	73.81 ± 8.01	79.72 ± 5.18	62.96 ± 14.04	70.03 ± 10.52
	MobileNet-V3Large	52.98 ± 9.20	55.66 ± 13.09	76.14 ± 27.62	60.84 ± 6.49
Performance metrics for the ventricular system	DenseASPP-RadFusion	60.52 ± 4.87	64.65 ± 8.03	47.64 ± 17.02	53.46 ± 11.04
	MobileASPP-RadFusion	59.07 ± 8.63	58.14 ± 8.86	70.18 ± 26.40	61.58 ± 13.00
	DenseNet-201	57.26 ± 7.05	58.54 ± 13.48	44.64 ± 20.66	49.37 ± 15.32
	MobileNet-V3Large	59.62 ± 5.06	56.69 ± 4.68	77.36 ± 12.94	65.28 ± 7.26

The upper section presents results for the sub-lentiform nucleus parenchyma with PVSs, while the lower section shows results for the ventricular system. Metrics evaluated across 5-fold cross-validation for each model.

features such as Energy, Maximum, Range, Long Run Emphasis, and High Gray Level Zone Emphasis repeatedly appeared across all folds.

Additionally, 16 features appeared in four folds and five features in three folds, most of which stemmed from LoG-filtered MRI with medium kernels (4 mm) or original T2-weighted images. These consistently selected features highlight the critical role of multiscale texture descriptors in capturing the heterogeneity of cerebrospinal fluid regions. Detailed feature selection results, including Feature Name, Image Source, Feature Class, and Frequency, are provided in the Supplementary file.

Upon investigating the classification results, the proposed fusion models (DenseASPP-RadFusion and MobileASPP-RadFusion) demonstrate improvements over their corresponding base architectures (DenseNet-201 and MobileNet-V3Large) in the sub-lentiform nucleus parenchyma region. DenseASPP-RadFusion achieved the highest mean accuracy (78.57 ± 4.76%) and precision (84.09 ± 3.31%), with relatively low standard deviations, indicating both high performance and consistency across folds. Although MobileASPP-RadFusion yielded the highest mean recall (77.05 ± 14.82%), the associated standard deviation was relatively large, suggesting instability in sensitivity across different validation folds.

In contrast, none of the models showed strong classification capability in the ventricular system. Accuracy values remained between 57.26 and 60.52%, while F1-scores were notably lower, particularly for DenseASPP-RadFusion (53.46 ± 11.04%) and DenseNet-201 (49.37 ± 15.32%). Moreover, the standard deviations in recall for all models were high (ranging from 12.94 to 20.66%), indicating a lack of reliability in detecting true positives in ventricular-level CSF signals.

The results show that the sub-lentiform nucleus parenchyma with PVSs provides more stable and discriminative information for classification tasks compared to the ventricular system. The performance of the models was statistically significantly different ($p < 0.05$). Detailed performance metrics for all models and anatomical regions are presented in Table 2, while the corresponding ROC curves are illustrated in Figure 4.

To complement the fold-level evaluation, patient-wise classification performance was also assessed under clinically motivated assumptions. To assess patient-level diagnostic performance under clinical assumptions, strict patient-wise accuracy was calculated separately for each class (infection and control) across all outer folds. Since each patient had two separate ROIs from the right and left sub-lentiform nucleus levels, the following decision rules were applied: for infection cases (class 1), a prediction was considered correct if at least one of the two ROIs was classified as infected, reflecting a clinically cautious approach to minimize false negatives. Conversely, for control cases (class 0), a prediction was deemed correct only if both ROIs were classified as non-infected, ensuring stricter criteria for healthy labeling. This binary patient-wise accuracy was computed per case and averaged within each fold for all models.

Figure 5 presents box plots illustrating the distribution of strict patient-wise accuracy values for each model, separately for the sub-lentiform nucleus parenchyma and ventricular system. In the sub-lentiform nucleus parenchyma with PVSs, the proposed model DenseASPP-RadFusion yielded the most stable and accurate performance, with infection class accuracies tightly clustered within the 80 to 90% interquartile range and control accuracies between 70 and 80%, both showing low interfold variability. Similarly, DenseNet-201 achieved high median values, though with a slightly wider spread in the control group. Notably, both models exhibited limited presence of outliers, suggesting consistency in predictions across patient subsets.

On the other hand, MobileNet-V3Large exhibited high variability and lower median accuracy, particularly for control patients. Its control group performance distribution dropped to a lower interquartile range (below 60%) and revealed several outliers, reflecting instability across folds. MobileASPP-RadFusion demonstrated acceptable median accuracy but higher dispersion, particularly in control cases, indicating less consistent generalization across folds.

In the ventricular system, all models demonstrated lower and more dispersed accuracy distributions, indicating reduced reliability in this anatomical region. For instance, although MobileNet-V3Large achieved reasonable infection accuracy, its control classification remained weak and inconsistent. MobileASPP-RadFusion and DenseNet-201 exhibited moderate accuracy with noticeably higher standard deviations, particularly in control predictions, highlighting the challenge of robust CSF signal interpretation in ventricular regions. The broader interquartile ranges and frequent outliers in the ventricular plots underscore the inconsistency of model behavior in this region. These findings further reinforce that the sub-lentiform nucleus parenchyma with PVSs provides a more clinically reliable classification, both at the level of the fold and the patient.

Under this realistic criterion, the proposed fusion models demonstrated high stability and accuracy, particularly in the

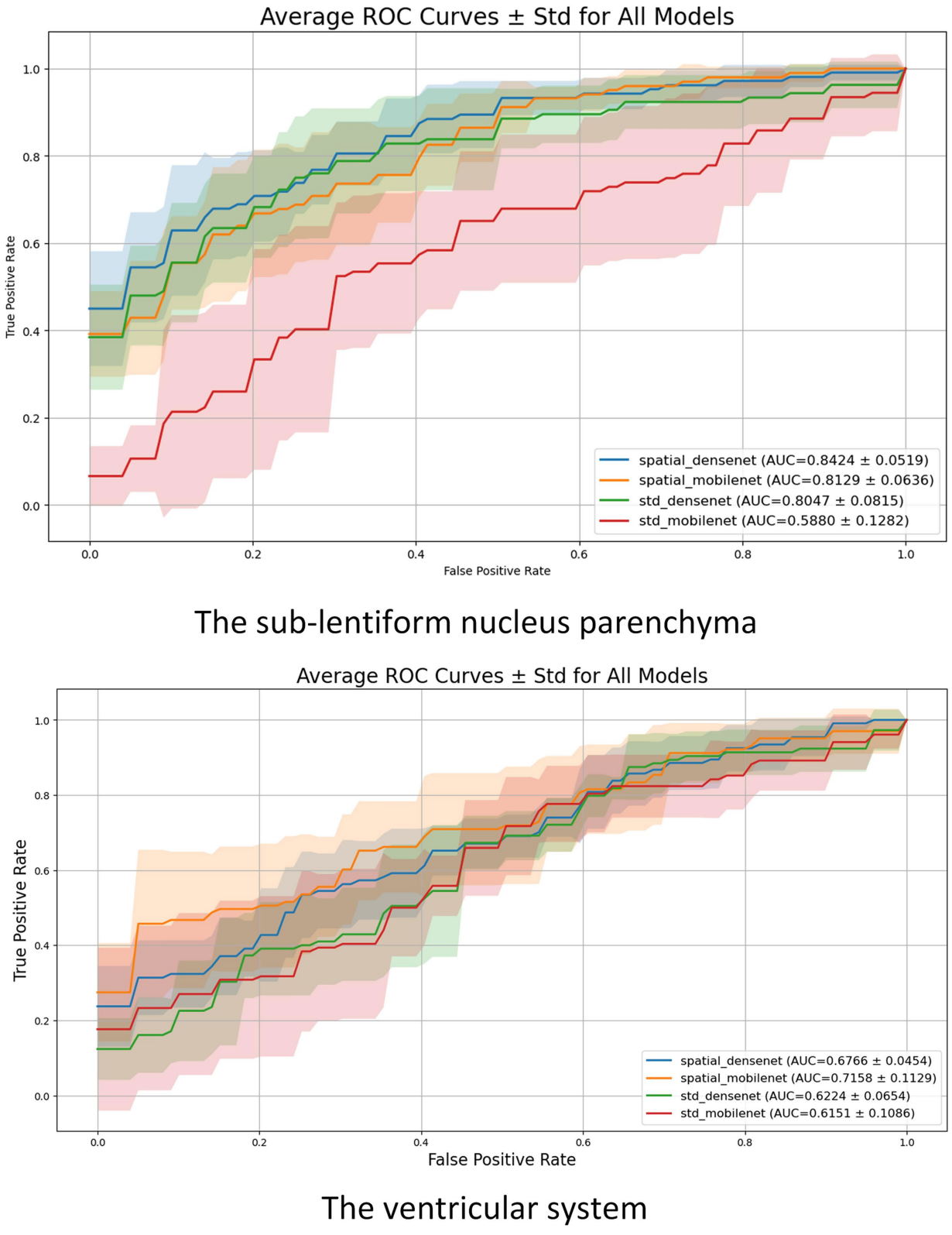


FIGURE 4
Mean ROC curves with standard deviation (shaded regions) for each model across the 5-fold cross-validation. The upper plot illustrates results for the sub-lentiform nucleus parenchyma with PVSs. The lower plot presents results for the ventricular system. Legend entries include average AUC ± standard deviation for each model.

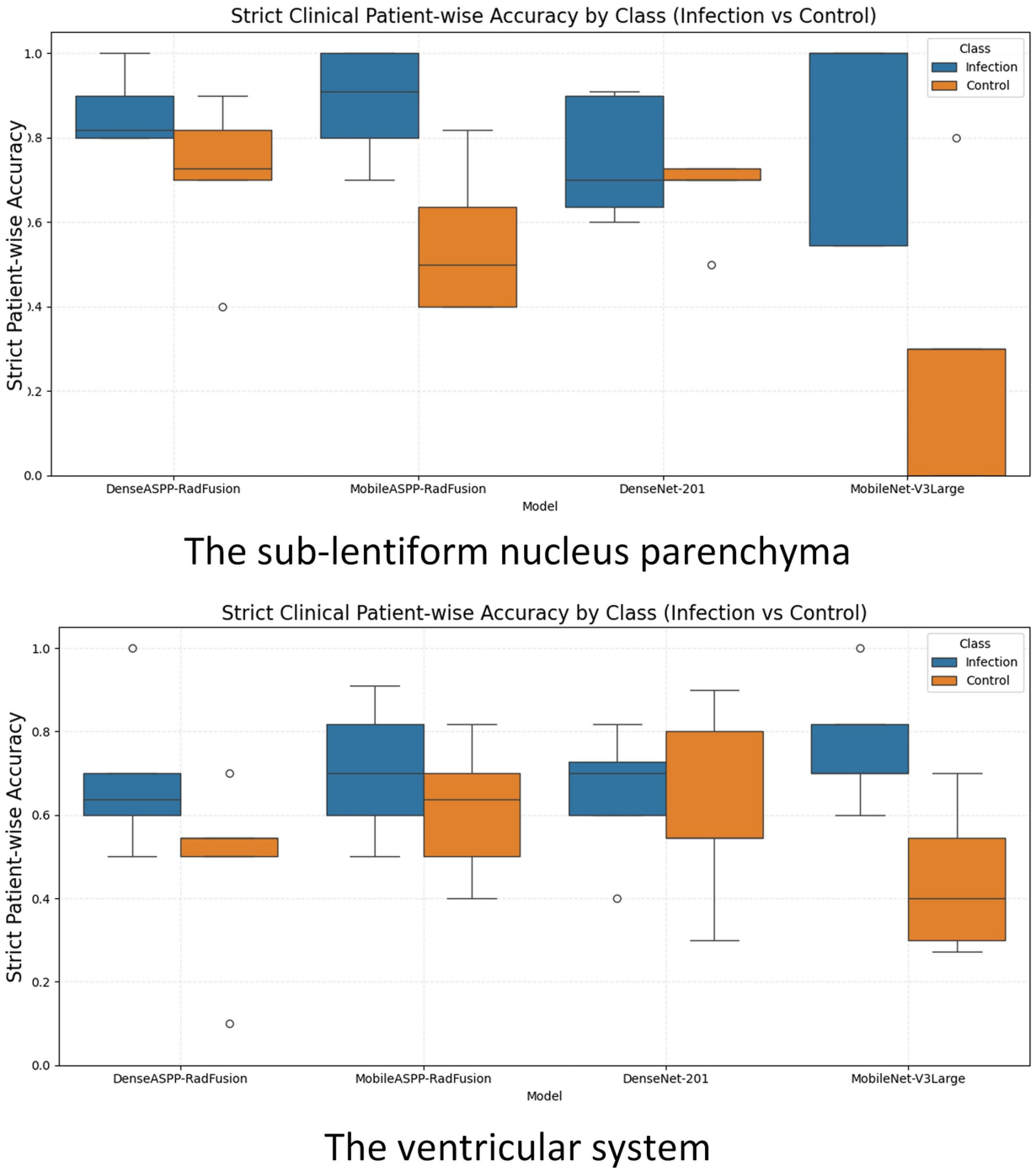


FIGURE 5
Box plots illustrating strict clinical patient-wise accuracy for CNSI and control classes across all models, evaluated separately for the sub-lentiform nucleus parenchyma (top) and the ventricular system (bottom). CNS, Central nervous system infection.

sub-lentiform nucleus. By contrast, all models showed reduced and inconsistent performance in ventricular CSF classification, further underscoring the diagnostic limitations of relying solely on ventricular analysis. Discordant predictions between left and right sub-lentiform nucleus evaluations occurred in 22.6 ± 3.1% for DenseASPP-RadFusion, 30.4 ± 6.0% for MobileASPP-RadFusion, 29.9 ± 4.7% for DenseNet-201, and 39.5 ± 5.9% for MobileNet-V3Large, indicating varying levels of stability in bilateral predictions.

4 Discussion

In this study, we developed and evaluated a novel MRI-based deep learning–radiomics framework to classify CSF signals in patients with acute CNSIs. Our findings demonstrate that the fusion of handcrafted radiomic descriptors with DL features enables more accurate and reliable classification of infected versus non-infected CSF, particularly when analyzing the sub-lentiform nucleus parenchyma region. These results offer promising evidence for the utility of noninvasive

imaging-based diagnostics as a potential complement or alternative to LP in selected clinical contexts.

Despite their central role in CNSI diagnosis, CSF analyses via LP remain invasive and carry procedural risks, including herniation, hemorrhage, or infection—especially in patients with intracranial mass lesions or bleeding disorders (1–3). Moreover, pleocytosis, often used as a surrogate marker of infection, may occasionally yield false-positive results, especially after repeated LPs or in immunocompromised individuals (1). These limitations necessitate the development of alternative diagnostic strategies that are rapid, noninvasive, and reproducible.

While MRI has proven valuable in detecting certain CNSI patterns—such as temporal lobe involvement in herpes encephalitis or leptomeningeal enhancement in meningitis—it lacks sufficient sensitivity to reliably detect all cases, particularly in early or ambiguous presentations (4–7). In our study, conventional visual inspection of ventricular CSF signals on MRI did not provide sufficient discriminatory power to distinguish infected from non-infected fluid. This is likely due to the inherent signal homogeneity and dynamic flow of CSF in the ventricles, which limits the effectiveness of static image-based analysis.

Indeed, previous AI-based studies evaluating body fluid segmentation—such as pleural or synovial effusions—have reported promising results (28, 29). However, these studies primarily focused on relatively static fluids that exhibit well-defined boundaries and textural consistency. CSF, on the other hand, is in constant motion, and its flow-dependent signal properties pose substantial challenges for conventional image segmentation and classification.

To address these limitations, our study focused on the sub-lentiform nucleus parenchyma, specifically targeting regions that include perivascular spaces (PVSs)—components of the glymphatic system that mediate convective CSF flow from penetrating arteries into the interstitial space. Unlike the ventricular system, these parenchymal regions are less affected by flow artifacts and may reflect more stable and informative imaging features. Additionally, inflammation in adjacent brain parenchyma during CNSI—though often invisible on routine MRI—may alter tissue texture and contribute to detectable radiomic changes.

Our results strongly support this hypothesis. The hybrid DenseASPP-RadFusion model, which integrates multiscale radiomics with spatially resolved DL features, achieved a mean classification accuracy of 78.6% in the sub-lentiform nucleus region—substantially outperforming both its base architecture (DenseNet-201) and all models applied to the ventricular system. Features derived from Laplacian of Gaussian (LoG)–filtered images, particularly with fine kernels (2 mm), contributed most significantly to model performance, suggesting that subtle intensity variations in the CSF-parenchyma interface are key discriminative elements.

Furthermore, we applied a clinically grounded, strict patient-wise classification strategy, wherein a diagnosis of infection was accepted if either hemisphere exhibited an infected CSF pattern, while a control classification required bilateral confirmation of non-infection. Under this realistic criterion, the proposed fusion models demonstrated high stability and accuracy, particularly in the sub-lentiform nucleus. By contrast, all models showed reduced and inconsistent performance in ventricular CSF classification, further underscoring the diagnostic limitations of relying solely on ventricular analysis.

The broader implication of our findings lies in the potential of hybrid DL–radiomics frameworks to improve CNSI diagnosis in settings where LP is delayed, contraindicated, or inconclusive. To our knowledge, this is the first study to apply a deep learning–radiomics fusion approach to analyze CSF signal patterns in brain MRI for the classification of CNSIs. Prior applications of AI to fluid-based diagnostics have largely centered around cancer-related effusions or synovial fluid segmentation in rheumatology (28, 29), whereas our study opens new directions for infectious disease imaging.

DenseNet-based models consistently outperformed MobileNet-based models across most performance metrics, likely due to their deeper and densely connected architectures enabling effective feature reuse and robust representation learning. Conversely, MobileNet's design prioritizes computational efficiency and fewer parameters, potentially limiting its capability to capture subtle radiomic patterns. Thus, DenseNet architectures may be preferable for tasks demanding detailed representation of subtle imaging features, whereas MobileNet remains beneficial under computational constraints.

Nevertheless, our study has limitations. The relatively modest sample size (n = 104) and single-center design may limit generalizability. However, all MRIs were acquired using a uniform 3 T scanner and standardized imaging protocol, enhancing internal consistency. Future research should validate these findings using multicenter datasets with larger, more diverse populations and include longitudinal evaluation across various CNSI subtypes (e.g., bacterial, viral, fungal). Additionally, the integration of clinical metadata (e.g., laboratory markers, symptoms) with imaging features may further improve classification performance. Moreover, we used a slice thickness of 5 mm, which is relatively thicker than the thin-cut images ($\leq$3 mm) typically preferred in current brain MRI research. Although this could potentially limit the segmentation accuracy and reliability in studies utilizing 3D modeling approaches, our analyses and segmentations were strictly performed on 2D images, reducing its impact within our study context. Future studies using thinner slice imaging might offer further improvements in segmentation detail and predictive performance. Future studies could further enhance the clinical impact and interpretability of the proposed fusion models by incorporating explainable AI (XAI) methodologies to identify and visualize the most influential radiomic and deep-learning-derived features. Integrating these techniques would significantly strengthen model transparency, improve clinical confidence, and facilitate a smoother translation into clinical practice.

In conclusion, our study introduces a novel, interpretable, and clinically relevant framework for noninvasive CNSI assessment using advanced radiomics and deep learning methods. The sub-lentiform nucleus parenchyma, inclusive of PVSs, emerges as a promising anatomical region for CSF evaluation. This approach has the potential to complement traditional LP-based diagnostics and support faster, safer, and more accurate CNSI management in clinical practice.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Clinical Research Ethics Committee of the University of Health Sciences, Gülhane Training and Research Hospital (Decision No.: 2023/18–amendment; Approval Date: 15 March 2023). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

FC: Conceptualization, Data curation, Resources, Writing – original draft. GT: Conceptualization, Methodology, Software, Visualization, Writing – original draft. MI: Data curation, Resources, Writing – original draft. MJ: Software, Writing – review & editing. GG: Conceptualization, Writing – review & editing. ÖK: Conceptualization, Project administration, Validation, Writing – review & editing. NB: Data curation, Resources, Validation, Writing – review & editing. BÖ: Data curation, Resources, Writing – review & editing. ÜS: Data curation, Validation, Writing – original draft. SŞ: Data curation, Resources, Validation, Writing – review & editing. AK: Conceptualization, Writing – review & editing. ET: Writing – review & editing. OO: Methodology, Software, Writing – review & editing. JR: Methodology, Software, 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 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

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.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2025.1659653/full#supplementary-material

References

1. Troendle M, Pettigrew A. A systematic review of cases of meningitis in the absence of cerebrospinal fluid pleocytosis on lumbar puncture. *BMC Infect Dis*. (2019) 19:692. doi: 10.1186/s12879-019-4204-z

2. Wright BLC, Lai JTF, Sinclair AJ. Cerebrospinal fluid and lumbar puncture: a practical review. *J Neurol*. (2012) 259:1530–45. doi: 10.1007/s00415-012-6413-x

3. Bedetti L, Marrozzini L, Baraldi A, Spezia E, Iughetti L, Lucaccioni L, et al. Pitfalls in the diagnosis of meningitis in neonates and young infants: the role of lumbar puncture. *J Matern Fetal Neonatal Med*. (2019) 32:4029–35. doi: 10.1080/14767058.2018.1481031

4. Nguyen I, Urbanczyk K, Mtui E, Li S. Intracranial CNS infections: a literature review and radiology case studies. *Semin Ultrasound CT MR*. (2020) 41:106–20. doi: 10.1053/j.sult.2019.09.003

5. Raza MA, Tufail M, Altuf L, Chaudhary K, Ghazanfar S, Bukhari SKA, et al. Diagnostic accuracy of magnetic resonance imaging for central nervous system associated infectious diseases. *J Health Rehabil Res*. (2024) 4:1–4. doi: 10.61919/jhrr.v4i3.1096

6. Kralik SF, Vallejo JG, Kukreja MK, Salman R, Orman G, Huisman TAGM, et al. Diagnostic accuracy of MRI for detection of meningitis in infants. *AJNR Am J Neuroradiol*. (2022) 43:1350–5. doi: 10.3174/ajnr.A7610

7. Vaswani AK, Nizamani WM, Ali M, Aneel G, Shahani BK, Hussain S. Diagnostic accuracy of contrast-enhanced FLAIR magnetic resonance imaging in diagnosis of meningitis correlated with CSF analysis. *Indus J Biosci Res*. (2025) 3. doi: 10.70749/ijbr.v3i5.1344

8. Van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights Imag*. (2020) 11:91. doi: 10.1186/s13244-020-00887-2

9. Jing R, Wang J, Li J, Wang X, Li B, Xue F, et al. A wavelet features derived radiomics nomogram for prediction of malignant and benign early-stage lung nodules. *Sci Rep*. (2021) 11:22330. doi: 10.1038/s41598-021-01470-5

10. Xia C, Zuo M, Lin Z, Deng L, Rao Y, Chen W, et al. Multimodal deep learning fusing clinical and Radiomics scores for prediction of early-stage lung adenocarcinoma lymph node metastasis. *Acad Radiol*. (2025) 32:2977–89. doi: 10.1016/j.acra.2024.12.018

11. Liu S, Liu S, Cai W, Che H, Pujol S, Kikinis R, et al. Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. *IEEE Trans Biomed Eng*. (2015) 62:1132–40. doi: 10.1109/TBME.2014.2372011

12. Chaddad A, Desrosiers C, Niazi T. Deep radiomic analysis of MRI related to Alzheimer's disease. *IEEE Access*. (2018) 6:58213–21. doi: 10.1109/ACCESS.2018.2871977

13. Qiu S, Chang GH, Panagia M, Gopal DM, Au R, Kolachalama VB. Fusion of deep learning models of MRI scans, Mini-mental state examination, and logical memory test enhances diagnosis of mild cognitive impairment. *Alzheimer's Dementia: Diagnosis, Assess Dis Monitor*. (2018) 10:737–49. doi: 10.1016/j.dadm.2018.08.013

14. Ye Z, George A, Wu AT, Niu X, Lin J, Adusumilli G, et al. Deep learning with diffusion basis spectrum imaging for classification of multiple sclerosis lesions. *Annals Clin Translational Neurol*. (2020) 7:695–706. doi: 10.1002/acn3.51037

15. Wu Y, Jiang JH, Chen L, Lu JY, Ge JJ, Liu FT, et al. Use of radiomic features and support vector machine to distinguish Parkinson's disease cases from normal controls. *Annals Translational Med*. (2019) 7:773. doi: 10.21037/atm.2019.11.26

16. Sun X, Ge J, Li L, Zhang Q, Lin W, Chen Y, et al. Use of deep learning-based radiomics to differentiate Parkinson's disease patients from normal controls: a study based on [18F]FDG PET imaging. *Eur Radiol*. (2022) 32:8008–18. doi: 10.1007/s00330-022-08799-z

17. Song L, Zhou H, Guo T, Qiu X, Tang D, Zou L, et al. Predicting hemorrhage progression in deep intracerebral hemorrhage: a multicenter retrospective cohort study. *World Neurosurg*. (2023) 170:e387–401. doi: 10.1016/j.wneu.2022.11.022

18. Hilbert A, Ramos LA, van Os HJA, Olabarriaga SD, Tolhuisen ML, Wermer MJH, et al. Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke. *Comput Biol Med*. (2019) 115:103516. doi: 10.1016/j.compbiomed.2019.103516

19. Zhang Q, Liao Y, Wang X, Zhang T, Feng J, Deng J, et al. A deep learning framework for 18F-FDG PET imaging diagnosis in pediatric patients with temporal lobe

epilepsy. *Eur J Nucl Med Mol Imaging*. (2021) 48:2476–85. doi: 10.1007/s00259-020-05108-y

20. Xie C, Yu X, Tan N, Zhang J, Su W, Ni W, et al. Combined deep learning and radiomics in pretreatment radiation esophagitis prediction for patients with esophageal cancer underwent volumetric modulated arc therapy. *Radiother Oncol*. (2024) 199:110438. doi: 10.1016/j.radonc.2024.110438

21. Cheng C, Wang Y, Zhao J, Wu D, Li H, Zhao H. Deep learning and Radiomics in triple-negative breast Cancer: predicting long-term prognosis and clinical outcomes. *J Multidiscip Healthc*. (2025) 18:319–27. doi: 10.2147/JMDH.S509004

22. Cè M, Chiriac MD, Cozzi A, Macrì L, Rabaiotti FL, Irmici G, et al. Decoding Radiomics: a step-by-step guide to machine learning workflow in handcrafted and deep learning Radiomics studies. *Diagnostics (Basel)*. (2024) 14:2473. doi: 10.3390/diagnostics14222473

23. Cuce F, Tulum G, Yılmaz KB, Osman O, Aralasmak A. Radiomics method in the differential diagnosis of diabetic foot osteomyelitis and charcot neuroarthropathy. *BJR*. (2023) 96:20220758. doi: 10.1259/bjr.20220758

24. Chan TF, Vese LA. Active contours without edges. *IEEE Trans on Image Process*. (2001) 10:266–77. doi: 10.1109/83.902291

25. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. *Proceed IEEE Conference Computer Vision Pattern Recog (CVPR)*. (2017):4700–8. doi: 10.1109/CVPR.2017.243

26. Howard A, Sandler M, Chu G, Chen L-C, Chen B, Tan M, et al. Searching for mobile NetV3. *ar Xiv*. (2019), 1314–1324. doi: 10.1109/ICCV.2019.00140

27. Ozcelik N, Ozcelik AE, Guner Zirih NM, Selimoglu I, Gumus A. Deep learning for diagnosis of malign pleural effusion on computed tomography images. *Clinics (Sao Paulo)*. (2023) 78:100210. doi: 10.1016/j.clinsp.2023.100210

28. Iqbal I, Shahzad G, Rafiq N, Mustafa G, Ma J. Deep learning-based automated detection of human knee joint's synovial fluid from magnetic resonance images with transfer learning. *IET Image Process*. (2020) 14:1990–8. doi: 10.1049/iet-ipr.2019.1646

29. Das P, Roy SD, Sangma K. SFRSeg-net: synovial fluid region segmentation from rheumatoid arthritis affected small joints using USG for early detection In: A Antonacopoulos, S Chaudhuri and R Chellappa, editors. Pattern recognition. Cham: Springer Nature Switzerland (2025). 127–46.

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 20 August 2025
DOI 10.3389/fmed.2025.1646249

Check for updates

OPEN ACCESS

EDITED BY
Pardeep Sangwan,
Maharaja Surajmal Institute of Technology, India

REVIEWED BY
Jia-Bao Liu,
Anhui Jianzhu University, China
Muhammad Adnan,
Kohat University of Science and Technology, Pakistan

*CORRESPONDENCE
Nesren S. Farhah
✉ n.farhah@seu.edu.sa
Nadhem Ebrahim
✉ nebrahim@uakron.edu
Sultan Ahmad
✉ s.alisher@psau.edu.sa

RECEIVED 16 June 2025
ACCEPTED 28 July 2025
PUBLISHED 20 August 2025

CITATION
Farhah NS, Alqarni AA, Ebrahim N and Ahmad S (2025) Diagnosing autism spectrum disorders using a double deep Q-Network framework based on social media footprints. *Front. Med.* 12:1646249. doi: 10.3389/fmed.2025.1646249

COPYRIGHT
© 2025 Farhah, Alqarni, Ebrahim and Ahmad. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Diagnosing autism spectrum disorders using a double deep Q-Network framework based on social media footprints

Nesren S. Farhah[1,2]*, Ahmed Abdullah Alqarni[2,3], Nadhem Ebrahim[4]* and Sultan Ahmad[5,6]*

[1]Department of Health Informatics, College of Health Science, Saudi Electronic University, Riyadh, Saudi Arabia, [2]King Salman Center for Disability Research, Riyadh, Saudi Arabia, [3]Department of Computer Sciences and Information Technology, Al-baha University, Al-baha, Saudi Arabia, [4]Department of Computer Science, College of Engineering and Polymer Science, University of Akron, OH, United States, [5]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia, [6]School of Computer Science and Engineering, Lovely Professional University, Phagwara, India

Introduction: Social media is increasingly used in many contexts within the healthcare sector. The improved prevalence of Internet use via computers or mobile devices presents an opportunity for social media to serve as a tool for the rapid and direct distribution of essential health information. Autism spectrum disorders (ASD) are a comprehensive neurodevelopmental syndrome with enduring effects. Twitter has become a platform for the ASD community, offering substantial assistance to its members by disseminating information on their beliefs and perspectives via language and emotional expression. Adults with ASD have considerable social and emotional challenges, while also demonstrating abilities and interests in screen-based technologies.

Methods: The novelty of this research lies in its use in the context of Twitter to analyze and identify ASD. This research used Twitter as the primary data source to examine the behavioral traits and immediate emotional expressions of persons with ASD. We applied Convolutional Neural Networks with Long Short-Term Memory (CNN-LSTM), LSTM, and Double Deep Q-network (DDQN-Inspired) using a standardized dataset including 172 tweets from the ASD class and 158 tweets from the non-ASD class. The dataset was processed to exclude lowercase text and special characters, followed by a tokenization approach to convert the text into integer word sequences. The encoding was used to transform the classes into binary labels. Following preprocessing, the proposed framework was implemented to identify ASD.

Results: The findings of the DDQN-inspired model demonstrate a high precision of 87% compared to the proposed model. This finding demonstrates the potential of the proposed approach for identifying ASD based on social media content.

Discussion: Ultimately, the proposed system was compared against the existing system that used the same dataset. The proposed approach is based on variations in the text of social media interactions, which can assist physicians and clinicians in performing symptom studies within digital footprint environments.

KEYWORDS

autism spectrum disorders, diagnosing, social media, deep learning, disabilities, artificial intelligence

1 Introduction

ASD is among the most prevalent neurodevelopmental disorders. ASD is often demonstrated in children by age three and is defined by impairments in social interactions and communication, repetitive sensory-motor activities, and stereotypical behavioral patterns (1). ASD is a congenital neurodevelopmental condition characterized by symptoms that are evident in early infancy. Autism, characterized by restricted interests, repetitive behaviors, and significant disparities in social communication and interaction, typically emerges during early developmental stages and presents challenges in various social functioning domains. A child with autism induces significant anxiety within the family due to several factors, including the ambiguity of the diagnosis, the intensity and persistence of the disease, and the child's nonconformity to social norms. In opposition, social awareness of autism is markedly inadequate, often conflated with intellectual disability and seen as an incurable ailment (2, 3). The ASD concept is displayed in Figure 1.

Content on social media, particularly videos and text disseminated by parents and caregivers, has emerged as a significant resource for facilitating the early identification of ASD (4, 5). Social media are technological tools designed for sharing, enabling users to create networks or engage in existing ones. In that order, the Pew Research Center identified the most popular social media sites as YouTube, Facebook, Instagram, Pinterest, LinkedIn, Snapchat, Twitter, and WhatsApp (6). Most consumers use these networks daily. This research utilizes Twitter data to assess the stigmatization of autism and associated terminology, picked based on accessibility and popularity, with analysis conducted using artificial intelligence technologies (7).

Conventional diagnostic methods, which primarily rely on observational and behavioral evaluations, often encounter issues with accessibility, consistency, and timeliness. Recent technology breakthroughs, especially in artificial intelligence (AI), and sensor-based techniques, provide novel opportunities for improving ASD identification. By developing more objective, accurate, and scalable approaches, these technologies transform diagnostic methodologies for autism spectrum disorder (ASD) (8–10). One new way to study the motor patterns, attentional processes, and physiological responses linked to ASD in real-time is wearable sensors, eye-tracking devices, and multimodal virtual reality settings. These technologies have the

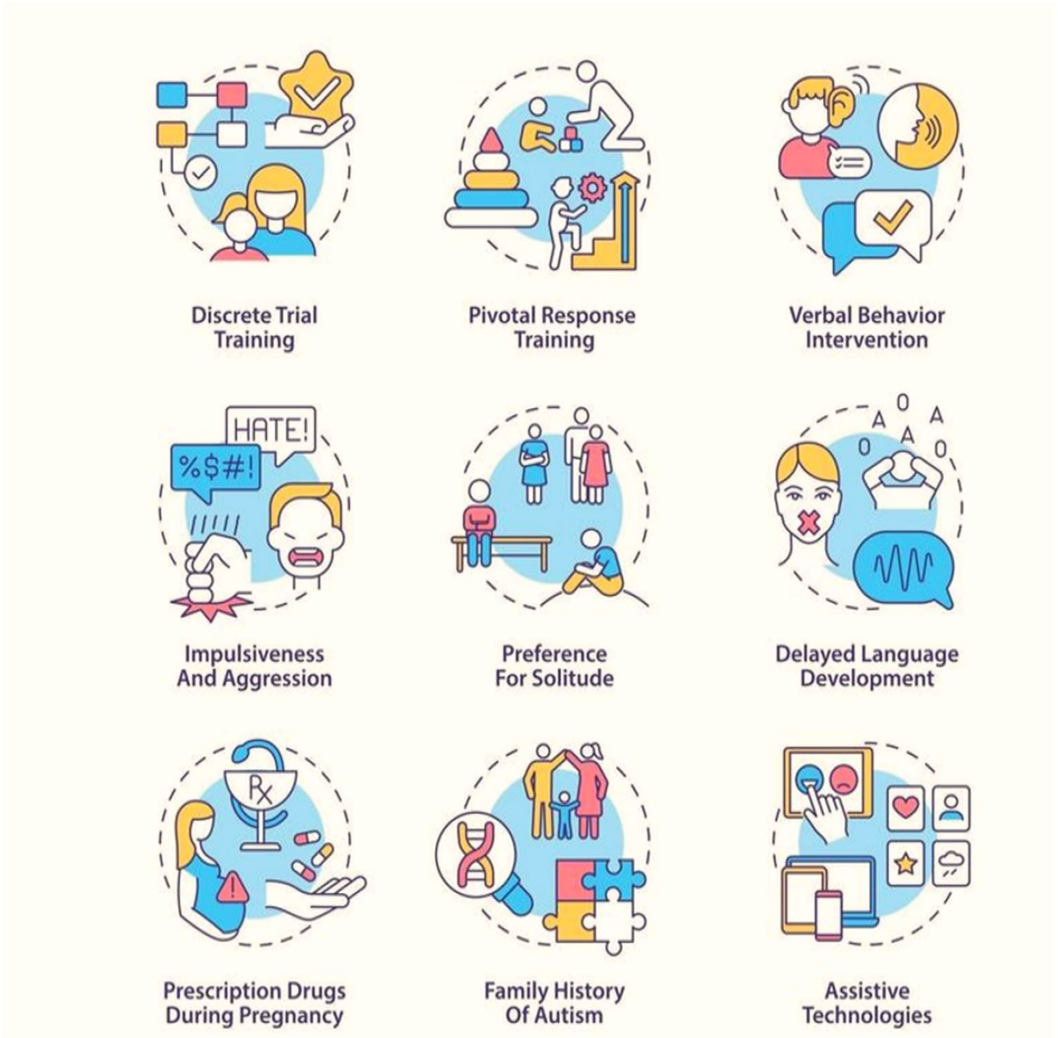


FIGURE 1
Displays the ASD concept.

potential to give non-invasive, continuous monitoring, which might help with the early diagnosis of ASD and shed light on neurological and behavioral traits that have been hard to document reliably.

Nevertheless, advancements in contemporary research are required to substantiate their efficacy. Sensor-based techniques may facilitate the identification of stereotyped behaviors and motor patterns linked to ASD in realistic environments, potentially yielding data that could guide timely and customized therapies (11). Neuroimaging and microbiome analysis further advance this technical domain by indicating neurological and biological traits specific to ASD. AI-enhanced neuroimaging aids in identifying structural and functional brain connection patterns associated with ASD, thereby enhancing the understanding of its neuroanatomical foundation (12).

The research conducted by Neeharika and Riyazuddin et al. (13) aimed to enhance the accuracy of ASD screening by using feature selection methods in conjunction with sophisticated machine learning classifiers. Their research included several datasets spanning infants, children, adolescents, and adults, enabling a thorough assessment of ASD characteristics across different age demographics. Authors' use of MLP model capacity to reliably and rapidly identify ASD, indicating a beneficial screening instrument suitable for various age groups, facilitating both clinical evaluations and extensive screenings. Wall et al. (14) investigated machine learning (ML) algorithms for diagnosing ASD using a standard dataset. The researchers focused on the Alternating Decision Tree classifier to identify a limited yet efficient set of queries that optimize the diagnostic procedure. Alzakari et al. (15) proposed a novel two-phase methodology to tackle the variability in ASD features with ML approaches, including behavioral, linguistic, and physical data. The first step concentrates on identifying ASD, using feature engineering methodologies and ML algorithms, including a logistic regression (LR) and support vector machine (SVM) ensemble, attaining a classification with high accuracy. EEG assesses brain activity and may identify children predisposed to developing ASD, hence facilitating early diagnosis. EEG data is used to compare ASD and HC (16–18). In (19), the CNN model was used for classification after transforming the data into a two-dimensional format. While EEG may facilitate the diagnosis of ASD, it is constrained by other factors, such as signal noise.

The research has used social media to investigate ASD. However, exploiting these prevalent platforms and innovative online data sources may be feasible to enhance the comprehension of these diseases. Previous research has utilized Twitter data to investigate discussions on ASD-related material, indicating that this subject is frequently addressed on this platform (20). Considering the use of social media for researching ASD is particularly significant, as a recent analysis indicated that around 80% of individuals with ASD engage with prominent social media platforms (21). This study aims to build upon previous research and enhance our comprehension of whether publicly accessible social media data from Twitter may provide insights into the existence of digital diagnostic indicators for ASD (22). Furthermore, we want to assess the viability of establishing a digital phenotype for ASD using social media.

Beykikhoshk et al. (20) examined Twitter's potential as a data-mining tool to comprehend the actions, challenges, and requirements of autistic individuals. The first finding pertained to the attributes of participants inside the autism subgroup of tweets, indicating that these tweets were highly informative and had considerable potential usefulness for public health experts and policymakers. Tomeny et al. (23) examined demographic correlations of autism-related anti-vaccine opinions on Twitter from 2009 to 2015. Their results indicated that the frequency of autism-related anti-vaccine views online was alarming, with anti-vaccine tweets connecting with news events and demonstrating geographical clustering. From 2015 to 2019, Tárraga-Mínguez et al. (24) examined the phrases "autism" and "Asperger" in Spain in relation to Google search peaks. The public view of autism was significantly impacted by how the condition was portrayed in the news and on social media, and the authors found that social marketing campaigns had a significant role in normalizing autism. In this research (25), looked at how people sought assistance. The results showed a strong correlation in Google search interest between the terms "Asperger syndrome" and "Greta Thunberg," reaching their highest point in 2019. Online traffic to the Asperger/Autism Network and Autism Speaks websites increased steadily from June to December 2019, indicating a correlation between help-seeking behavior and Thunberg's fame, according to the research. According to the results, the stigma associated with Asperger's disorder may have been positively affected by Thunberg's public exposure.

1.1 Contribution

The use of tweets from Twitter for the detection of ASD is substantial, since it offers extensive, real-time, user-generated data that facilitates the early identification of ASD-related behaviors, particularly via self-reported experiences and parental observations. This methodology promotes the advancement of suggested models, namely LSTM, CNN-LSTM, and inspired DDQN, for natural language processing to examine linguistic patterns, feelings, and keywords related to ASD. It provides insights into popular views, stigma, and misconceptions around autism, guiding awareness initiatives and public health measures. Twitter data is a powerful and accessible resource for enhancing early detection and understanding of ASD in diverse groups. Utilizing social media in this manner may offer more accessible and timely screening, particularly in regions with limited healthcare resources.

2 Materials and methods

Figure 2 shows the pipeline of the proposed system to provide a broader perspective to researchers and developers. The framework delineates the processing phases for the pipeline that utilizes social media content to diagnose ASD. Below, we present a comprehensive assessment of each step.

2.1 Dataset

To help with the early diagnosis of ASD by using proposed systems, the TASD-Dataset includes comprehensive textual sequences that depict the everyday lives of children with and without ASD. It offers new elements, including Noise Sensitivity, Sharing Interest, Sign Communication, and Tiptoe Flapping, It combines critical ASD assessment aspects like Attention Response, Word Repetition, and Emotional Empathy, as shown in Figure 3. Parents may get detailed insights and better identify signs of autism spectrum disorder (ASD) due to the deepening of certain behaviors. The dataset contains 172 tweets from the ASD class and 158 non-ASD tweets. Figure 4 shows the class of the dataset.

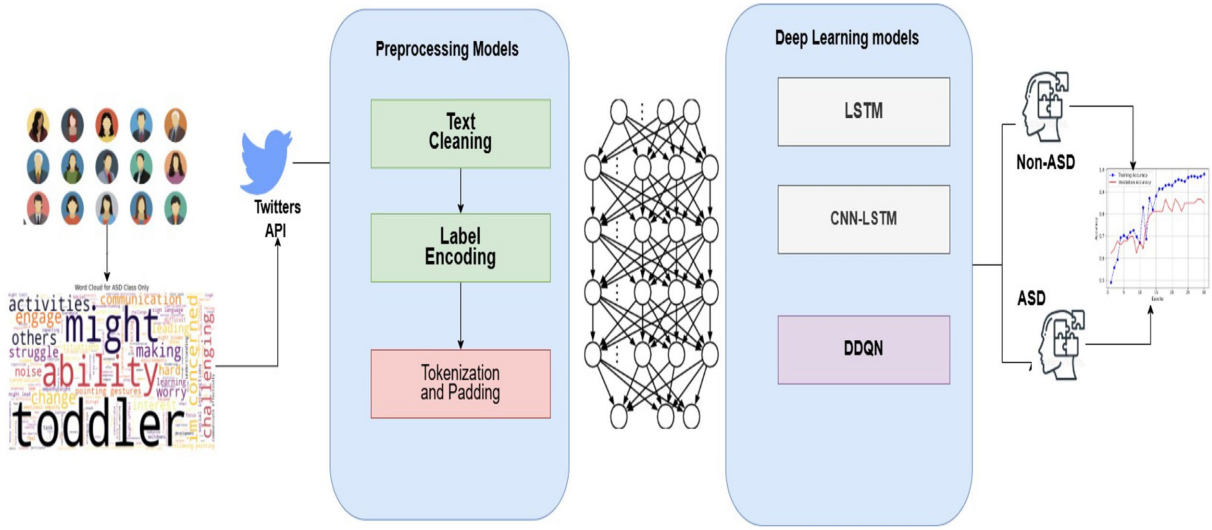


FIGURE 2
Farmwork of ASD system.

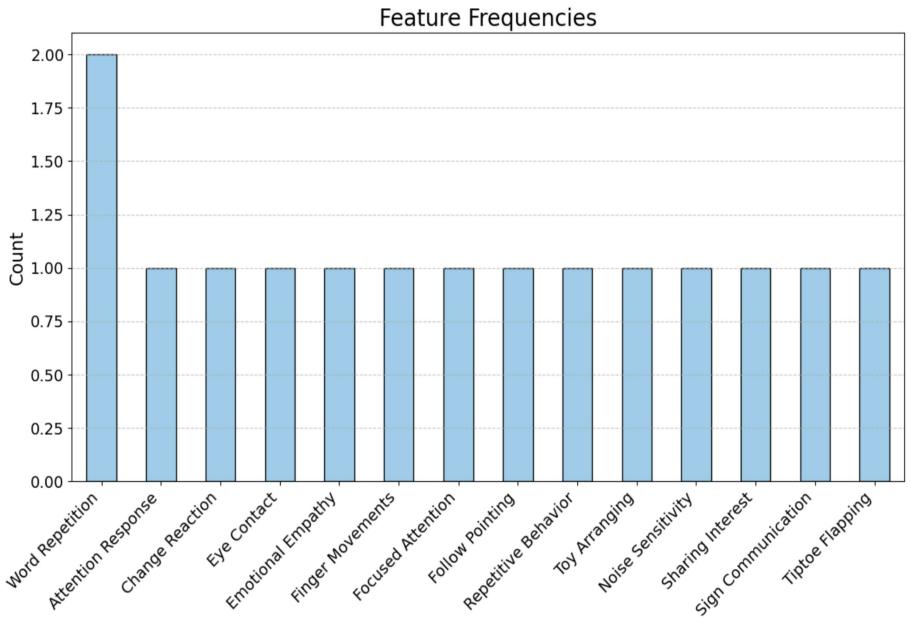


FIGURE 3
Features of the dataset.

2.2 Preprocessing

Text preprocessing is an essential step in the text processing process. Words, sentences, and paragraphs can all be found in a text, which is defined as a meaningful sequence of characters. Preprocessing methods feed text data to a proposed algorithm in a better form than in its natural state. A tweet can contain different viewpoints on the data it represents. Tweets that have not been preprocessed are highly unstructured and contain redundant data. To address these issues, several steps are taken to preprocess tweets for detecting ASD, as shown in Figure 5.

2.3 Text cleaning

The clean text preprocessing method is a significant step in text datasets because the text contains several extra contexts to preprocess and normalize raw text data for analysis. In these

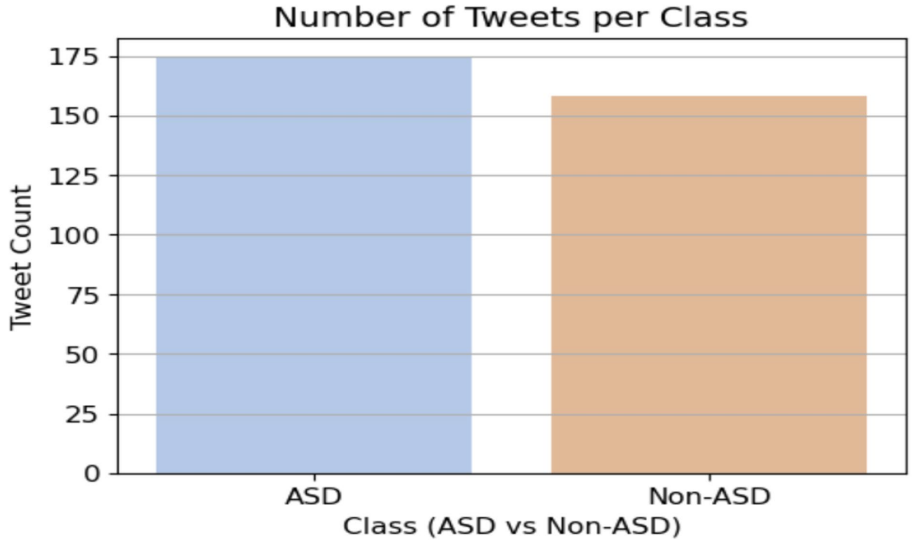


FIGURE 4
Label of the dataset.

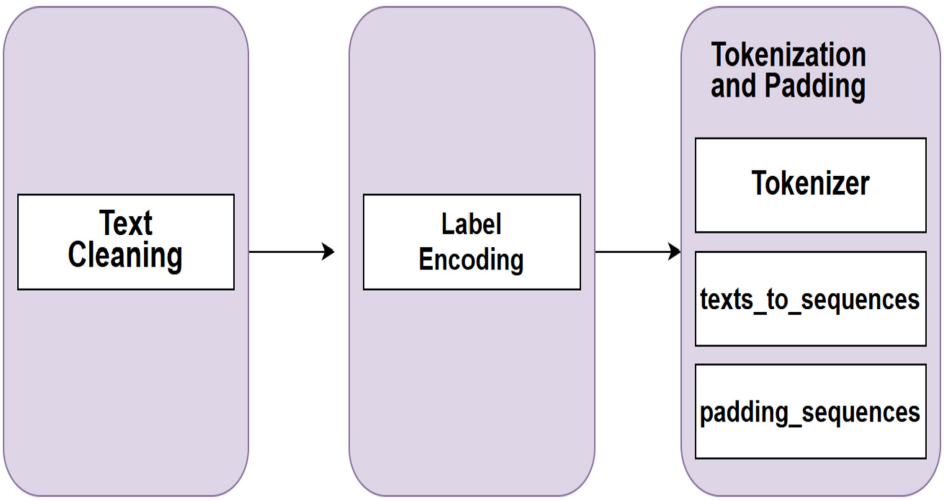


FIGURE 5
Preprocessing ASD text analysis.

steps, the use is transformed to lowercase to guarantee consistency and prevent differentiation between "ASD" and "Non-ASD." Subsequently, any characters that are not letters, numerals, or spaces are eliminated by a regular expression, so punctuation and other symbols that might create extraneous noise are removed. This method is ultimately applied to the 'Text' column of the Data Frame, ensuring that all text elements are sanitized and prepared for feature extraction. Figure 6 displays the clean text process.

2.4 Label encoding

The LabelEncoder method converts text class (ASD and Non-ASD) into numbers, designating 0 for ASD and 1 for Non-ASD. This transformation updates the classification effort by enabling the model to see the labels as numerical values instead of text. Equations 1, 2 show the label encoding.

$$yclassification \in \left(ASD, Non-ASD \right) Then \tag{1}$$

$$y = labelEncoder\left(yclassifcation \right) \rightarrow y \in \left\{ 0,1 \right\} \tag{2}$$

2.5 Tokenization and padding

Tokenization and padding are essential NLP preprocessing procedures that transform unprocessed text into a numerical representation appropriate for machine learning models,

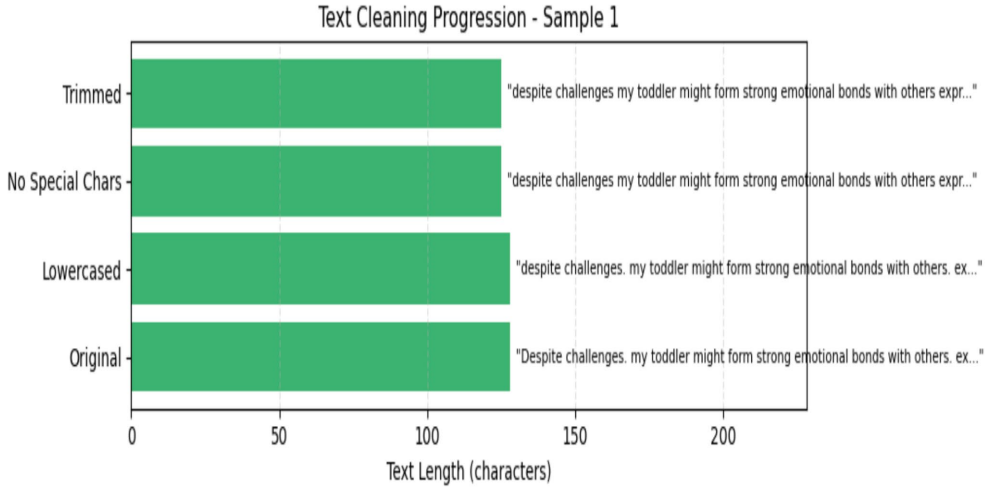


FIGURE 6
Clean text.

particularly neural networks. Figure 7 shows the tokenization and padding Equation 3.

2.5.1 Tokenizer

Tokenizer procedures transform textual data into a numerical representation suitable for input into neural networks. They convert a text corpus into integer sequences, assigning a distinct index to each unique word according to its frequency, as shown in Equation 3. The tokenizer processing is shown in Figure 8.

$$index(w) = rank_f(w)\ if\ rank_f(w) \le V \tag{3}$$

Where $rank_f(w)$ is rank w frequency $f(w)$ and V is the maximum number of words.

2.5.2 Fit texts

This phase is crucial for transforming unprocessed text into numerical sequences suitable for input into the proposed system.

2.5.3 Texts_to_sequences

To convert unprocessed text input into sequences of word indices according to the mapping acquired via as shown in Equation 4.

$$sequence(T_i) = \left[index(w_1), index(w_2), \ldots\ldots, index(w_m)\right] \tag{4}$$

Where is the T_i is the sentence of the text contained, and w is the words of the text, whereas the $index(w_1)$ is an index of the words in the context.

2.5.4 Padding_sequences

Normalize sequence lengths, which may differ post-tokenization, by padding shorter sequences and truncating larger ones to a predetermined length as shown in Equation 5. The padding and truncated b are fixed on the length. $L = 200$. The padding processing is shown in Figure 7.

$$x = \begin{vmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ . \\ y \end{vmatrix} \in \mathbb{R}^{nxL} \tag{5}$$

Where x is features contain padding and are tokenized, L is the length of the vector. The number of texts is indicated n, and $\in \mathbb{R}^{nxL}$ is matrix lues.

2.6 Proposed systems

2.6.1 Convolutional neural networks

The CNN model is at the core of all advanced machine learning and deep learning applications. They can successfully address text classification, image recognition, object identification, and semantic segmentation. Using the same method with a task as different as Natural Language Processing is counterintuitive (7). The structure is presented in Figure 9. Equation 6 presents the convolution layer of CNN.

$$O(x,y) = \sum_{i=1}^{H}\sum_{j=1}^{W} I(x+i, y+j) * K(i,j) + b \tag{6}$$

Where the features of text $O(x,y)$The feature of the text is mapped by using.$I(x+i, y+j)$ is weighted by a neural network and b is biased to adjust the neural. The ReLU activation function is Equation 7, the max pooling function is presented in Equation 8. The Dense Layer is given in Equation 9.

$$f(x) = \max(0, x) \tag{7}$$

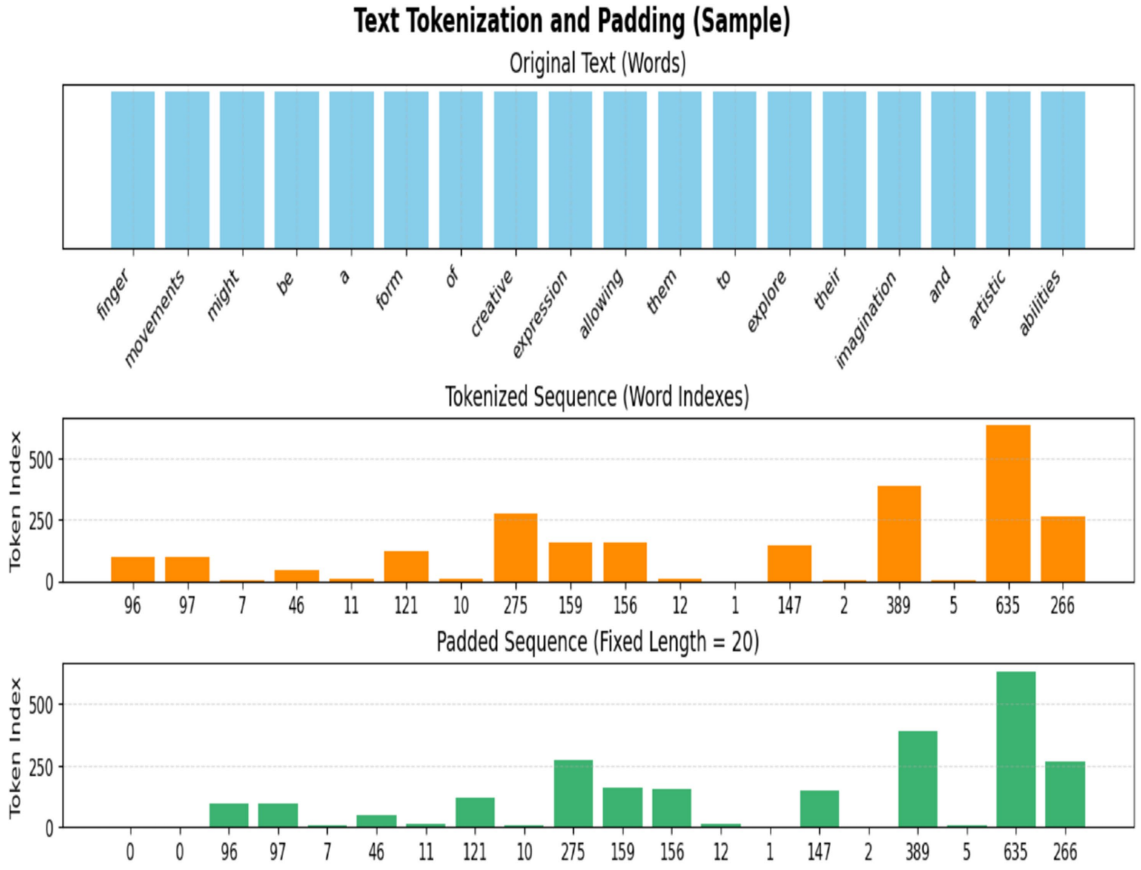


FIGURE 7
Sample of text tokenization and padding.

$$O(x,y)=\sum_{i=1}^{H}\sum_{j=1}^{W} I(x+i,y+j)*K(i,j)+b \tag{8}$$

$$\mathrm{O}=W\cdot X+b \tag{9}$$

2.6.2 Long short-term memory network

An LSTM network is an advanced form of a sequential neural network. It fixes the problem of RNN gradients fading over time. RNNs often handle long-term storage. At a high level, the operation of an LSTM is comparable to that of a single RNN neuron. The inner workings of the LSTM network are outlined in this section. The LSTM consists of three parts, each performing a particular function, as seen in Figure 10 below. In the first step, it is decided whether the information from the previous time stamp is significant enough to be saved or if it is harmless enough to be deleted. In the second step, the cell will try to acquire new information by analyzing the data that has been presented to it. In the third and final step, the cell incorporates the data from the most recent time stamp into the data stored in the next time stamp. These three components constitute what is referred to as a gate for an LSTM cell. The "Forget" gate comes first, followed by the "Input" section, and then the "Output" section is used to define the last portion as shown in Equations 10– 14.

$$\text{Forget gate}: f_t=\sigma\left(W_f.X_t+W_f.h_{t-1}+b_f\right) \tag{10}$$

$$\text{Input gate}: i_t=\sigma\left(W_c.X_t+W_i.h_{t-1}+b_i\right) \tag{11}$$

$$\text{Cell gate}: C_t=\left(W_f*\left(.h_{t-1},x_t\right)b_f\right) \tag{12}$$

$$\text{Output gate}: o_t=\sigma\left(W_o+X_t+W_o.h_{t-1}+V_o.C_t+b_o\right) \tag{13}$$

$$\text{Hidden layer}: h_t=o_t+\tanh\left(C_t\right) \tag{14}$$

In Figure 10, C_t Represent the prior and current states of the cell, respectively. Both h_{t-1} and h Represents the cell output that was processed before the one now being processed. It is common practice to disregard f_t As a gate, even though it is the input gate. The output of a sigmoid gate is symbolized here by o_t. The cable that connects the cell gates is where all the data collected by the cell gates is sent to and from C. The f_t Layer decides to remember anything, and the f_t*The* Output is multiplied by c to do so (t-1). After that, c (t-1) is multiplied by the product of the sigmoid layer gate and the tanh layer gate, and the output h t is generated by point-wise multiplication of o_t and tanh.

The LSTM architecture is intended to capture long-term relationships in Twitter text data. The preprocessing converts input

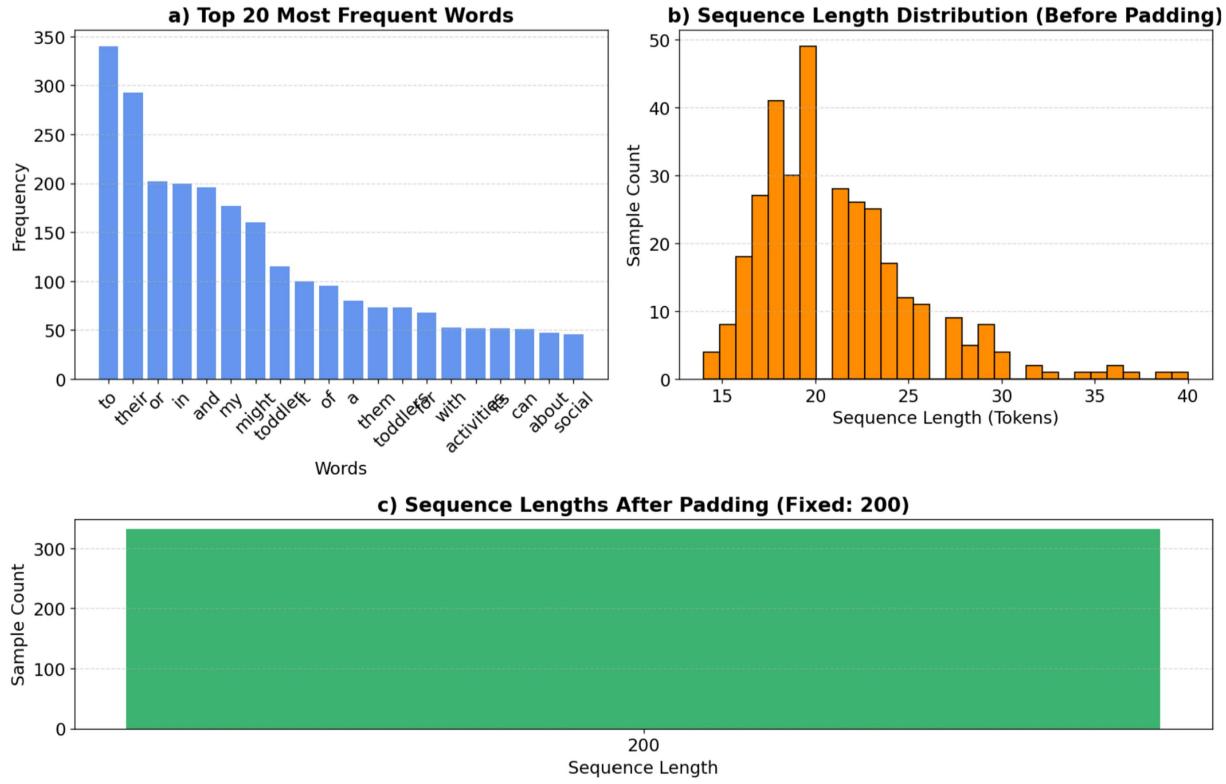


FIGURE 8
Tokenizer analysis: word frequencies and sequence lengths.

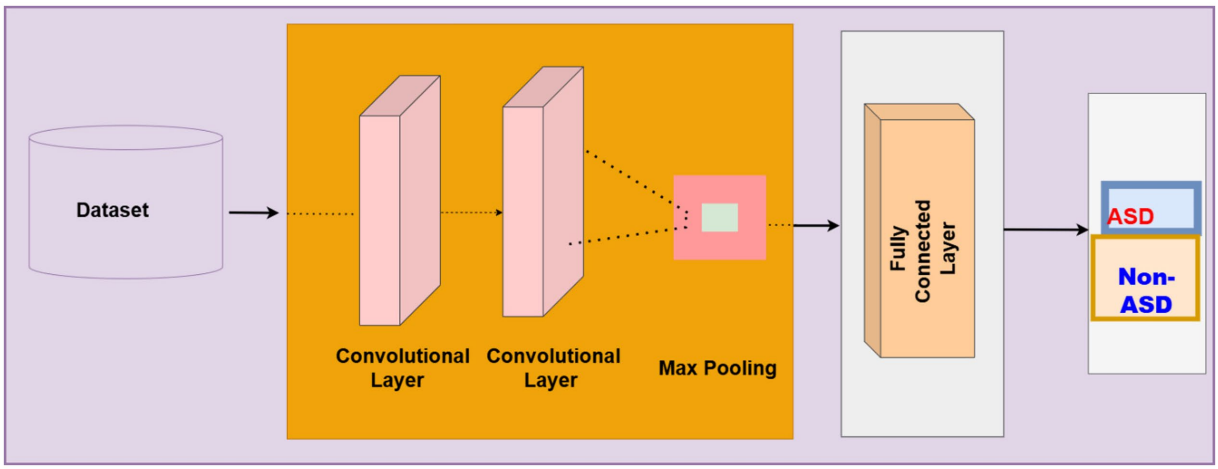


FIGURE 9
Structure CNN.

words that start with an embedding layer into 128-dimensional dense vectors. The LSTM layer with 64 units is then used to mitigate overfitting, integrating dropout and recurrent dropout with 0.5. An L2 regularization term is further included in the LSTM and output dense layer. Table 1 shows parameters of the LSTM model.

2.6.3 CNN-LSTM model

The CNN-LSTM model is a hybrid architecture that combines convolutional neural networks (CNN) for spatial feature extraction and long short-term memory (LSTM) networks for sequential learning, making it highly effective for analyzing text data such as tweets. The model begins with an embedding layer

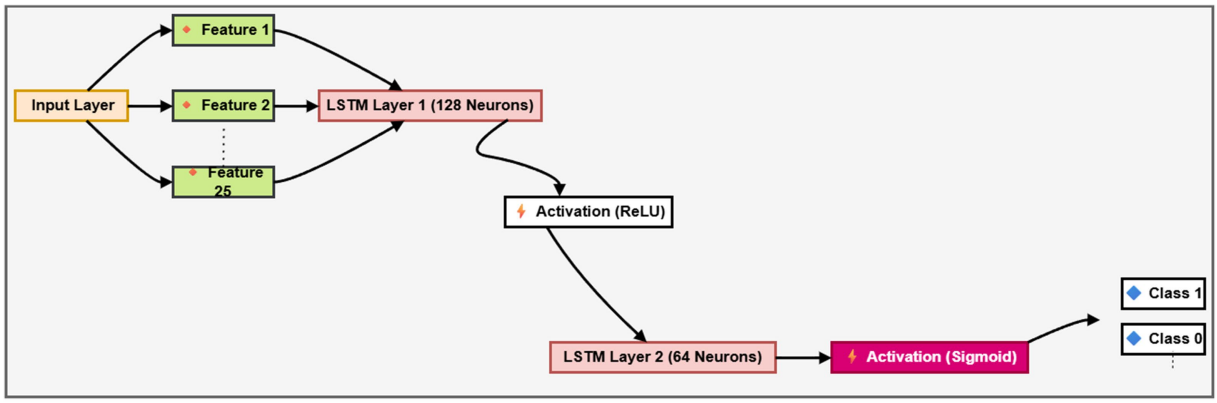


FIGURE 10
LSTM model.

TABLE 1 LSTM parameters model.

Input	Values
Embedding dimension	256
LSTM unit	64
Conv1D	64, K = 5
MaxPooling ID	yes
Dropout_rate	0.5
Dense_Unites	32
Activation_function	ReLU
L2	0.001
Optimizer	Adam
Loss	Binary
Epoch	30
Batch size	16

TABLE 2 CNN-LSTM parameters.

Input	Values
Embedding dimension	128
LSTM unit	64
Conv1D	No
MaxPooling ID	No
Dropout_rate	0.5
Dense_Unites	32
Activation_function	ReLU
L2	0.001
Optimizer	Adam
Loss	Binary
Epoch	30
Batch size	16

that transforms each word into a 256-dimensional dense vector, capturing the semantic meaning of words. This is followed by a 1D convolutional layer with 64 filters and a kernel size of 5, which scans through the text to detect local patterns and n-gram features such as common word combinations or phrases often associated with ASD. A batch normalization layer is applied to stabilize and accelerate training, followed by a max pooling layer that reduces the dimensionality and computational load by selecting the most prominent features. A dropout layer with a rate of 0.5 is then used to prevent overfitting by randomly deactivating some neurons during training. The output is passed into a 64-unit LSTM layer that captures the temporal dependencies and contextual relationships across the tweet sequence. Finally, a dense layer with sigmoid activation performs binary classification to predict whether the tweet indicates ASD-related content. The model is trained using the Adam optimizer, binary cross-entropy loss, class weights, and regularization to handle imbalanced data and improve generalization. The critical parameters of the CNN-LSTM model are displayed in Table 2.

2.6.4 Double deep Q-network (DDQN-inspired)

The Double Q-Learning model was introduced by H. van Hasselt in 2010, addressing the issue of significant overestimations of action value (Q-value) inherent in traditional Q-Learning. In fundamental Q-learning, the Agent's optimal strategy is consistently to select the most advantageous action in any specific state. This concept's premise is that the optimal action corresponds to the highest expected or estimated Q-value. Initially, the Agent lacks any knowledge of the environment; it must first estimate Q(s, a) and subsequently update these estimates with each iteration. The Q-values exhibit considerable noise, leading to uncertainty about whether the action associated with the highest expected or estimated Q-value is genuinely the optimal choice.

Double Q-Learning employs two distinct action-value functions, Q and Q', as estimators. Even if Q and Q' exhibit noise, this noise can be interpreted as a uniform distribution as shown Figure 11 The update procedure exhibits some variations compared to the basic version. The action selection and action evaluation processes are separated into two distinct maximum function estimators. shown in Equations 15, 16.

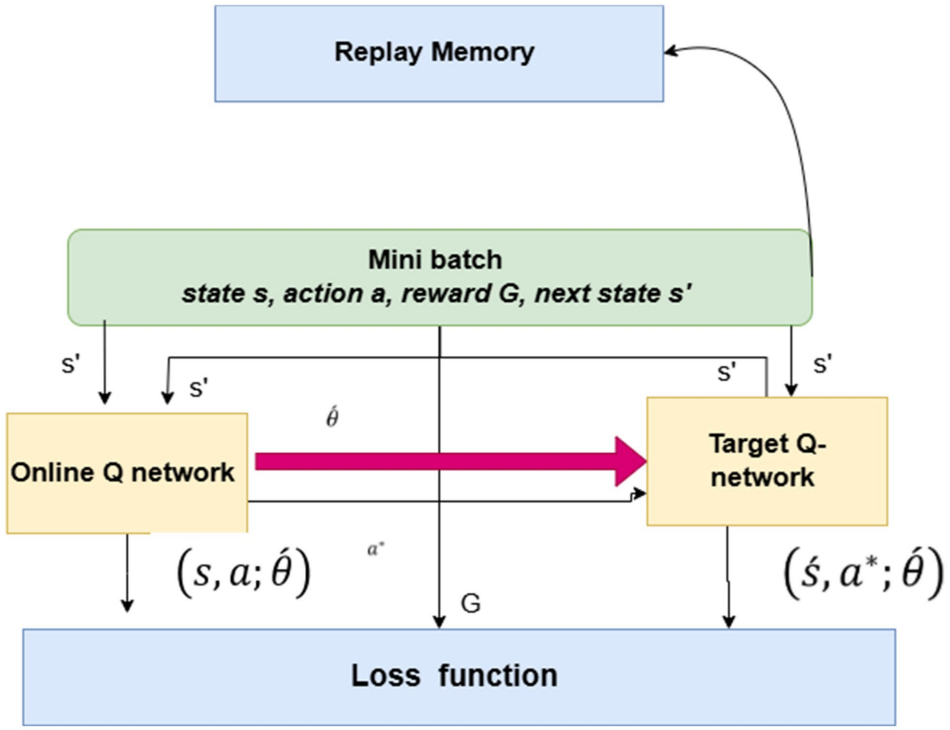


FIGURE 11
DDQN-inspired model.

Let the vector of a neural network's weights be represented by θ. We establish two Q-networks: the online Q-network Q (s, a; θ(t)) and the target Q-network Q (s, a; θ (t)). To be more specific, the training of Q (s, a; X (t)) is done by modifying the weights (t) at time slot t in relation to the goal value y(t).

$$y(t) = G(t) + \left(s', \arg\max Q\left(s', a^*; \theta'(t)\right); \theta'(t)\right) \quad (15)$$

$$y(t) = G(t) + \left(s', \arg\max Q\left(s', a^*; \theta'\right); \theta_{i-1}\right) \quad (16)$$

The reinforcement learning mechanism integrates generative artificial intelligence for decision-making and prediction tasks, as shown in Equations 15, 16. This equation indicates the generative which produces the estimation or hypothesis at a given time t. *Double Q – Learning* Used next state, whereas the s' is exit state and $\arg\max Q(s', a^*; \theta'(t))$ defined as the action of a^* to maximize the predicted Q-value based on the current parameters. To estimate the Q-value of this selected action in the next state, the outer Q-function Q' employs the older parameters. $\theta_{i-1}1$, which helps reduce overestimation bias. This combination makes applications for predicting ASD from social media content domains possible.

The DDQN model is used to classify ASD and non-ASD cases utilizing text data. The model utilizes a preprocessing step for text processing that encompasses data loading, cleaning (including lowercasing, removal of special characters, and normalization of spaces), and tokenization, constrained by a maximum vocabulary of 10,000 words and a sequence length of 200. The model architecture, drawing from the Double Deep Q-Network (DDQN) model comprises an input layer, an embedding layer with 256 dimensions, and two parallel LSTM branches, each containing 64 units, a dropout rate of 0.5, and L2 regularization to capture sequential patterns effectively. The model uses the Adam optimizer with a learning rate of 1e-4 and employs binary cross-entropy loss. It is trained for 30 epochs, incorporating early stopping and learning rate reduction callbacks to mitigate overfitting. Parameters of DDQN-Inspired are shown in Table 3.

3 Performance of the framework

3.1 Performance of LSTM

Figure 12 presents the accuracy and loss metrics used to train and validate an LSTM model over 30 epochs. The validation accuracy of the LSTM model, displayed in red, begins at a lower value and increases to about 81%. The blue line in the accuracy plot (a) shows the training accuracy of the LSTM model; it increases gradually from around 50% to almost 99%, showing that the model learns the training data well over time. The plot (b) shows the loss of the LSTM model; the blue line represents the training loss, which drops gradually from around 0.7 to less than 0.2, suggesting that the model is getting a better fit to the training data. Meanwhile, the red validation loss line declines from around 0.7 to about 0.3. While the training loss continues

to grow, the validation loss reaches a level and exhibits small oscillations, suggesting that the model's generalizability may stabilize.

The ROC curve illustrated in Figure 13 shows the efficacy of the LSTM model in differentiating between the classes. The graph illustrates the TP rate (sensitivity) in relation to the FP Rate across different threshold levels. The LSTM model attains an AUC of 0.95, demonstrating exceptional classification capability. The AUC of 1.0, but a result of 0.5 indicates random chance.

3.2 Performance of the CNN-LSTM model

Figure 14 presents plots illustrating the performance of a CNN-LSTM model over 25 epochs, showing its training and validation metrics for accuracy and loss. The accuracy plot (a) illustrates the training accuracy (blue line), which increases progressively from approximately 51.42% to nearly 99.53%, indicating effective learning from the training data. In contrast, the validation accuracy (red line) rises to about 83.02% with some variability, indicating satisfactory but imperfect generalization. The loss plot (b) shows the training loss (blue line) declining steadily from 0.7140 to below 0.0760, indicating enhanced model fit. In contrast, the validation loss (red line) decreases from 0.7130 to approximately 0.3530, with a slight decline toward the conclusion. This notification indicates that the CNN-LSTM model demonstrates efficient learning, as evidenced by the difference between the training and validation measures.

Figure 15 illustrates the ROC curve for the CNN-LSTM model, illustrating its classification performance at various thresholds. The graph illustrates the TP Rate (Sensitivity) in relation to the FP Rate, with the AUC recorded at 92%. The elevated AUC value indicates the model has robust discriminative capability in differentiating between the ASD and Non-ASD classes. The ROC ascends rapidly toward the top-left corner, as seen in the figure, indicating a high TP rate with few false positives.

TABLE 3 Parameters of DDQN-inspired.

Input	Values
Max-sequence length	200
Vocabulary	10,000
Embedding_dimension	256
Dropout_rate	0.5
Dense_Unites	32
Activation_function	ReLU
LS	0.0001
Optimizer	Adam
Loss	Binary
Epoch	30
Batch size	16

3.3 Performance of DDQN-inspired model

Graphs 16 illustrate the performance of a DDQN throughout 30 epochs. The accuracy plot (a) demonstrates that the training accuracy increases from around 58.02% to almost 98.58%, indicating the DDQN model successful learning from the training data over time. The validation accuracy of the DDQN is about 87, showing the best performance compared to different models like LSTM and CNN-LSTM. The plot (b) illustrates that the training loss decreases from about 0.8155 to around 0.1477, indicating a robust fit to the training data. The validation loss begins at 0.3831 with many fluctuations throughout (Figure 16).

Figure 17 shows the ROC curve for the DDQN model; it shows a visual representation of its classification capability, with the curve toward the top-left corner, indicating strong predictive power. The AUC value of the DDQN model is 96%, demonstrating that the model can distinguish between the positive and negative classes.

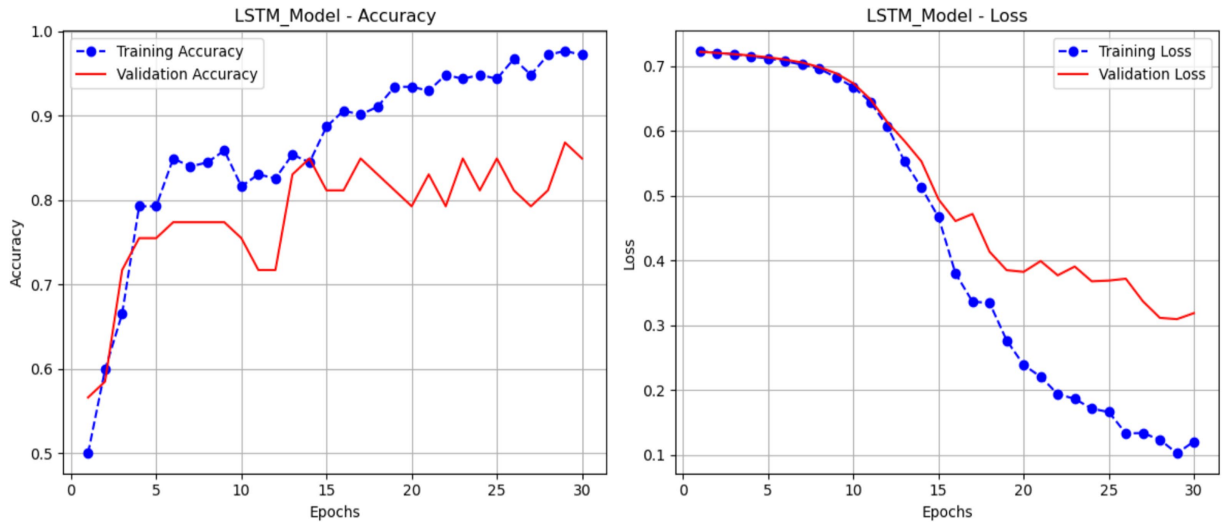


FIGURE 12
Performance of the LSTM model.

4 Experiment and discussion results

Both the Jupyter deep learning framework and the Windows 10 operating system were utilized during the testing process. Experiments were conducted using a machine with 16 gigabytes of RAM and an Intel Core i7 central processing unit. The input dimensions of the experiment were a standard text dataset collected from the Twitter API related to ASD. The test was utilized in our database, while the remaining 20% was used as part of our validation set. The three DL models, namely LSTM, CNN-LSTM, and DDQN-Inspired, were proposed for detecting ASD from social media content.

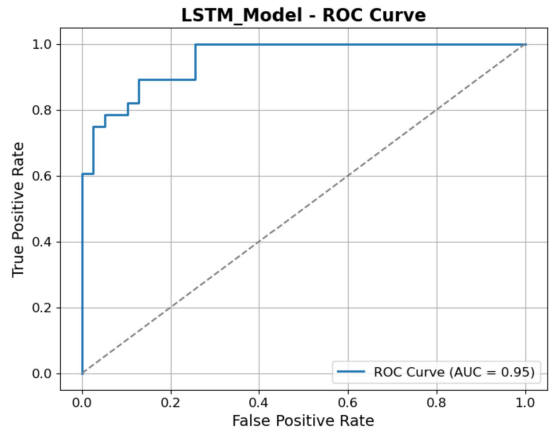


FIGURE 13
ROC of the LSTM model.

4.1 Measuring the model's performance

Sensitivity, specificity, accuracy, recall, and F1 scores are assessment measures used to determine how successfully the algorithms identify ASD. The related equations from 17 to 21:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \quad (17)$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \quad (18)$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (19)$$

$$specificity = \frac{TN}{TN + FP} \times 100 \quad (20)$$

$$F1 - score = 2 * \frac{precision \times Sensitivity}{precision + Sensitivity} \times 100 \quad (21)$$

4.2 Result of the LSTM model

The classification LSTM model, presented in Table 4, summarizes its performance in differentiating between ASD and Non-ASD patients, attaining an overall accuracy of 81%. The LSTM model demonstrates in ASD class a precision of 91%, indicating a high accuracy in identifying predicted ASD cases. The LSTM with recall metric scored 77% and an F1-score of 82% for detecting the ASD class. The LSTM model with Non-ASD class demonstrates a precision of 71%, a recall of 89%, and an F1-score of 79%, to identify Non-ASD cases. The macro average of the LSTM model for all metrics is (precision: 81%, recall: 82%, F1-score:

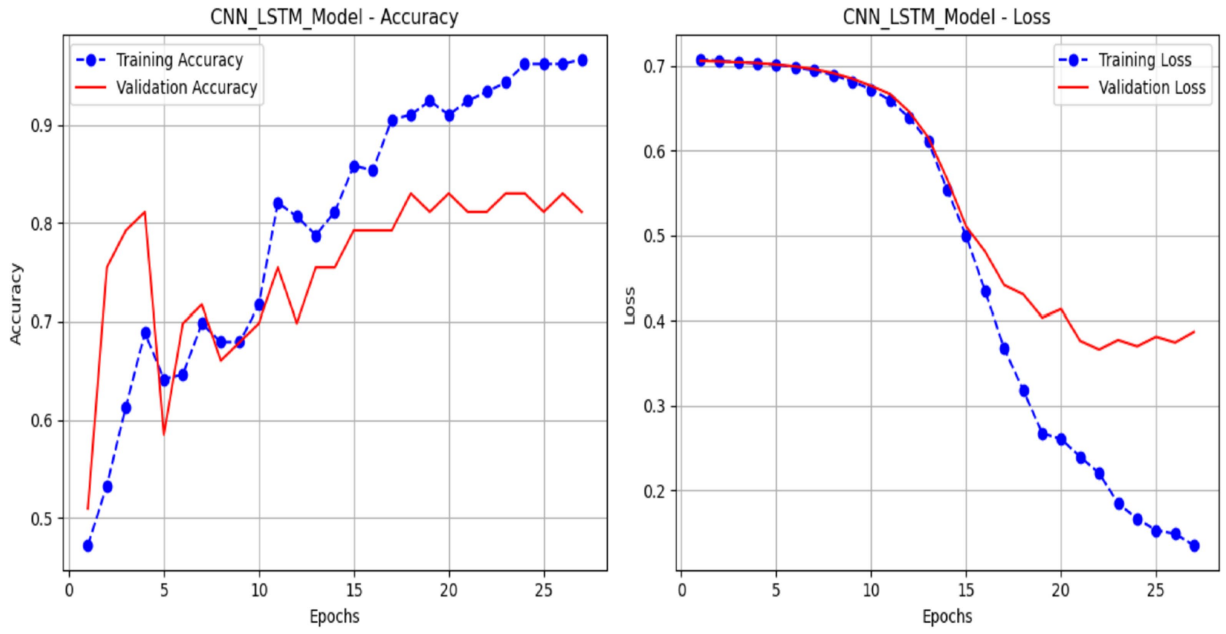


FIGURE 14
Performance of the LSTM model.

81%). LSTM model is recognized for its efficiency and scalability as a model for social media content.

The confusion matrix for the LSTM model is provided in Figure 18. It is presented in a clear manner. Among the confirmed ASD cases, 29 were accurately identified as ASD, whereas 10 were incorrectly classified as Non-ASD, indicating strong performance with minor errors. In the true non-ASD cases, 25 were correctly identified, while 3 were misclassified as ASD, suggesting a generally effective detection process. The deep blue and light shades produce a tranquil visual, illustrating the model's balanced approach in classifying the 67 total instances, demonstrating notable strength in identifying Non-ASD cases, while exhibiting marginally lower accuracy for ASD. This matrix effectively illustrates the LSTM model's systematic approach to managing sequential data, such as text or time-series inputs, in a clear and comprehensible manner.

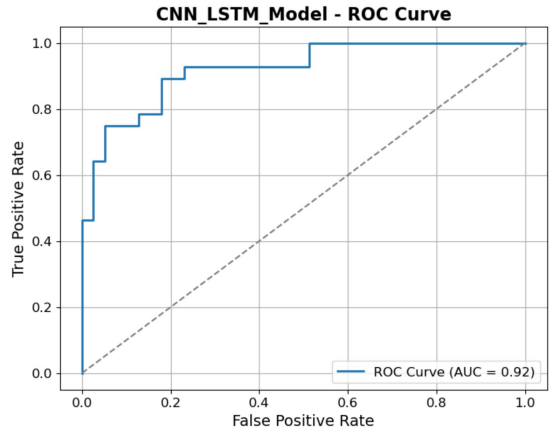


FIGURE 15
ROC of the CNN-LSTM model.

4.3 Result of the CNN-LSTM model

Table 5 displays the CNN-LSTM model's performance in distinguishing between ASD and non-ASD classes. The CNN-LSTM model attained an overall accuracy of 85% across the dataset. In the ASD label, a precision of 91% was achieved, a high percentage for predicting ASD cases that were accurately recognized. The recall indicates that the model identified 82% of all genuine ASD cases, resulting in an F1 score of 86%, better than the recall metric. The CNN-LSTM model attained 78% accuracy, 89% recall, and an 83% F1 score for the Non-ASD class. The macro average, representing the unweighted mean of precision, recall, and F1 score across both classes, was 85, 86, and 85%, respectively. The findings indicate that the CNN-LSTM model performs satisfactorily, exhibiting a marginally superior capacity to identify ASD cases relative to non-ASD cases accurately.

The confusion matrix of a CNN-LSTM model is presented in Figure 19, for classifying instances into ASD and Non-ASD. The matrix is structured with true labels on the vertical axis and predicted labels on the horizontal axis, providing a clear summary of the model's classification outcomes. The matrix shows that out of the instances truly labeled as ASD, the model correctly predicted 32 as ASD TP while 7 were incorrectly classified as Non-ASD FN. For the instances truly labeled as Non-ASD, the model accurately identified 25 as Non-ASD TN but 3 were misclassified as ASD FP. This indicates that the model demonstrates a relatively strong ability to correctly identify ASD and Non-ASD cases, with higher accuracy for true positives (32 out of 39 ASD cases) and true negatives (25 out of 28 Non-ASD cases). Overall, the model exhibits promising performance with minimal misclassification errors.

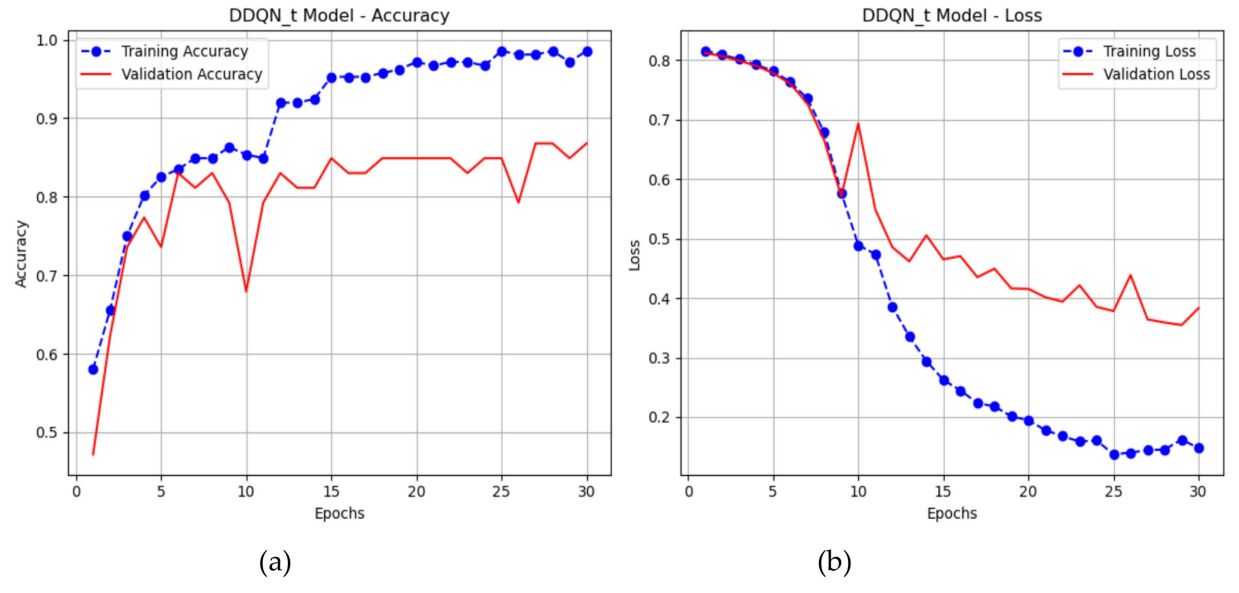


FIGURE 16
Performance of the *DDQN-inspired* model.

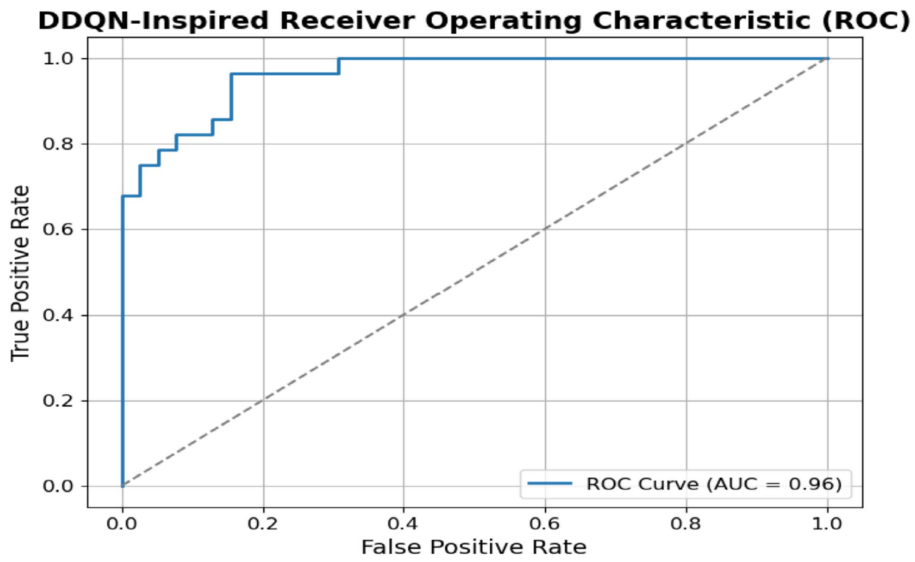


FIGURE 17
ROC *DDQN-inspired model.*

TABLE 4 LSTM results.

Class name	Precision (%)	Recall (%)	F1 Score (%)	Support
ASD	91	74	82	39
Non-ASD	71	89	79	28
Accuracy		81		
Macro Avg	81	82	81	67

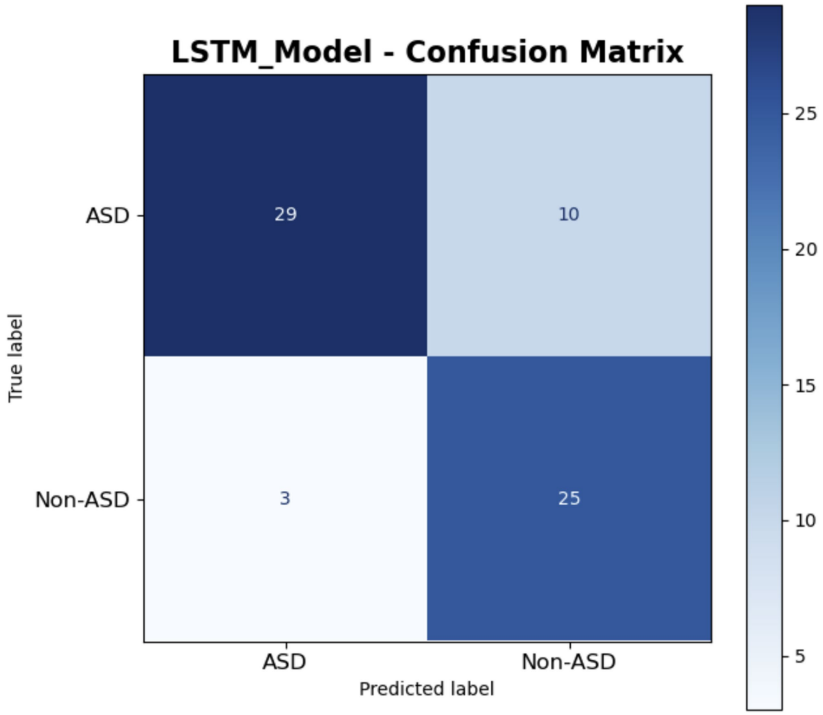


FIGURE 18
LSTM model.

TABLE 5 Results of the CNN-LSTM model.

Class name	Precision (%)	Recall (%)	F1 Score (%)	Support
ASD	91	82	86	39
Non-ASD	78	89	83	29
Accuracy		85		
Macro Avg	85	86	85	67

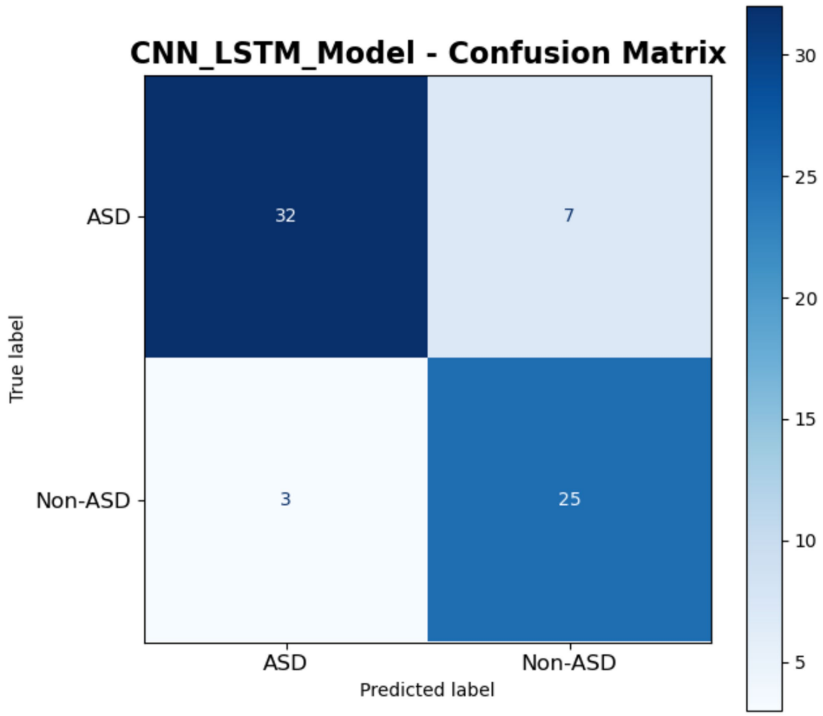


FIGURE 19
Results of CNN-LSTM model.

TABLE 6 Result of DDQN-inspired.

Class name	Precision (%)	Recall (%)	F1 Score (%)	Support
ASD	95	79	87	39
Non-ASD	77	96	86	28
Accuracy		87		67
Macro Avg	87	88	87	67

4.4 Results of double deep Q-network

The findings of the DDQN model are shown in Table 6, achieving a high precision of 87% compared to the other models. This finding demonstrates the potential of the proposed DDQN approach for identifying ASD based on social media content. Ultimately, the proposed system was compared against the existing one using the same dataset. The proposed approach may assist physicians in detecting ASD and conducting symptomology research in a natural environment, attaining an overall accuracy of 87. The model for the ASD class shows a precision of 95%, a recall of 79%, and an F1-score of 87%, indicating robust efficacy in accurately identifying ASD patients. The Non-ASD class has a precision of 77%, a recall of 96%, and an F1-score of 86%, indicating somewhat reduced accuracy with robust recall. The macro average measures (precision 87%, recall 88%, F1-score 87%) indicate performance across both classes.

The confusion matrix of the DDQN model is shown in Figure 20 for the classification task between ASD and non-ASD cases. For correct classification of ASD cases, the model correctly classified 31 instances as ASD, represented by the top-left quadrant (TP). However, the DDQN model, misclassified 8

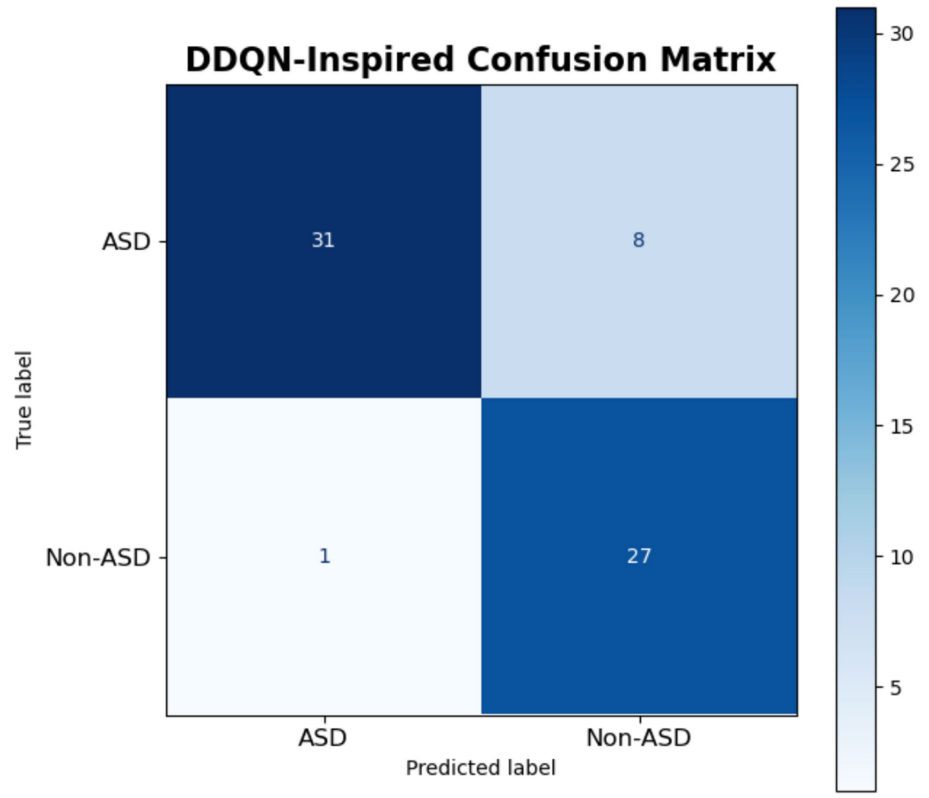


FIGURE 20
Result of DDQN-inspired model.

instances misclassifying true ASD cases as Non-ASD, shown in the top-right quadrant (FN). On the other hand, the DDQN showed the true Non-ASD cases, accurately identified 27 instances as Non-ASD, depicted in the bottom-right quadrant (TN). At the same time, 1 instance was incorrectly labeled as ASD, as shown in the bottom-left quadrant (FP). The confusion matrix of DDQN model highlights that it performs well overall, with a strong ability to correctly identify both ASD and Non-ASD cases, as evidenced by the high counts of TP (31) and TN (27).

In the digital era, people frequently write content on social media to express their feelings, opinions, beliefs, and activities. This makes social media one of the most significant sources of data generation, allowing you to explore its opportunities and challenges. Today, social media has become a mediator between people and the healthcare sector, enabling them to search for information about any specific disease and methods for diagnosing it.

Individuals within the mental health community use social media platforms such as Twitter to seek information, exchange experiences, and get assistance about ASD in an environment that is seen as more approachable and informal than conventional medical contexts. They often seek immediate, relevant information—whether to understand symptoms, identify coping mechanisms, or connect with others facing similar difficulties. Figure 21 illustrates that Word clouds are visual representations of text that highlight key terms and their frequency of use. We used WordCloud to compare ASD and Non-ASD texts for instances of word repetition.

The deployment model based on the Deep Q-Network (DQN) model for diagnosing ASD is shown in Figure 22.

Step 1: Data Collections, including cleaning, normalization, and tokenization.

Step 2: Model Development: The preprocessed data is used to train and validate a Deep Q-Network (DQN) model for classifying tweets as indicative of ASD or non-ASD patterns.

Step 3: Application Interface: An application interface is developed once the model has been trained. It integrates with users' Twitter accounts and continuously analyzes their tweets.

Step 4: Deployment: The proposed system is deployed in the cloud for storing tweets, enabling real-time monitoring of incoming tweets. Predictions are flagged for review by healthcare professionals, who validate the model's output before categorizing individuals as potentially having ASD or non-ASD.

This digital imprint may serve as an ancillary resource for mental health practitioners, providing insights into an individual's emotional state and social behaviors in a natural environment, potentially facilitating early detection or corroborating a diagnosis. This method is a non-invasive means of data collection, particularly beneficial for individuals who lack rapid access to clinical assessments due to financial constraints, stigma, or resource scarcity. However, it should not replace professional diagnoses and must be conducted with ethical consideration to prevent misunderstanding. Table 7 shows the findings of the proposed framework on the Twitter dataset. It demonstrates that the suggested method outperforms the current systems in terms of accuracy, proving its efficacy and potential for performance improvements.

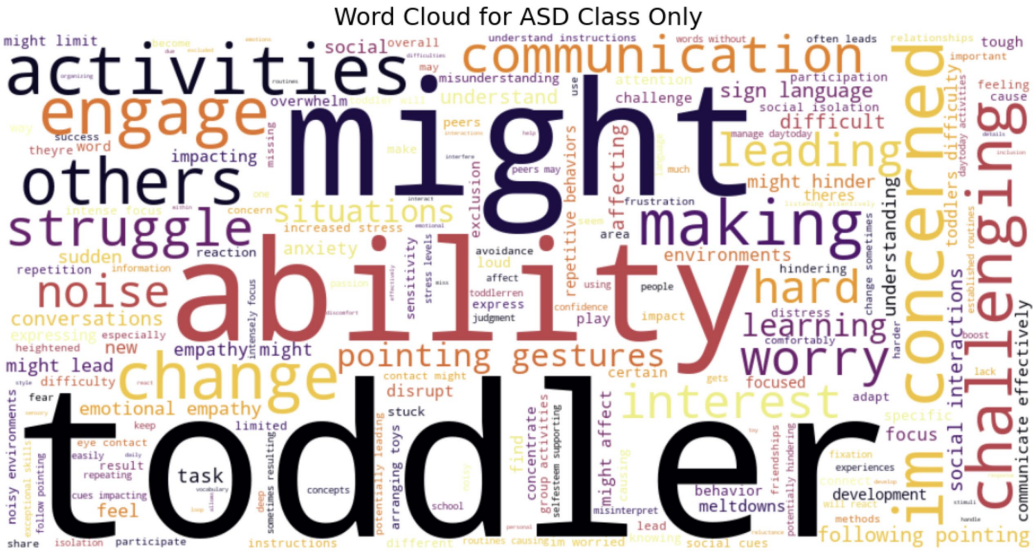


FIGURE 21
ASD word cloud.

FIGURE 22
Deployment system-based text for detecting ASD.

TABLE 7 Compared with the proposed ASD system.

References	Dataset	Model	ACC %
Rubio-Martín et al. (26)	Twitters dataset	BERT	84
Jaiswal and Washington (27)	Twitters dataset	ML	78

5 Conclusion

To assist people in identifying trends in their behavior, such as social challenges or sensory sensitivities, which may encourage them to pursue a formal diagnosis. The main objective of examining tweets for identifying ASD is its ability to provide behavioral and emotional indicators associated with the disorder. This research was used to analyze the textual analysis of tweets to detect the behaviors in self-identified autistic individuals relative to others. The suggested framework was evaluated using information from the social media platform "Twitter" collected from a public repository. Before examining the proposed system, several preprocessing steps must be implemented in the text. The 'Text' column is cleaned by converting it to lowercase, eliminating non-alphanumeric characters (excluding spaces) through regular expressions, normalizing whitespace to a single space, and removing any leading or trailing spaces. The ASD and Non-ASD labels are converted into a numerical format (0 or 1) with LabelEncoder to accommodate the binary classification requirement. Tokenization of the text data is performed using a tokenizer, restricting the vocabulary to 10,000 words, and then transforming the text into sequences of numbers. The sequences are padded to a standardized length of 200 tokens to maintain consistency for the proposed model input. The proposed data is ultimately divided into an 80% training and 20% testing ratio, and class weights are calculated to resolve any class imbalance. This preparation pipeline efficiently converts raw text data into a structured numerical representation appropriate for the proposed framework, while preserving academic integrity. The output of these preprocessing steps was processed using three DL models, such as Short-Term Memory (CNN-LSTM) and a Double Deep Q-network (DDQN). The results of these proposals were proven, revealing that the DDQN model achieved a high accuracy score of 87% with respect to the accuracy measure. The proposed framework, based on real textual data, can be helpful for real-time offering natural, behavioral, and emotional data that might indicate ASD-related characteristics. Finally, we have observed that social media (Twitter) postings include linguistic patterns, emotional expressions, and social interactions that can help official health officials detect ASD based on the thorough symptoms of ASD that are posted on the platform. This study utilized a conventional dataset sourced only from the Twitter network. We will emphasize the necessity of gathering datasets from many platforms to enhance the model's generalizability in the future.

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/s. The dataset used in this study can be found at https://data.mendeley.com/datasets/87s2br3ptb/1.

Ethics statement

Ethical approval was not required for the study involving human data in accordance with the local legislation and institutional requirements. Written informed consent was not required, for either participation in the study or for the publication of potentially/indirectly identifying information, in accordance with the local legislation and institutional requirements. The social media data was accessed and analysed in accordance with the platform's terms of use and all relevant institutional/national regulations.

Author contributions

NF: Supervision, Conceptualization, Software, Methodology, Writing – original draft, Investigation, Visualization, Formal analysis, Validation. AA: Data curation, Visualization, Methodology, Conceptualization, Validation, Software, Formal analysis, Writing – original draft. NE: Investigation, Writing – review & editing, Validation, Formal analysis. SA: Visualization, Software, Formal analysis, Writing – review & editing, Data curation.

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-2024-288.

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.

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

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. ASD. (2023). Available online at: https://www.healthline.com/health/signs-of-autism-in-3-year-old.

2. Matson JL, Goldin RL. What is the future of assessment for autism spectrum disorders: short and long term. *Res Autism Spectr Disord*. (2014) 8:209–13. doi: 10.1016/j.rasd.2013.01.007

3. Srivastava AK, Schwartz CE. Intellectual disability and autism spectrum disorders: causal genes and molecular mechanisms. *Neurosci Biobehav Rev*. (2014) 46:161–74. doi: 10.1016/j.neubiorev.2014.02.015

4. Alhujaili N, Platt E, Khalid-Khan S, Groll D. Comparison of social media use among adolescents with autism Spectrum disorder and non-ASD adolescents. *Adolesc Health Med Ther*. (2022) 13:15–21. doi: 10.2147/AHMT.S344591

5. Angulo-Jiménez H, DeThorne L. Narratives about autism: an analysis of YouTube videos by individuals who self-identify as autistic. *Am J Speech Lang Pathol*. (2019) 28:569–90. doi: 10.1044/2018_AJSLP-18-0045

6. Pew Research Center. (2021). Available online at: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

7. Liu J-B, Guan L, Cao J, Chen L. Coherence analysis for a class of polygon networks with the noise disturbance. *IEEE Trans Syst Man Cybern Syst*. (2025) 2025:326. doi: 10.1109/TSMC.2025.3559326

8. Al-Nefaie AH, Aldhyani TH, Sultan HA, Alzahrani Eidah M. Application of artificial intelligence in modern healthcare for diagnosis of autism spectrum disorder. *Front Med*. (2025) 12:1569464. doi: 10.3389/fmed.2025.1569464

9. Kim B, Jeong D, Kim JG, Hong H, Han K. *V-DAT (virtual reality data analysis tool): supporting self-awareness for autistic people from multimodal VR sensor data*. In: Proceedings of the UIST 2023—Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, San Francisco, CA, USA, 29 October–1 November 2023. (2023).

10. Chen C, Chander A, Uchino K. *Guided play: digital sensing and coaching for stereotypical play behavior in children with autism*. In Proceedings of the International Conference on Intelligent User Interfaces, Proceedings IUI, Marina del Ray, CA, USA, 17–20 March 2019; Part F1476. pp. 208–217. (2019).

11. Parui S, Samanta D, Chakravorty N, Ghosh U, Rodrigues JJ. Artificial intelligence and sensor-based autism spectrum disorder diagnosis using brain connectivity analysis. *Comput Electr Eng*. (2023) 108:108720. doi: 10.1016/j.compeleceng.2023.108720

12. Golestan S, Soleiman P, Moradi H. A comprehensive review of technologies used for screening, assessment, and rehabilitation of autism Spectrum disorder. *arXiv*. (2018) 2018:10986. doi: 10.48550/arXiv.1807.10986

13. Neeharika CH, Riyazuddin YM. Developing an artificial intelligence based model for autism Spectrum disorder detection in children. *J Adv Res Appl Sci Eng Technol*. (2023) 32:57–72. doi: 10.37934/araset.32.1.5772

14. Wall DP, Dally R, Luyster R, Jung J-Y, DeLuca TF. Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PLoS One*. (2012) 7:e43855. doi: 10.1371/journal.pone.0043855

15. Alzakari SA, Allinjawi A, Aldrees A, Zamzami N, Umer M, Innab N, et al. Early detection of autism spectrum disorder using explainable AI and optimized teaching strategies. *J Neurosci Methods*. (2025) 413:110315. doi: 10.1016/j.jneumeth.2024.110315

16. Heunis T, Aldrich C, Peters J, Jeste S, Sahin M, Scheffer C, et al. Recurrence quantification analysis of resting state EEG signals in autism spectrum disorder—a systematic methodological exploration of technical and demographic confounders in the search for biomarkers. *BMC Med*. (2018) 16:101. doi: 10.1186/s12916-018-1086-7

17. Vicnesh J, Wei JKE, Oh SL, Arunkumar N, Abdulhay E, Ciaccio EJ, et al. Autism spectrum disorder diagnostic system using HOS bispectrum with EEG signals. *Int J Environ Res Public Health*. (2020) 17:971. doi: 10.3390/ijerph17030971

18. Novielli P, Romano D, Magarelli M, Diacono D, Monaco A, Amoroso N, et al. Personalized identification of autism-related bacteria in the gut microbiome using explainable artificial intelligence. *iScience*. (2024) 27:110709. doi: 10.1016/j.isci.2024.110709

19. 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

20. Beykikhoshk A, Arandjelović O, Phung D, Venkatesh S, Caelli T. Using twitter to learn about the autism community. *Soc Netw Anal Min*. (2015) 5:261. doi: 10.1007/s13278-015-0261-

21. Mazurek MO. Social media use among adults with autism spectrum disorders. *Comput Hum Behav*. (2013) 29:1709–14. doi: 10.1016/j.chb.2013.02.004

22. Onnela J, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology*. (2016) 41:1691–6. doi: 10.1038/npp.2016.7.npp20167

23. Tomeny TS, Vargo CJ, El-Toukhy S. Geographic and demographic correlates of autism-related anti-vaccine beliefs on twitter, 2009–2015. *Soc Sci Med*. (2017) 191:168–75. doi: 10.1016/j.socscimed.2017.08.041

24. Tárraga-Mínguez R, Gómez-Marí I, Sanz-Cervera P. What motivates internet users to search for Asperger syndrome and autism on Google? *Int J Environ Res Public Health*. (2020) 17:9386. doi: 10.3390/ijerph17249386

25. Hartwell M, Keener A, Coffey S, Chesher T, Torgerson T, Vassar M. Brief report: public awareness of Asperger syndrome following Greta Thunberg appearances. *J Autism Dev Disord*. (2020) 51:2104–8. doi: 10.1007/s10803-020-04651-9

26. Rubio-Martín S, García-Ordás MT, Bayón-Gutiérrez M, Prieto-Fernández N, Benítez-Andrades JA. "*Early detection of autism Spectrum disorder through AI-powered analysis of social media texts*," 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS), L'Aquila, Italy, pp. 235–240. (2023).

27. Jaiswal A, Washington P. Using #ActuallyAutistic on twitter for precision diagnosis of autism Spectrum disorder: machine learning study. *JMIR Form Res*. (2024) 8:e52660. doi: 10.2196/52660

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 29 August 2025
DOI 10.3389/fonc.2025.1643208

Check for updates

OPEN ACCESS

EDITED BY
Pardeep Sangwan,
Maharaja Surajmal Institute of Technology, India

REVIEWED BY
Burcu Selçuk,
Yeditepe University, Türkiye
Abdul Rahman,
VIT Bhopal University, India

*CORRESPONDENCE
Weijuan Han
hanweijuan@zykj.edu.cn

RECEIVED 10 June 2025
ACCEPTED 11 August 2025
PUBLISHED 29 August 2025

CITATION
Han W, Dong X, Wang G, Ding Y and Yang A (2025) Application and improvement of YOLO11 for brain tumor detection in medical images.
Front. Oncol. 15:1643208.
doi: 10.3389/fonc.2025.1643208

COPYRIGHT
© 2025 Han, Dong, Wang, Ding and Yang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Application and improvement of YOLO11 for brain tumor detection in medical images

Weijuan Han[1]*, Xinjie Dong[2], Guixia Wang[1], Yuwen Ding[3] and Aolin Yang[4]

[1]School of Mechanical and Electronic Engineering, Zhongyuan Institute of Science and Technology, Zhengzhou, China, [2]Information and Communication Department, Henan Public Security Department, Zhengzhou, China, [3]Editorial Department, Henan Medical College, Zhengzhou, China, [4]School of Public Administration, Henan University of Economics and Law, Zhengzhou, China

Brain tumors pose a critical threat to human health, and early detection is essential for improving patient outcomes. This study presents two key enhancements to the YOLOv11 architecture aimed at improving brain tumor detection from MRI images. First, we integrated a set of novel attention modules (Shuffle3D and Dual-channel attention) into the network to enhance its feature extraction capability. Second, we modified the loss function by combining the Complete Intersection over Union (CIoU) with a Hook function (HKCIoU). Experiments conducted on a public Kaggle dataset demonstrated that our improved model reduced parameters and computations by 2.7% and 7.8%, respectively, while achieving mAP50 and mAP50–95 improvements of 1.0% and 1.4%, respectively, over the baseline. Comparative analysis with existing models validated the robustness and accuracy of our approach.

KEYWORDS

brain tumor, object detection, you only look once (YOLO), attention, intersection over union (IoU), mean average precision (MAP), giga floating point operations per second (GFLOPs)

1 Introduction

Brain tumors present a serious risk to human health with potentially devastating consequences. Abnormal growth can interfere with brain function, causing severe neurological symptoms, cognitive impairment, and in many cases, mortality (1, 2). The classification of brain tumors serves as the foundation for clinical diagnosis, treatment planning, and prognostic assessment. The most authoritative international system is the World Health Organization (WHO) Classification of Tumors of the Central Nervous System, with the latest 5th edition (WHO CNS5) published in 2021 (3, 4). This classification integrates histopathology, molecular genetics, and clinical phenotypes to form an integrated diagnosis framework, replacing the previous morphology-based classification model. Based on tissue origin and biological characteristics, WHO CNS5 categorizes brain tumors into the following 6 categories: Neuroepithelial Tumors (Gliomas

and Related Tumors), Meningeal Tumors, Cranial and Peripheral Nerve Tumors, Germ Cell Tumors, Sellar Region Tumors, and Metastatic Brain Tumors.

Tumor characteristics such as location, size, and grade are critical determinants of neurological impairments and functional deficits in patients with brain tumors. Location directly influences the specific deficits due to the brain's functional specialization. For example, tumors in the motor cortex often cause contralateral limb weakness or paralysis, while lesions in the cerebellum may lead to ataxia and coordination difficulties. Size correlates with the severity of mass effect and peritumoral edema. Larger tumors (e.g., diameters >4 cm) exert greater mechanical pressure on surrounding tissues, causing midline shift, ventricular compression, and increased intracranial pressure, which manifest as headaches, nausea, altered consciousness, and even herniation. Grade reflects tumor aggressiveness and biological behavior. Low-grade tumors grow slowly and may remain asymptomatic for years, while high-grade tumors exhibit rapid infiltration, angiogenesis, and necrosis, leading to severe and progressive deficits. In summary, tumor location dictates the type of neurological deficits, size determines the extent of mass effect and increased intracranial pressure-related complications, and grade predicts the tempo and severity of clinical progression. Multidisciplinary management (5), including surgical planning, adjuvant therapies, and neurorehabilitation, must account for these interdependent factors to optimize outcomes.

Early and accurate detection of brain tumors is essential for treatment planning, as timely intervention can significantly improve patient prognosis and quality of life. Magnetic resonance imaging (MRI) (6) has become a primary diagnostic tool for brain tumors owing to its high soft-tissue contrast and detailed anatomical resolution. However, manual analysis of MRI scans for tumor detection is time-consuming and prone to error, relying heavily on medical expertise. Therefore, developing automated and reliable object-detection algorithms for brain tumors in MRI images has become a critical research priority.

Traditional machine learning algorithms to detect brain tumors in medical images, such as Haar cascades (7) and histograms of oriented gradients (HOG) (8) combined with support vector machines (SVM), have been applied to brain tumor detection. These methods depend on handcrafted features that require extensive domain knowledge and careful design. However, these methods often fail to generalize across datasets and imaging modalities, as performance is constrained by the complexity and variability of brain tumor appearance on MRI scans. The inability to extract high-level semantic information limits the accuracy and robustness of traditional machine-learning-based detection methods.

Deep learning has introduced transformative advances in object detection. Region-based convolutional neural networks (R-CNNs) (9), introduced by Girshick et al., marked a significant milestone by applying a data-driven approach to object detection. Faster R-CNN (10), an improved version of R-CNN, integrated a region proposal network (RPN), which reduces computational cost and increases detection speed while preserving accuracy. For brain tumor detection, Faster R-CNNs have demonstrated promise in accurately identifying tumor regions by leveraging deep convolutional features (11). However, its slow processing and complex two-stage architecture limit practical use in real-time medical diagnostics.

The single-shot multibox detector (SSD) (12) developed by Liu et al. has proven to be an efficient alternative to two-stage detectors. The model predicts the bounding boxes and class probabilities within a single network, enabling faster inference. By utilizing feature maps from different layers, an SSD can effectively detect objects of various scales, achieving a good balance between speed and accuracy. In brain tumor detection using MRI images, SSD has demonstrated the ability to detect tumors of different sizes; however, it still faces challenges in accurately detecting small and irregularly shaped tumors because of the limited receptive field of shallow layers and loss of spatial information in deeper layers.

The You Only Look Once (YOLO) series (13), introduced by Redmon et al., has attracted wide attention for its significant advantages in object detection. Firstly, the single-stage architecture of YOLO endows it with high computational efficiency, and is capable of real-time or near-real-time detection. This is highly valuable in clinical settings, where rapid results help doctors make timely diagnostic decisions. For example, in the context of brain tumor detection from MRI images (14), doctors can promptly access results, and quickly specify examinations or treatment. Secondly, YOLO captures global contextual information from the entire input image. In contrast to other methods that focus on local regions separately, the holistic approach of YOLO helps better understand the relationships between different parts of an image. Simultaneously, YOLO can accurately identify the location and category of tumors, even when they have complex shapes. This holistic understanding is particularly valuable for addressing the complexity of brain tumors in MRI scans. Moreover, the YOLO series has demonstrated strong generalization across different datasets and scenarios such as COCO (15), PASCAL VOC2012, NEU-DET, RSOD (16), LOCO dataset (17), Figshare dataset (18), and so on. With continuous improvements in its architecture and training strategies over successive versions (19–22), it can adapt well to the variations in image quality, tumor appearance, and imaging parameters commonly encountered in real-world medical imaging applications. This adaptability renders YOLO a reliable tool for detecting brain tumors in varied MRI datasets.

The original YOLO can achieve real-time performance on standard graphics processing units (GPUs), rendering it suitable for applications requiring rapid detection. Subsequent versions of YOLO, such as YOLOv5, YOLOv8, and beyond, have continuously improved the architecture and introduced advanced techniques, further enhancing detection performance.

This study focuses on YOLOv11 (23), an iteration of the YOLO series released in 2024. Building on the achievements of its predecessors, YOLOv11 integrates advanced architectures and optimization strategies to overcome limitations in handling the complex and diverse characteristics of brain tumors in MRI images. Given the increasing demand for efficient and accurate brain tumor detection in clinical practice, YOLOv11 holds considerable

potential for achieving superior performance in terms of detection speed, accuracy, and the ability to identify tumors of various shapes and sizes. This study aimed to explore the capabilities of YOLOv11 in brain tumor detection from MRI images and conduct comprehensive experiments to evaluate its effectiveness using a publicly available Kaggle dataset.

The structure of this paper is organized as follows: Section 2 describes the related work. Section 3 provides a detailed description of our methodology and improvement measures. Section 4 presents the experimental results, a comprehensive performance analysis and comparison with other models. Section 5 provides an overall discussion. Finally, Section 6 concludes the paper.

2 Related work

In recent years, numerous studies have been conducted on the detection of brain tumors in MRI images using deep learning algorithms, particularly the YOLO series algorithms, which have demonstrated excellent performance.

Kharb et al. (24) proposed a hybrid model for brain tumor classification that combined faster R-CNN and EfficientNet. The hybrid model achieved a notable accuracy of 98.96% during the training phase and 99.2% during the testing phase on the Figshare (25) Datasets.

Hikmah et al. (26) introduced a novel approach for precise brain tumor detection, combining various approaches such as morphological operations for tumor segmentation, image enhancement, and a deep learning architecture based on MobileNetV2-SSD with feature pyramid network (FPN), where the FPN level originally set to 3 had been modified to level 2, which enhanced the detection of smaller objects. The proposed model obtained a recall value of around 98% and a precision value of around 89%.

Alsufyani (27) explored the use of several deep-learning models, including YOLOv8, YOLOv9, Faster R-CNN, and ResNet18, for the detection of brain tumors from MRI images. The results on the Kaggle's Medical Image Dataset for Brain Tumor Detection, consisting of 3903 brain MRI images, demonstrate that YOLOv9 outperforms the other models in terms of mAP (0.826) and accuracy (0.784), highlighting its potential as the most effective deep-learning approach for brain tumor detection.

Chen et al. (28) proposed the YOLO-NeuroBoost model, combining the improved YOLOv8 algorithm with innovative techniques, such as the dynamic convolution kernel warehouse, attention mechanism CBAM, and inner-GIoU loss function. It achieved mean average precision (mAP) scores of 99.48% and 97.71% on the BR35H (29) and RoboFlow (30) datasets. High mAP scores indicate the high accuracy and efficiency of the model in detecting brain tumors in MRI images. However, the model has more parameters and GFLOPs than YOLOv11, resulting in a larger model size.

Kang et al. (31) proposed PK-YOLO, which included the following three components: a pretrained, pure lightweight CNN-based backbone via sparse masked modeling, a YOLO architecture with a pretrained backbone, and a regression loss function for improving small object detection. PK-YOLO achieved a mAP of 58.2% on the BR35H dataset.

Monisha and Rahman et al. (32) proposed a federated learning architecture to enhance brain tumor detection by incorporating the YOLOv11 algorithm. The federated learning approach safeguards patient data while enabling collaborative deep-learning model training across multiple institutions. On a synthetic brain tumor dataset with about 10,000 MRI images, the model achieved a mean average precision (mAP) of 90.8% and an mAP50–95 of 65.3%.

Dulal et al. (33) proposed an enhanced version of YOLOv8. Their work significantly advances automated brain tumor detection by introducing an improved YOLOv8 model. Through strategic modifications, including the integration of a Vision Transformer block, Ghost Convolution, and RT-DETR, their model achieved 91% mAP0.5 on a public Kaggle dataset.

Wahidin et al. (34) used several of the latest versions of the YOLO model, namely YOLOv11m, YOLOv10m, YOLOv9m, and YOLOv8m, to detect brain tumors such as gliomas, meningiomas, and pituitary tumors in MRI images. Hyperparameter tuning was conducted using the Bayesian optimization and HyperBand (BOHB) search algorithm with ray tuning through 16 trials. YOLOv11m achieved the highest accuracy, with an mAP50 of 0.934 and an inference speed of 70.550 FPS. In contrast, YOLOv8m delivered the fastest inference speed of 80.471 FPS.

Bai et al. (35) proposed the SCC-YOLO architecture, integrating the SCConv module into YOLOv9. The SCConv module improves convolutional efficiency by reducing spatial and channel redundancy and enhancing image feature learning. This study examined the effects of different attention mechanisms with YOLOv9 on brain tumor detection using Br35H and custom datasets. The results indicate that SCC-YOLO improves mAP50 by 0.3 to 95.7% on the BR35H dataset and by 0.5 to 86% compared with YOLOv9. SCC-YOLO demonstrated strong performance in brain tumor detection.

This study involved two primary improvements. First, the YOLOv11 network architecture was enhanced by integrating several newly designed attention modules to strengthen the feature extraction capabilities of the network. Second, the loss function was modified to increase the loss value of low-quality prediction boxes, and promote rapid convergence of the model.

3 Materials and methods

The YOLO series of algorithms has demonstrated strong performance in detecting brain tumors in MRI images, particularly in terms of accuracy and efficiency. However, the algorithms may have different performances in different datasets and application scenarios, and further research and improvements are needed to improve the accuracy and efficiency of brain tumor detection and to serve clinical diagnosis better.

3.1 YOLOv11

The YOLOv11 structure (Figure 1) comprises three main components: the backbone, neck, and head (36, 37). The backbone contains 0–10 convolution modules, the neck layer comprises 11–22 parts, and the rest are three parallel detection heads that detect feature maps of 20 × 20, 40 × 40, and 80 × 80, and generate 8,400 possible detection results.

As the core of feature extraction, the backbone of YOLOv11 replaces YOLOv8's C2f module with an improved C3K2 module and standard convolution (CBS). C3K2 module uses multi-scale convolution kernel C3K, where K is an adjustable convolution kernel size, such as 3 × 3, 5 × 5, etc. This design can expand the receptive field, allowing the model to capture a wider range of contextual information, especially suitable for large object detection or scenes with complex backgrounds. The CBS module mainly consists of three parts: Conv (convolution layer), BN (Batch Normalization) and SiLU (activation function). It also adds a C2PSA (Cross-Level Pyramid Slice Attention) module after SPPF, enhancing global feature modeling capabilities through a multi-head attention mechanism. This design enables the network to more effectively capture long-range dependencies, which is particularly important for occluded objects and complex scenes. The Feature Pyramid Network (FPN) structure is retained at the neck layer. The neck layer also uses C3K2 and CBS convolutions for extraction, with feature fusion performed using the Concat operation. The head layer, like previous versions, also includes three detection heads. Each head employs depthwise separable convolution (DWC) and standard convolution (CBS).

YOLOv11's loss function continues the YOLO series' pursuit of a balance between detection accuracy and speed. Targeted at the decoupled head structure, the loss function is divided into three parts: bounding box regression loss, confidence loss and classification loss, Bounding box regression loss enables the model to accurately locate the target, confidence loss can optimize the accuracy of the prediction box and improve the model's ability to judge whether the target exists in the prediction box, and classification loss determines the category of the image in the prediction box. Bounding box regression includes the CIoU (Complete Intersection over Union) (38) loss and the DFL (Distribution Focal Loss) (39), which take into account the overlap, position, and shape of the bounding boxes. The total loss

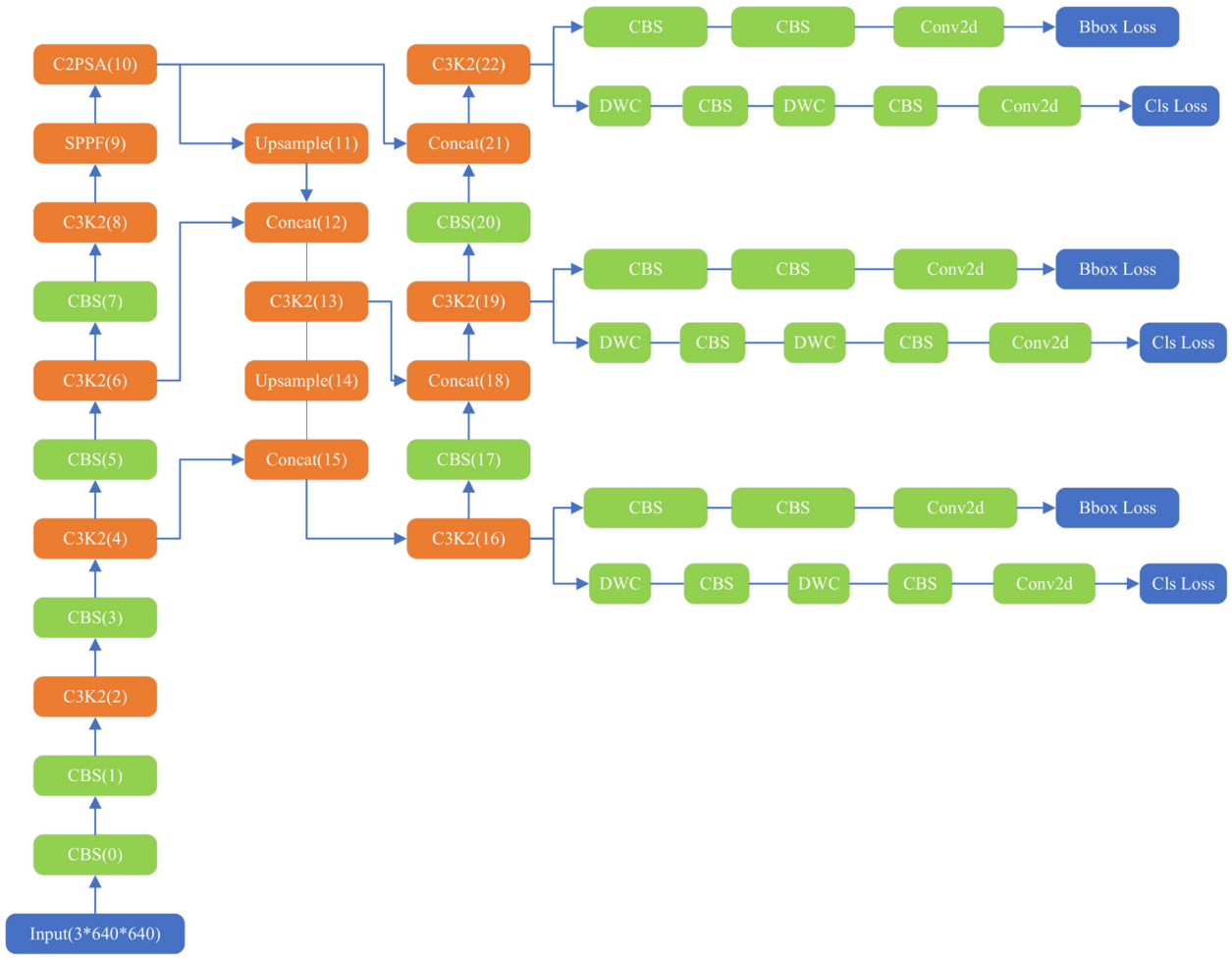


FIGURE 1
Structure of original YOLOv11.

is a weighted sum of these three losses. The loss function calculation formula is shown in Equations 1, 2. In the equations, L_{box} represents bounding box regression loss, L_{obj} represents confidence loss, L_{cls} represents classification loss, L_{CIoU} represents CIoU loss, L_{DFL} represents DFL loss and α, β, and γ represent weight parameters.

$$L_{total} = \alpha L_{box} + \beta L_{obj} + \gamma L_{cls} \quad (1)$$

$$L_{box} = L_{CIoU} + L_{DFL} \quad (2)$$

3.2 Main methods

Due to hardware limitations in clinical application environments and the demand for faster speeds, we are committed to reducing the number of model parameters and computational complexity, and improving detection accuracy. we integrated a set of novel attention modules into the network. This study replaces the original self-attention module C2PSA with a newly designed spatial attention module. At the same time, this study uses an improved loss function instead of the original loss function CIoU.

3.3 Attention

This study employed three attention mechanisms: Spatial attention, Shuffle3D attention, and Dual-channel attention. The latter two are newly designed attention mechanisms.

3.3.1 Shuffle3D attention

This study draws on the concepts of the Shuffle (40) and SimAM (41) attention mechanisms to propose a novel attention mechanism, designated as Shuffle3D (Figure 2). On the one hand, channel rearrangement is applied to disrupt the original channel order, introducing random diversity and enabling joint modeling of different features. This module increases information exchange and balance between channels. On the other hand, a spatial inhibition mechanism is used. In neuroscience, information-rich neurons often exhibit different discharge patterns from the surrounding neurons. Moreover, activated neurons commonly inhibit neighboring neurons. Thus, neurons exhibiting spatial inhibition should receive greater emphasis. The calculation formulae of inhibition effects are presented in Equations 3–5, where x represents the input feature map, x_{ij} represents a point in the feature map, e represents the mean, H represents the height of the feature map, W represents the width of the feature map, u represents the degree of deviation from the mean at a certain point on the feature map, and α and β are the regulators, which are set to the -4th power of 10 and 0.5, respectively. Neurons that deviate more from the mean yield higher activation function values.

$$e = \frac{1}{H*W-1}\sum_{j=1}^{H}\sum_{i=1}^{W}x_{ij} \quad (3)$$

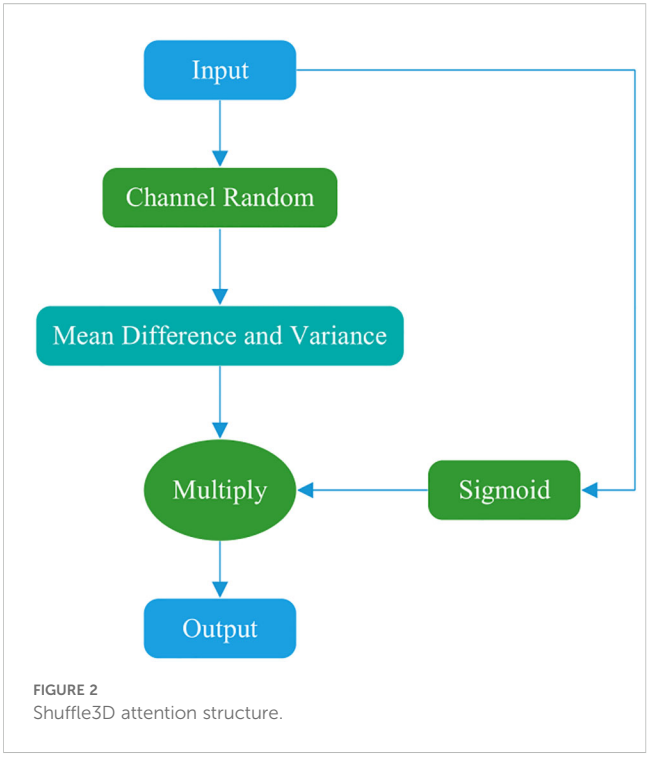


FIGURE 2
Shuffle3D attention structure.

$$u = \frac{(x-e)^2}{4(\sum_{j=1}^{H}\sum_{i=1}^{W}x_{ij}/(H*W-1)+\alpha)} + \beta \quad (4)$$

$$x = sigmoid(u)*x \quad (5)$$

3.3.2 Spatial attention

The main goal of the Spatial attention module (Figure 3) is to explicitly model the dependencies between spatial locations and generate a spatial attention map. First, the input features are max-pooled and average-pooled in the channel dimension to generate two spatial descriptors. These two spatial descriptors are then concatenated in the channel dimension and passed through a convolutional layer to generate a spatial attention map. Finally, the values of the spatial attention map are normalized to the range (0, 1) using a sigmoid function and multiplied by the input tensor to generate the output.

The Conv2d module in the figure uses a kernel of (7,7), a stride of 1, padding of 3, 2 input channels, and 1 output channel (number of filters). These parameters ensure that the spatial dimensions (w, h) of the input and output feature maps are consistent and combine the results of average pooling and max pooling.

3.3.3 Dual-channel attention

Figure 4 illustrates the Dual-channel attention, which comprises two main components. The Dual-channel attention borrows the idea of parallel convolution of different sizes of kernels from Inception (42). The first part uses two parallel convolution operations with different convolution kernel sizes to capture

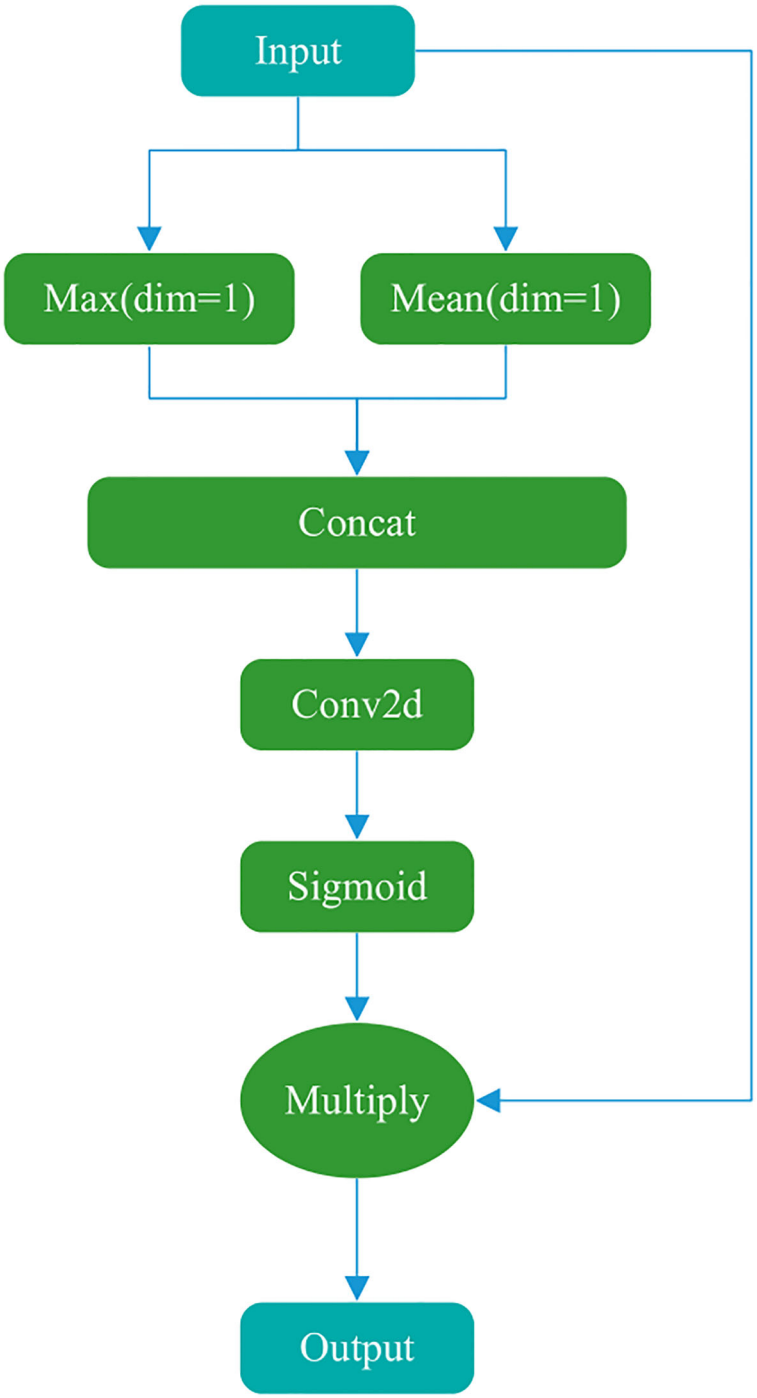


FIGURE 3
Spatial attention structure.

additional feature information. The second part involves concatenation, convolution, and spatial attention computation. The final result is multiplied by the input to produce the output.

3.3.4 New structure of the YOLOv11 networks

To enhance feature extraction in convolutional neural networks, we integrated the newly designed Shuffle3D with Spatial and Dual-channel attention. The positions of the attention modules are shown in Figure 5. The blue areas represent attention modules that are newly added or that replace the original ones. Dual-channel replaces the original self-attention module C2PSA, greatly reducing the computational load. Shuffle3D replaces the first CBS and DWC convolution modules on each detection head, enhancing the ability of the model to extract features from key regions.

10.3389/fonc.2025.1643208

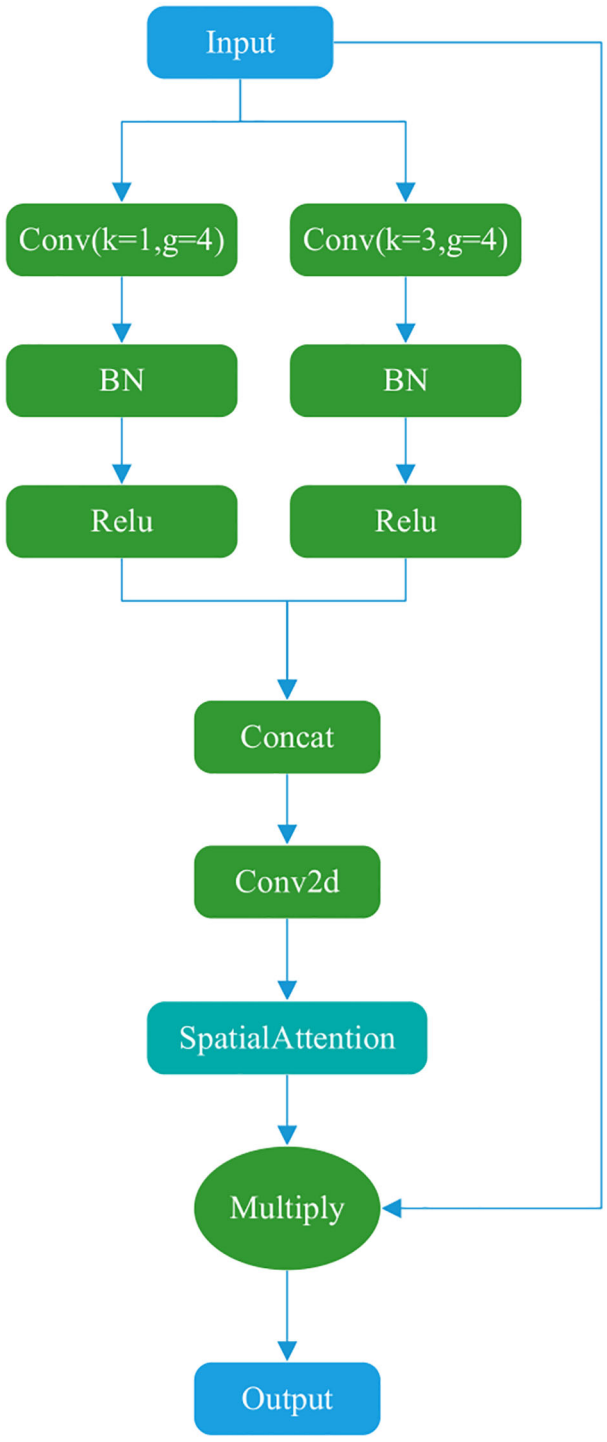


FIGURE 4
Dual-channel attention structure.

3.4 HKCIoU

In the original YOLOv11, complete intersection over union (CIoU) serves as the boundary regression loss function, as shown in Equations 6–8. The *CIoU* loss refers to the loss during training and validation. The *IoU* stands for Intersection over Union. The ρ represents the distance between the center points of the predicted box and the true box, and c represents the diagonal distance of the minimum closure area that can contain both the predicted and true boxes. b^p and b^t represent the center points of the predicted box and the true box respectively. w^t represents the width of the true box, and w^p represents the width of the predicted box. h^t represents the height of the true box, and h^p represents the height of the predicted box. CIoU adds the penalty term of α and β, which are parameters used to measure the consistency of the aspect ratio.

$$CIoU = IoU - \frac{\rho^2(b^p, b^t)}{c^2} - \alpha * \beta \tag{6}$$

$$\beta = \frac{4}{\pi^2}(arctan\frac{w^t}{h^t} - arctan\frac{w^p}{h^p})^2 \tag{7}$$

$$\alpha = \frac{\beta}{1 - IoU + \beta} \tag{8}$$

The hook function opens upward in the first quadrant (Figure 6). It is used to adjust the CIoU value, forming the HKCIoU. For a smaller CIoU, the loss is relatively amplified, and for a larger CIoU, the loss is relatively reduced, thereby accelerating the network convergence and enabling the network parameters to reach the optimal value faster. The calculation is given in Equations 9, 10. x represents the loss of CIoU, a and b are hyperparameters. a and b are both set to 0.5 where the value of equation has reached the minimum when x equals 1.

$$f(x) = ax + \frac{b}{x} \ (ab > 0) \tag{9}$$

$$HKCIoU = (a * CIoU + \frac{b}{CIoU}) * CIoU \tag{10}$$

4 Results

The experimental hardware setup includes a 13th Gen Intel(R) Core(TM) i5-13600KF, 3500 MHz, 14 cores, 32 GB of RAM, and an RTX 4060Ti GPU with 16 GB of VRAM. The software environment included Windows 11, Python 3.8, Torch 1.13.1, CUDA 11.7, and PyCharm 2021.3. Each model was trained for 100 epochs, with a batch size of 32. The model employed SGD as the optimizer, with an initial learning rate of 0.01, a momentum of 0.937, and a weight decay of 0.0005.

YOLOv11 extensively utilizes various data augmentation techniques in training, including but not limited to HSV adjustment (hue, saturation, brightness transformation), random flipping/rotation, scaling, geometric affine transformation, random erasure, and Mosaic enhancement, significantly improving the model's adaptability to scale changes, occluded scenes, and small targets. YOLOv11 closes Mosaic at the end of training and switches to standard image training in the last 10 epochs to avoid overfitting caused by differences in distribution between synthesized images and real data.

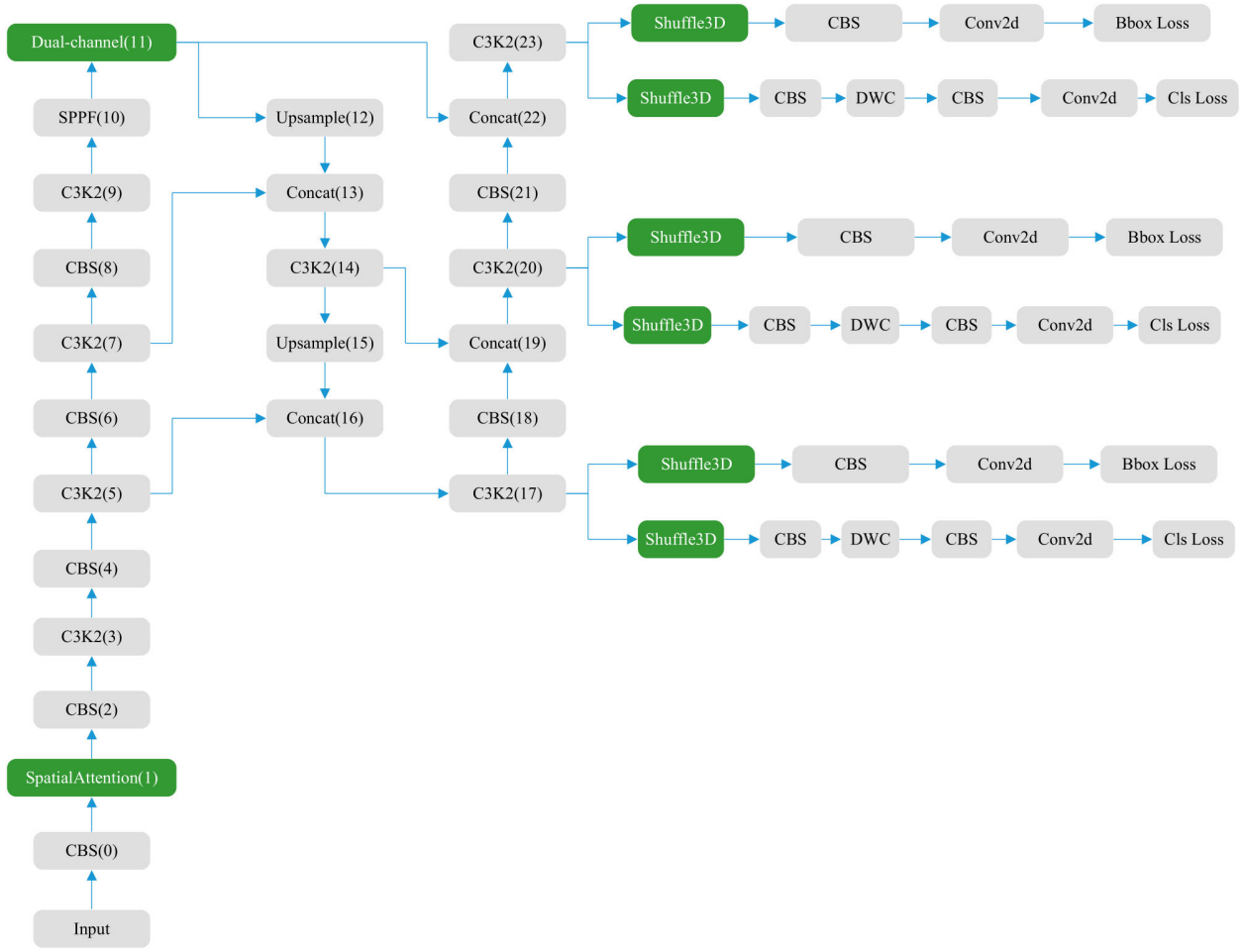


FIGURE 5
New structure of YOLOv11 with attention.

A Brain Tumor Detection Dataset (43) from Kaggle was used as experimental data. The dataset contains 5,249 MRI images divided into training and validation sets. The training set consists of 4,737 images, including 1,153 Glioma, 1,449 Meningioma, 711 No Tumor, and 1,424 Pituitary images. The validation set consists of 512 images, including 136 Glioma, 140 Meningioma, 100 No Tumor, and 136 Pituitary images. Each image was annotated with YOLO-format bounding boxes and labeled with one of four brain

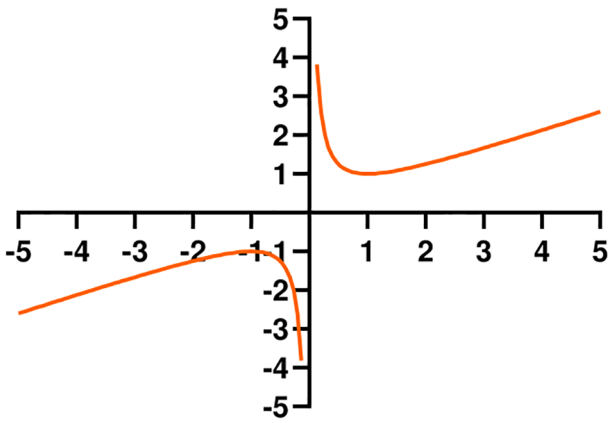


FIGURE 6
Hook function.

tumor classes. The evaluation indicators of the model include parameter count, computational complexity, mAP50, mAP50-95, and FPS (Frames Per Second).

4.1 Attention ablation experiment

In the experiment, we used three attention mechanisms, and the ablation results of the three attention mechanisms are shown in Table 1. From the table, it can be seen that the use of attention mechanism resulted in varying degrees of increase in mAP indicators. Compared to the model numbered 8, the models numbered 2, 3, and 5 achieved higher performance, but their parameter and computational complexity increased significantly. Although the parameter quantity and computational complexity of models numbered 4, 6, and 7 are lower than model 8, their mAP indicators are not as good as model 8. Their results are very close, and there is some fluctuation in the results of different experiments in the same model. Taking all factors into consideration, we have chosen to use the model 8 with three types of attention, namely Spatial, Dual-channel, Shuffle attention.

4.2 Ablation experiment

Ablation experiments (Table 2, Figure 7) demonstrated that when only the hook function was used, both mAP50 and mAP50–95 were improved by 0.8%. When only the attention mechanism was used, mAP50 and mAP50–95 were improved by 0.7% and 0.5%, respectively. The model using both Hook and Attention, named YOLOv11n-HA, improved mAP50 and mAP50–95 by 1% and 1.4%, respectively, with a 2.7% reduction in parameters and a 7.8% reduction in calculations. Simultaneously, in terms of FPS, YOLOv11n-HA achieved a 1.5% rise compared to the baseline model. The PR curve of YOLOv11n-HA on the test set is shown in Figure 8, which includes the mAP50 values of each subclass.

To demonstrate the robustness of the model, we conducted three experiments on the final model, YOLOv11n-HA, which includes two improvements. The results are shown in Table 3. From the table, it can be seen that there is some fluctuation in the results of the model. This study speculates that this phenomenon is not only related to the jitter of the neural network but also to the random channel rearrangement of Shuffle3D attention, which increases the randomness of the model. Based on the mAP50 metric, we selected the experiment with the median value as the result. That is the one with an mAP50 value of 96.8%.

4.3 Comparison

Table 4 presents results comparing YOLOv11n-HA with other models, including non-YOLO and YOLO series deep learning models. The models and data involved were retrained and validated using the same dataset for this study.

4.3.1 Comparison with non-YOLO series

Faster-RCNN and SSD not only have lower mAP50 and mAP50–95 indicators than YOLOv11n-HA but also have several

TABLE 1 Attention ablation experiment based on YOLOv11n.

Number	Attention	Parameters (million)	GFLOPs	mAP50 (%)	mAP50-95 (%)
1	YOLOv11n	2.59	6.4	95.8	78.1
2	+Spatial	2.59	6.5	96.5	79.7
3	+Dual-channel	2.64	6.5	96.3	78.7
4	+Shuffle3D	2.47	5.8	96.4	78.4
5	+Spatial +Dual-channel	2.64	6.5	96.7	79.5
6	+Spatial +Shuffle3D	2.47	5.8	96.3	78.6
7	+Dual-channel +Shuffle3D	2.52	5.8	96.5	78.5
8	+ALL	2.52	5.9	96.5	78.6

TABLE 2 Improved ablation experiment based on YOLOv11n.

Model	Parameters (million)	GFLOPs	mAP50 (%)	mAP50-95 (%)	FPS (f/s)
YOLOv11n	2.59	6.4	95.8	78.1	66.91
+Hook	2.59	6.4	96.6	78.9	65.96
+Attention	2.52	5.9	96.5	78.6	67.85
+Hook +Attention (YOLOv11n-HA)	2.52	5.9	96.8	79.5	67.90

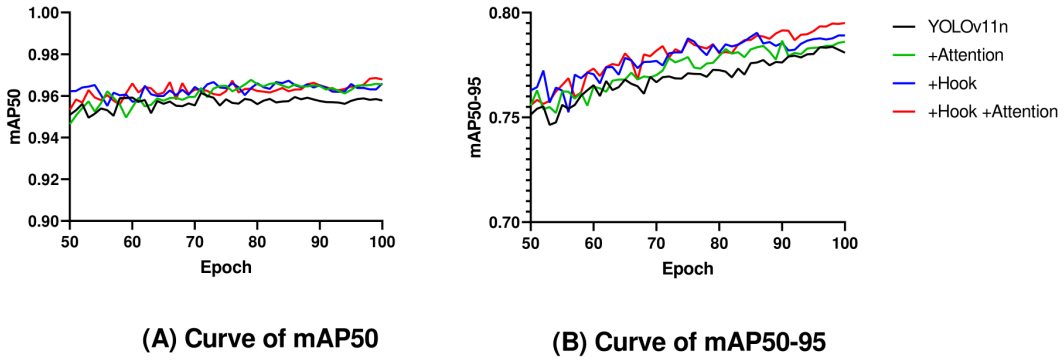


FIGURE 7
Curves of mAP50 and mAP50–95 with epoch in ablation experiments.

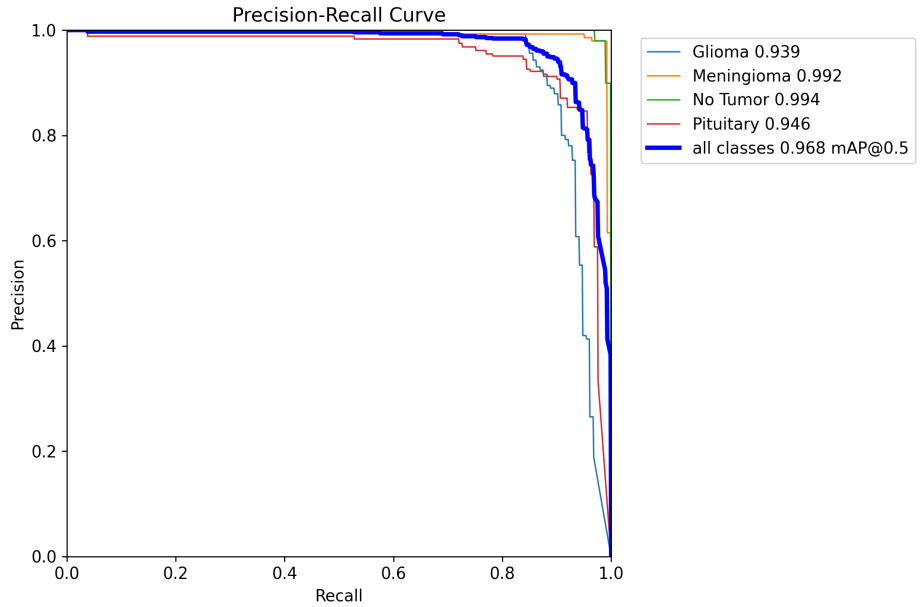


FIGURE 8
PR curve of YOLOv11n-HA on test set.

times more parameters and computational complexity. Compared with RT-DETR(L), YOLOv11n-HA uses only 7.7% of the parameters and 3.3% of the computational complexity, while achieving increases of 3.9% in mAP50 and 7.9% in mAP50-95.

TABLE 3 Results of three experiment based on YOLOv11n-HA.

Number	mAP50 (%)	mAP50-95 (%)
1	96.7	79.1
2	96.8	79.5
3	96.9	79

4.3.2 Comparison with YOLO series

Comparing the metrics of YOLOv11n-HA with that of YOLOv5n, we observe that the GFLOPs of YOLOv11n-HA remain the same, the number of parameters increases by 15.6% from 2.18M to 2.52M, and the mAP50 and mAP50–95 indicators increase by 0.5% and 1.4%, respectively. Compared with that of YOLOv8n, the number of parameters in YOLOv11n-HA decreased by 6.3%, computational GFLOPs decreased by 14.5%, and the mAP50 and mAP50–95 indicators increased by 0.6% and 0.5%, respectively. Compared to that of YOLOv9s, the number of parameters of YOLOv11n-HA decreased by 60.1%, the number of calculations decreased by 74%, mAP50 increased by 0.4%, and mAP50–95 decreased by 0.2%. Under the condition of a significant decrease in the number of parameters and the cost of calculations,

TABLE 4 Comparison results with other state-of-the-art models used in the detection of brain tumors.

Model	Parameters (million)	GFLOPs	mAP50 (%)	mAP50-95 (%)
Faster-RCNN (ResNet50)	28.30	470.48	91.2	59
SSD (VGG)	24.01	61.06	93.7	70.7
YOLOv5n	2.18	5.9	96.3	78.1
YOLOv8n	2.69	6.9	96.2	79
YOLOv9s	6.32	22.7	96.4	79.7
YOLOv10n	2.71	8.4	95.4	78.4
RT-DETR (L)	32.8	108.0	92.9	71.6
YOLOv11n-HA	2.52	5.9	96.8	79.5

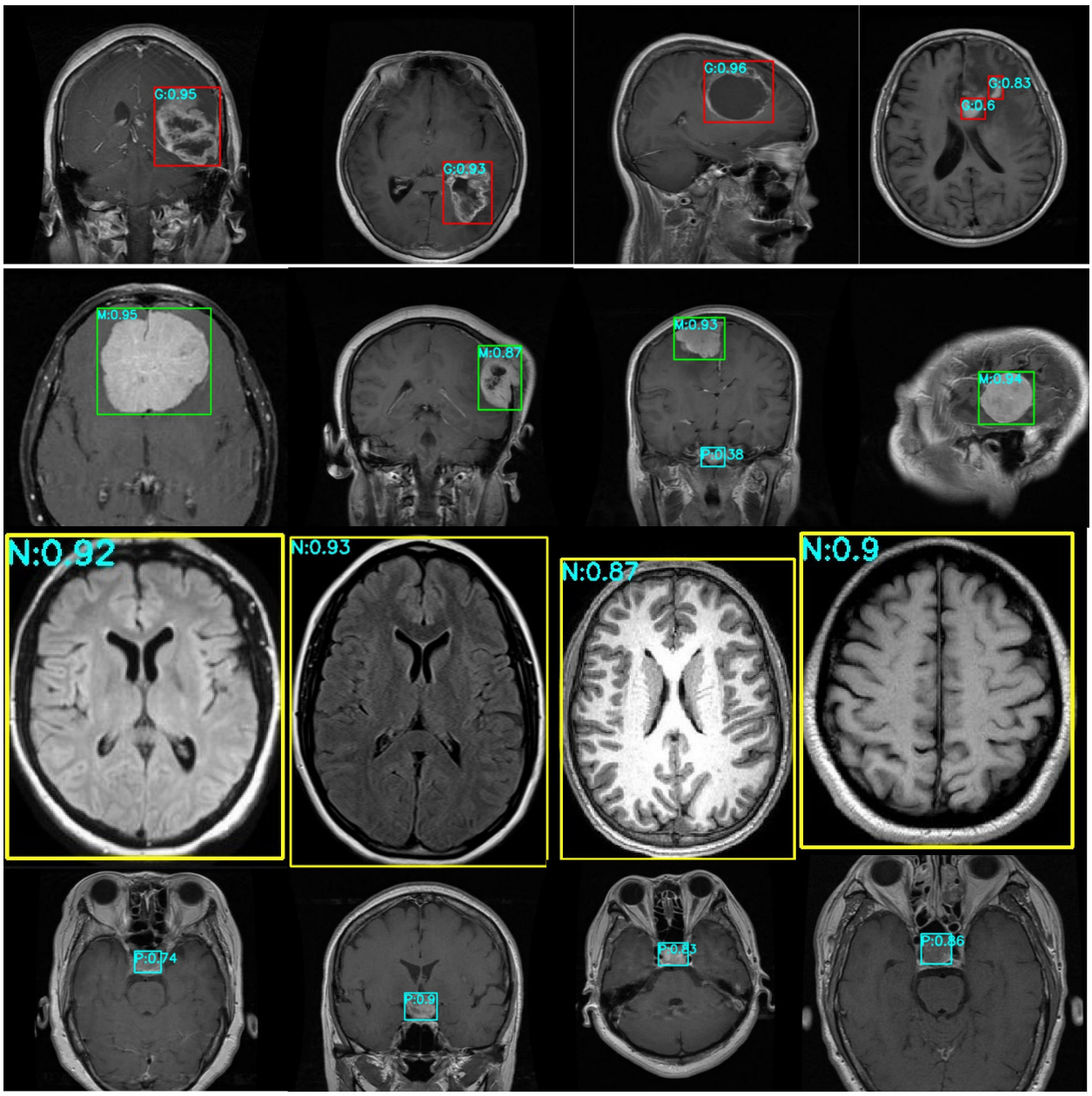


FIGURE 9
Effect diagram of brain tumor detection on the dataset.

YOLOv11n-HA is still better than YOLOv9s in terms of mAP50. Compared with that in YOLOv10n, the number of parameters in YOLOv11n-HA decreased by 7.0%, computational GFLOPs decreased by 29.8%, and the mAP50 and mAP50–95 indicators increased by 1.4% and 1.1%, respectively.

5 Discussion

This study introduces two key improvements to the original YOLOv11 model. First, it improves the YOLOv11 network structure by adding the Spatial attention, two newly designed Shuffle3D attention schemes, and Dual-channel attention. Second, it improves the loss function by introducing a hook function to adjust the CIoU loss, amplify penalties for low-quality predictions, and accelerate network convergence. The ablation experiment proved that, compared with native YOLOv11n, YOLOv11n-HA increased mAP50 and mAP50–95 by 1% and 1.4%, respectively, while the model parameters and computational GFLOPs decreased by 1.4% and 2.7%, respectively. Compared to other state-of-the-art models, YOLOv11n-HA achieved a superior recognition rate.

Figure 9 presents the test results for the Kaggle brain tumor dataset. The red box and G represent Glioma, the green box and M represents Meningioma, the yellow box and N represent No tumor, the cyan box and P represents Pituitary. The numbers behind represent the probability value of belonging to this class.

This study makes a significant contribution to the literature because it introduces a lightweight, computationally efficient model that achieves superior detection performance compared to state-of-the-art methods, thereby offering a practical solution for clinical applications with hardware constraints.

Further, this study addresses a critical challenge in medical imaging, accurate and rapid detection of brain tumors, by combining deep learning innovations with clinical relevance, offering insights that bridge technical development and healthcare impact. The proposed model achieves a strong balance between detection performance and computational efficiency, making it especially suitable for clinical deployment where hardware limitations exist. By providing accurate, real-time tumor localization in MRI images, this work contributes toward scalable and practical AI-assisted diagnostic solutions for healthcare settings.

6 Conclusion

This study used YOLOv11n to detect brain tumors in a public MRI dataset from Kaggle and introduced two key improvements. The first enhanced the network structure by integrating attention mechanisms, namely Shuffle3D attention and Dual-channel attention, which are newly designed in this study. The second introduces a new loss function, HKCIoU, which amplifies the loss for poorly predicted boxes via the hook function to accelerate network convergence. Ablation experiments demonstrate that mAP50 increased to 96.8% and mAP50–95 to 79.5%, with a 2.7% decrease in the number of parameters and a 7.8% decrease in GFLOPs.

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

WH: Methodology, Conceptualization, Investigation, Funding acquisition, Writing – review & editing, Project administration, Writing – original draft. XD: Writing – original draft, Visualization, Validation, Methodology, Software. GW: Validation, Methodology, Conceptualization, Resources, Writing – review & editing, Investigation, Funding acquisition. YD: Data curation, Visualization, Validation, Investigation, Writing – review & editing, Formal analysis, Resources. AY: Data curation, Formal analysis, Resources, Writing – review & editing, Visualization, Project administration.

Funding

The author(s) declare financial support was received for the research and/or publication of this article. This study was supported by the Key Research and Promotion Special Project of Xuchang City in 2025 (No. 2025090), 2026 Key Scientific Research Projects of Higher Education Institutions in Henan Province(No. 26B460032), and the Young Key Teacher Training Program (No. ZYKJQNGG2510) of Zhongyuan Institute of Science and Technology.

Acknowledgments

The authors would like to thank the Zhongyuan Institute of Science and Technology for its strong support of this study.

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 Generative AI was used in the creation of this manuscript. Generative AI was used in medical background knowledge.

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

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. McFaline-Figueroa JR, Lee EQ. Brain tumors. *Am J Med*. (2018) 131:874–82. doi: 10.1016/j.amjmed.2017.12.039

2. Jung AY. Basics for pediatric brain tumor imaging: techniques and protocol recommendations. *Brain Tumor Res Treat*. (2024) 12:1. doi: 10.14791/btrt.2023.0037

3. Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, et al. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro Oncol*. (2021) 23:1231–51. doi: 10.1093/neuonc/noab106

4. Olszak J, Zalewa K, Orłowska D, Bartoszek L, Kapłan W, Poleszczuk K, et al. Somatic and psychiatric symptoms of brain tumors - a review of the literature. *J Educ Health Sport*. (2024) 71:55845. doi: 10.12775/jehs.2024.71.55845

5. Lundy P, Domino J, Ryken T, Fouke S, McCracken DJ, Ormond DR, et al. The role of imaging for the management of newly diagnosed glioblastoma in adults: a systematic review and evidence-based clinical practice guideline update. *J Neurooncol*. (2020) 150:95–120. doi: 10.1007/s11060-020-03597-3

6. Priyadarshini P, Kanungo P, Kar T. Multigrade brain tumor classification in MRI images using Fine tuned efficientnet. In: *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, vol. 8. Amsterdam: Elsevier Ltd (2024). p. 100498. doi: 10.1016/j.prime.2024.100498

7. Viola P, Jones M. Rapid object detection using a boosted cascade of Simple features. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* Piscataway, NJ: IEEE (2001). doi: 10.1109/cvpr.2001.990517

8. Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Piscataway, NJ: IEEE (2005). p. 886–93. doi: 10.1109/cvpr.2005.177

9. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE (2014). p. 580–7. doi: 10.1109/cvpr.2014.81

10. Girshick R. Fast R-CNN. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. Piscataway, NJ: IEEE (2015). p. 1440–8. doi: 10.1109/iccv.2015.169

11. Ezhilarasi R, Varalakshmi P. Tumor detection in the brain using faster R-CNN. In: *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)* Piscataway, NJ: IEEE (2018). p. 388–92. doi: 10.1109/i-smac.2018.8653705

12. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot multiBox detector. In: *Computer Vision - ECCV 2016*. Amsterdam, Netherlands. Heidelberg: Springer-Verlag (2016) p. 21–37. doi: 10.1007/978-3-319-46448-0_2

13. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* Piscataway, NJ: IEEE (2016). p. 779–88. doi: 10.1109/cvpr.2016.91

14. Almufareh MF, Imran M, Khan A, Humayun M, Asim M. Automated brain tumor segmentation and classification in MRI using YOLO-based deep learning. *IEEE Access*. (2024) 12:16189–207. doi: 10.1109/access.2024.3359418

15. Su P, Han H, Liu M, Yang T, Liu S. MOD-YOLO: Rethinking the YOLO architecture at the level of feature information and applying it to crack detection. *Expert Syst Appl*. (2024) 237:121346. doi: 10.1016/j.eswa.2023.121346

16. Ren Z, Yao K, Sheng S, Wang B, Lang X, Wan D, et al. YOLO-SDH: improved YOLOv5 using scaled decoupled head for object detection. *Int J Mach Learn Cyber*. (2024) 16:1643–60. doi: 10.1007/s13042-024-02357-3

17. Tadjine C, Ouafi A, Taleb-Ahmed A, El Hillali Y, Rivenq A. Object detection based on Logistic Objects in Context (LOCO) dataset: an improved dataset split and performance on NVIDIA Jetson Nano. *J Real-Time Image Proc*. (2025) 22:98. doi: 10.1007/s11554-025-01673-3

18. Taha AM, Aly SA, Darwish MF. Detecting Glioma, Meningioma, and Pituitary Tumors, and Normal Brain Tissues based on Yolov11 and Yolov8 Deep Learning Models. *arXiv*. (2025). Available online at: https://arxiv.org/abs/2504.00189ss (Accessed August 20, 2025).

19. Yang T, Lu X, Yang L, Yang M, Chen J, Zhao H. Application of MRI image segmentation algorithm for brain tumors based on improved Yolo. *Front Neurosci*. (2025) 18:1510175. doi: 10.3389/fnins.2024.1510175

20. Islam J, Furqon EN, Farady I, Lung C-W, Lin C-Y. Early alzheimer's disease detection through YOLO-based detection of hippocampus region in MRI images. In: *2023 Sixth International Symposium on Computer, Consumer and Control (IS3C)* Piscataway, NJ: IEEE (2023). p. 32–5. doi: 10.1109/is3c57901.2023.00017

21. Diwan T, Anirudh G, Tembhurne JV. Object detection using yolo: Challenges, architectural successors, datasets and applications. *Multimedia Tools Applications*. (2022) 82:9243–75. doi: 10.1007/s11042-022-13644-y

22. Hussain M. YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. *Machines*. (2023) 11:677. doi: 10.3390/machines11070677

23. Khanam R, Hussain M. Yolov11: An overview of the key architectural enhancements. *arXiv*. (2024). doi: 10.48550/arXiv.2410.17725

24. Kharb A, Chaudhary P. Designing efficient brain tumor classifier using hybrid EfficientNet-faster R-CNN deep learning model. *Eng Res Express*. (2024) 6:035216. doi: 10.1088/2631-8695/ad63fa

25. Cheng J. brain tumor dataset(2017). Available online at: https://figshare.com/articles/dataset/brain_tumor_dataset/1512427 (Accessed August 20, 2025).

26. Hikmah NF, Hajjanto AD A, Surbakti AF, Prakosa NA, Asmaria T, Sardjono TA. Brain tumor detection using a MobileNetV2-SSD model with modified feature pyramid network levels. *IJECE*. (2024) 14:3995. doi: 10.11591/ijece.v14i4.pp3995-4004

27. Alsufyani A. Performance comparison of deep learning models for MRI-based brain tumor detection. *AIMS Bioengineering*. (2025) 12:1–21. doi: 10.3934/bioeng.2025001

28. Chen A, Lin D, Gao Q. Enhancing brain tumor detection in MRI images using YOLO-NeuroBoost model. *Front Neurol*. (2024) 15:1445882. doi: 10.3389/fneur.2024.1445882

29. Merlin. Br35H:Brain tumor detection(2020). Available online at: https://www.heywhale.com/mw/dataset/61d3e5682d30dc001701f728 (Accessed August 20, 2025).

30. Magesh. *Brain Tumor Dataset*. Des Moines: Roboflow (2024). Available online at: https://universe.roboflow.com/magesh-kctcd/brain-tumor-3rrwu (Accessed August 20, 2025).

31. Kang M, Ting FF, Phan RC-W, Ting C-M. PK-yolo: Pretrained knowledge guided Yolo for brain tumor detection in multiplanar MRI slices. In: *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* Piscataway, NJ: IEEE (2025). p. 3732–41. doi: 10.1109/wacv61041.2025.00367

32. Monisha SMA, Rahman R. Brain tumor detection in MRI based on federated learning with YOLOv11. *arXiv*. (2025). Available online at: https://arxiv.org/abs/2503.04087 (Accessed August 20, 2025).

33. Dulal R, Dulal R. Brain tumour identification using improved yolov8. *Int J Complexity Appl Sci Technol*. (2025) 1. doi: 10.1504/ijcast.2025.10071167

34. Wahidin MF, Kosala G. Brain tumor detection using YOLO models in MRI images. In: *2025 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS)* Piscataway, NJ: IEEE (2025). p. 1–6. doi: 10.1109/icadeis65852.2025.10933433

35. Bai R, Xu G, Shi Y. SCC-YOLO: An improved object detector for assisting in Brain tumor diagnosis. *Proc 2025 Int Conf Health Big Data*. (2025), 114–20. doi: 10.1145/3733006.3733026

36. Zhao Y, Jiang Z. Yolo-WWBI: An optimized YOLO11 algorithm for PCB defect detection. *IEEE Access*. (2025) 13:74288–97. doi: 10.1109/access.2025.3564734

37. Rao H, Zhan H, Wang R, Yu J. A lightweight and enhanced YOLO11-based method for small object surface defect detection. (2025). doi: 10.21203/rs.3.rs-6093937/v1

38. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D. Distance-IOU loss: Faster and better learning for bounding box regression. *Proc AAAI Conf Artif Intelligence*. (2020) 34:12993–3000. doi: 10.1609/aaai.v34i07.6999

39. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. In: *2017 IEEE International Conference on Computer Vision (ICCV)* Piscataway, NJ: IEEE (2017). doi: 10.1109/iccv.2017.324

40. Zhang QL, Yang YB. Sa-net: Shuffle attention for deep convolutional neural networks. In: *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* Piscataway, NJ: IEEE (2021). p. 2235–9. doi: 10.1109/icassp39728.2021.9414568

41. Yang L, Zhang RY, Li L, Xie X. SimAM: A simple. *Parameter-Free Attention Module Convolutional Neural Networks (PMLR)*. (2021) 139:11863–74. Available online at: https://proceedings.mlr.press/v139/yang21o.html (Accessed August 20, 2025).

42. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* Piscataway, NJ: IEEE (2016). p. 2818–26. doi: 10.1109/cvpr.2016.308

43. Sorour1 A. MRI for brain tumor with bounding boxes(2024). Available online at: https://www.kaggle.com/datasets/ahmedsorour1/mri-for-brain-tumor-with-bounding-boxes (Accessed August 20, 2025).

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 29 August 2025
DOI 10.3389/fmed.2025.1636059

Check for updates

OPEN ACCESS

EDITED BY
Deepti Deshwal,
Maharaja Surajmal Institute of Technology, India

REVIEWED BY
Ahmad ALI,
Shenzhen University, China
S. Suchitra,
Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, India

*CORRESPONDENCE
Jawad Rasheed
jawad.rasheed@izu.edu.tr

RECEIVED 27 May 2025
ACCEPTED 18 August 2025
PUBLISHED 29 August 2025

CITATION
Naeem AB, Osman O, Alsubai S, Cevik T, Zaidi A and Rasheed J (2025) Lightweight CNN for accurate brain tumor detection from MRI with limited training data. *Front. Med.* 12:1636059. doi: 10.3389/fmed.2025.1636059

COPYRIGHT
© 2025 Naeem, Osman, Alsubai, Cevik, Zaidi and Rasheed. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Lightweight CNN for accurate brain tumor detection from MRI with limited training data

Awad Bin Naeem[1,2], Onur Osman[3], Shtwai Alsubai[4], Taner Cevik[5], Abdelhamid Zaidi[6] and Jawad Rasheed[7,8,9,10]*

[1]Department of Computer Science, National College of Business Administration and Economics, Multan, Pakistan, [2]Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna, Italy, [3]Department of Electrical and Electronics Engineering, 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 Rumeli University, Istanbul, Türkiye, [6]Department of Mathematics, College of Science, Qassim University, Buraydah, Saudi Arabia, [7]Department of Computer Engineering, Istanbul Sabahattin Zaim University, Istanbul, Türkiye, [8]Department of Software Engineering, Istanbul Nisantasi University, Istanbul, Türkiye, [9]Research Institute, Istanbul Medipol University, Istanbul, Türkiye, [10]Applied Science Research Center, Applied Science Private University, Amman, Jordan

Aim: This study aims to develop a robust and lightweight deep learning model for early brain tumor detection using magnetic resonance imaging (MRI), particularly under constraints of limited data availability. Objective: To design a CNN-based diagnostic model that accurately classifies MRI brain scans into tumor-positive and tumor-negative categories with high clinical relevance, despite a small dataset. Methods: A five-layer CNN architecture—comprising three convolutional layers, two pooling layers, and a fully connected dense layer—was implemented using TensorFlow and TFlearn. A dataset of 189 grayscale brain MRI images was used, with balanced classes. The model was trained over 10 epochs and 202 iterations using the Adam optimizer. Evaluation metrics included accuracy, precision, recall, F1 Score, and ROC AUC.
Results: The proposed model achieved 99% accuracy in both training and validation. Key performance metrics, including precision (98.75%), recall (99.20%), F1-score (98.87%), and ROC-AUC (0.99), affirmed the model's reliability. The loss decreased from 0.412 to near zero. A comparative analysis with a baseline TensorFlow model trained on 1,800 images showed the superior performance of the proposed model.
Conclusion: The results demonstrate that accurate brain tumor detection can be achieved with limited data using a carefully optimized CNN. Future work will expand datasets and integrate explainable AI for enhanced clinical integration.

KEYWORDS

MRI images, deep learning, medical diagnosis, computer-aided diagnosis, healthcare, neuroimaging

1 Introduction

A technique for training a computer to create original representations from unprocessed data is called deep learning. The network's popularity may be attributed to its hierarchical and layered structure. Convolutional Neural Networks (CNNs) acquire properties through an object compositional hierarchy, starting with simple edges and progressing to more intricate forms. By layering convolutional and pooling layers, this is achieved. By lowering the feature

map, pooling combines similar traits into one, and each convolutional layer identifies local conjunctions of features from the preceding layer. Researchers in neuroscience have also benefited from deep learning, as they are starting to address issues related to neuroimaging. Deep Learning has garnered significant interest due to its ability to address problems across various domains, including medical image analysis. In Palestine, cancer is now the second leading cause of death for both men and women, but over the next decades, it is predicted to overtake all other causes of death (1).

Research has shown that the most effective means of lowering death from brain cancer is early diagnosis and treatment. A low-grade growth that develops slowly will eventually evolve into a neoplasm that grows rapidly. As a result, the first tumor identification and categorization helped to anticipate the prognosis and treatment plan by supporting the assessment of the tumor's grade and aggressiveness. The diagnosis of brain tumors is mostly reliant on medical imaging (2). One of the most efficient methods currently used for tumor detection is magnetic resonance imaging (MRI). A powerful magnetic flux, radiofrequency pulses, and a laptop is employed to process tomography imaging data to produce detailed images of soft tissues and organs. It aids medical professionals in treating illnesses. The main reason for tomography's popularity is that it is a more suitable designation than X-rays (3).

Noise significantly degrades medical images, including MRIs. This is largely due to knowledge acquisition systems, multiple sources of interference, operator error, and other factors that impact imaging mensuration processes and can lead to significant classification errors (4). This approach typically requires a basic microscope and may result in a different or incorrect diagnosis, yet it is often inappropriate when dealing with human life. It emphasizes the need for power-assisted systems, high-precision systems, or diagnostic systems (CADx) (5). The CADx system is essential for medical institutions, as it supports the judgments made by doctors and radiologists. It may be challenging to create a highly automated and economical diagnostic system as a result (6).

Gliomas are the most prevalent and aggressive kind of brain tumor, with a very short survival time for the highest grade. Therefore, therapy planning may be a crucial step in raising the medical patients' standards of living. One popular imaging modality for evaluating these tumors may be MRI (7). These days, with numerous instances and massive volumes of objective data analysis, computer-based medical image analysis is gaining popularity due to its speed and intelligence, surpassing manual methods. By varying the excitation and repetition durations, magnetic resonance imaging may produce notably unique tissue types, making it an incredibly adaptable tool for studying various structures of interest. A single magnetic resonance imaging scan is insufficient to phase the growth and all of its subregions fully. Convolutional Neural Networks (CNNs) have demonstrated high effectiveness in identifying cell division events in two-dimensional microscopic anatomy pictures within the field of medical image analysis. When it comes to machine learning strategies, deep learning is undoubtedly the best option for many imaging tasks. The possibility of deep learning-based automated diagnosis of brain illnesses will arise from the availability of large neuroimaging data sets for training. MRI is a frequently used medical imaging method that offers information on the identification of brain tumors (8). One of the main challenges a physician has after reviewing the tomography data is determining how much time and effort to devote to tumor detection. These days, CNNs are used for the majority of picture classification problems due to their superior accuracy and precision over other currently used techniques. The accuracy and precision of tumor detection and identification have increased due to the use of CNNs for image classification (9).

2 Related work

Over the last 20 years, the detection of brain cancers using MRI has undergone significant advancements, thanks to the integration of deep learning (DL), traditional machine learning (ML), and conventional image processing techniques. This section discusses the main categories of methodologies and provides an overview of how our research contributes to and expands upon the existing body of literature.

2.1 Conventional techniques for machine learning and segmentation

Most of the early work uses unsupervised clustering and custom feature extraction. Due to their ability to separate picture intensities into clusters that represent normal and diseased tissue regions, segmentation techniques like fuzzy C-Means (FCM) and K-Means clustering have been widely used (10–12). Despite achieving basic localization, these methods were very susceptible to noise and required human parameter adjustment. Changes aimed at improving segmentation accuracy, such as region-expanding algorithms (13, 14) and gray-level histograms (15), were computationally expensive and inconsistent, particularly in low-contrast or early-stage tumors where borders were not obvious. For feature extraction and classification, further research employs learning vector quantization, support vector machines (SVMs), and artificial neural networks (ANNs) (16, 17). These earlier methods, however, sometimes did not work with diverse patient datasets and needed careful feature engineering.

2.2 Techniques based on deep learning and CNN

CNNs have been used extensively in medical imaging applications due to their effectiveness in computer vision (26, 27). CNNs eliminate the requirement for human feature design by automatically extracting hierarchical features. Models like AlexNet, VGG16, and ResNet have been modified to perform tasks related to brain tumor classification and segmentation (18, 19). Although these designs have demonstrated outstanding performance, they often rely on large, annotated datasets, which are challenging to collect in the medical field due to privacy concerns and high labeling costs. To manage volumetric MRI data and capture spatial relationships between image slices, 3D CNNs have been the subject of several studies (20). Although these models improve the accuracy of segmentation tasks, their computational cost makes them unsuitable for real-time applications or situations with limited resources. Similar studies have been conducted on Stacked Autoencoders (SAEs) and Deep Belief Networks (DBNs) (21), but in the lack of suitable data, training these deep models from scratch may lead to overfitting.

2.3 Domain adaptation and learning transfer

By utilizing pre-trained networks as feature extractors for MRI classification, which have been trained on natural image datasets such as ImageNet, researchers have employed transfer learning to reduce the need for large datasets (22, 23). When paired with domain-specific fine-tuning, it can accelerate training and enhance generalization. However, insufficient feature representations may result from the domain mismatch between natural and medical images. ResNet or InceptionV3 versions that have been carefully altered and work well on binary classification tasks are used in certain studies. Clinical safety criteria, such as recall and AUC, which are essential for real-world diagnosis, are seldom used to evaluate models.

2.4 Methods for multimodal MRI and synthesis

To collect different tissue contrasts, advanced segmentation algorithms often use several MRI modalities. Studies like the BraTS Challenge and BraSyn Benchmark (24, 25) demonstrate the challenges that arise when sequences are erratic or nonexistent, while also emphasizing the advantages of multimodal input. To fill in the gaps, several studies have explored the creation of synthetic MRIs using GANs or autoencoders; however, these methods require a complex design and are not ideal for use in situations with limited data.

3 Materials and methods

Cancer remains one of the most life-threatening diseases worldwide, and early detection is critical for effective treatment. MRI is a widely used, non-invasive imaging technique that helps identify abnormalities in the brain, including cancerous tumors. In recent years, machine learning—particularly image classification techniques—has demonstrated significant promise in improving the accuracy and speed of cancer detection using MRI. This study examines the DL-based application in developing a CNN for brain tumor detection using MRI scans. The proposed CNN architecture consists of five layers, specifically designed to classify MRI images into cancerous and non-cancerous categories with high accuracy.

3.1 Data acquisition

Data plays a crucial part in machine learning systems. The dataset utilized in this work was available from the UCI Machine Learning Repository and Kaggle, both of which are publicly accessible. The dataset downloaded from Kaggle and is accessible at https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection/data (last accessed: January 10, 2025), and the second dataset is available at https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri (last accessed: January 20, 2025).

3.2 Methodology and model architecture

The architecture employed in this study is based on a CNN design, which is particularly effective for image classification tasks. CNNs typically include the following core components:

- Convolutional Layers: Extract feature maps from the input image using learned filters and apply non-linear activation functions (e.g., ReLU).
- Pooling Layers: Reduce the spatial size of feature maps, enhance computational efficiency, and mitigate overfitting—max-pooling is the most commonly used technique.
- Fully Connected (Dense) Layers: Interpret the extracted features and produce classification decisions; each neuron is connected to all neurons in the previous layer.

The proposed model consists of five primary layers: three convolutional layers, two max-pooling layers, and a fully connected dense layer. The architecture is implemented using the high-level TensorFlow Layers API, which streamlines the creation of neural networks by offering functions to define convolutional, pooling, and dense layers, along with activation functions and regularization options such as dropout.

Figure 1 illustrates the sequential layer-wise architecture of the CNN, clarifying the dimensional transformation of MRI data from input through convolution, pooling, and dense layers to the final binary classification. The model was trained using the Adam optimizer with the following parameters: ε = 1e-8, β_1 = 0.9, β_2 = 0.999, and a learning rate of 0.001. To avoid overfitting, a dropout layer with a 0.5 rate was added after the dense layer.

The model processes grayscale MRI images resized to 128 × 128 × 1. The first convolutional layer applies 32 filters (3 × 3) with ReLU activation, followed by a 2 × 2 max-pooling operation. The second convolutional layer utilizes 64 filters (3 × 3) with ReLU activation and an additional 2 × 2 max-pooling operation. The third convolutional layer consists of 128 filters (3 × 3), followed by another pooling operation. The output of the convolutional stages is flattened and passed to a dense layer with 128 neurons, also using ReLU activation. Ultimately, a single output neuron with sigmoid activation yields a binary classification decision (tumor-positive or tumor-negative).

Figure 2 illustrates the initial layers of the CNN, including the first convolution and pooling layers. The initial convolutional and pooling layers extract low-level spatial features, such as edges and texture gradients, which are essential for differentiating tumor boundaries from normal tissue in MRI images, including edges, lines, and simple textures. The visual representation highlights how spatial information is preserved while dimensionality is reduced.

Figure 3 illustrates the intermediate layers of the CNN, which include deeper convolutional layers with a greater number of filters. These layers extract high-level, abstract features such as tumor shapes, boundaries, and textures. These deeper layers abstract high-level semantic features such as irregular tumor shapes, enhancing the model's ability to distinguish pathological from healthy brain structures.

Figure 4 focuses on the final layers of the CNN, including the fully connected dense layer and the output neuron. These layers are responsible for interpreting the extracted features and making the final classification decision. The use of sigmoid activation in the

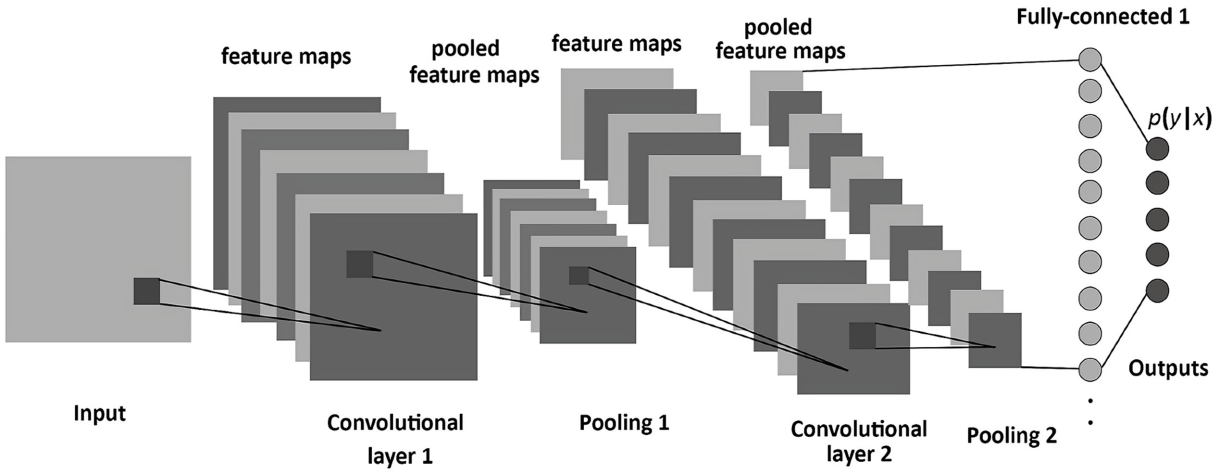


FIGURE 1
CNN architecture for brain tumor classification, showing layers for feature extraction and final classification from MRI input images.

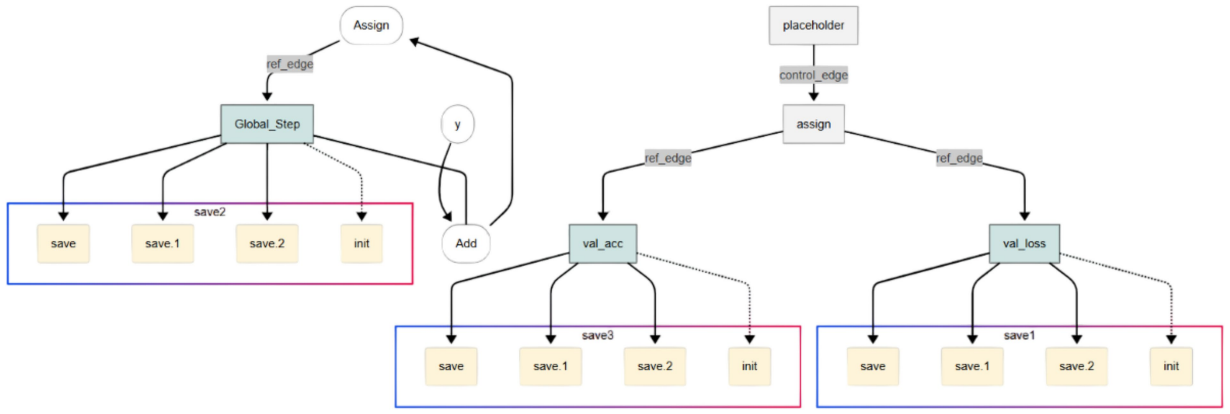


FIGURE 2
Feature extraction in early CNN layers showing low-level spatial features such as edges and textures derived from tumor MRI images.

output neuron enables the model to output a probability score indicating the presence or absence of a brain tumor.

To complement these visual representations, Table 1 provides a detailed layer-wise summary of the CNN model, listing input/output dimensions, number of filters or neurons, kernel and pooling sizes, and activation functions used at each stage. Moreover, it offers a concise yet thorough reference for understanding the architecture's design and function.

The TensorFlow Layers API enables the construction of these components with functions such as:

- conv2d(): Defines 2D convolutional layers with specified parameters.
- max_pooling2d(): Creates pooling layers to down-sample feature maps.
- dense(): Builds fully connected layers for classification.

Due to the complexity of the computational graph, it is segmented for clarity across Figures 2–4, with each segment representing a critical stage in the data transformation and classification process.

4 Experimental setup and results

The proposed CNN model was trained and evaluated using a dataset comprising 189 MRI images, with an equal balance between cancerous and non-cancerous cases. The dataset was stratified into training, validation, and testing subsets to maintain balanced representation of tumor-positive and tumor-negative cases. Table 2 presents the data distribution according to the train and test splits. Training was performed for 10 epochs with a batch size of 18, yielding approximately 202 iterations. Key performance metrics, including accuracy, loss, and ROC-AUC, were monitored via TensorBoard throughout training. Hyperparameters were consistently maintained across experiments to enhance reproducibility. Tracking accuracy and loss over 202 iterations with TensorBoard enabled validation of stable convergence and early detection of overfitting, which is critical given the limited dataset size.

Because of the small sample size, we utilized TensorFlow's "ImageDataGenerator" to supplement data in real time and increase generalization. The augmentation pipeline used horizontal flipping ($p = 0.5$) to mimic mirrored brain orientations, small-angle rotations

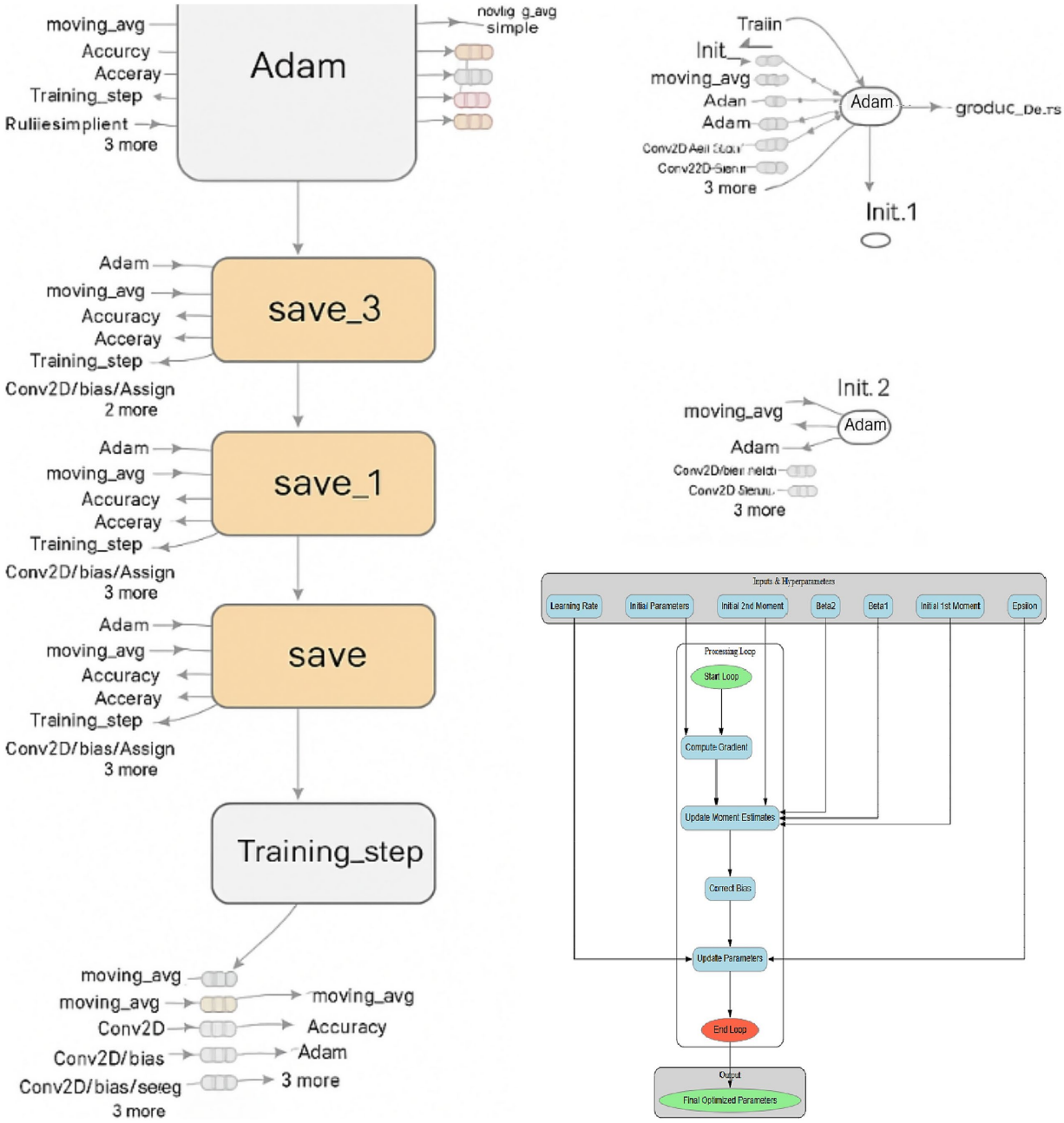


FIGURE 3
Intermediate CNN layers highlighting deeper convolutions and expanded feature maps that capture high-level tumor features.

(±10°) to account for head tilt variability, random zoom (±5%) and translations (±5% of image dimensions) to simulate patient positioning differences, and Gaussian noise injection ($\sigma = 0.01$) to simulate MRI scanner acquisition noise. The augmentation pipeline contained:

- Horizontal Flipping: To represent mirrored anatomical configurations, has a chance of 0.5.
- Rotation: Random small-angle rotations within ±10°, to account for minor patient head tilts.
- Zoom: To mimic size differences across scanners, zoom in and out by up to 5%.
- Translation: An image dimension from vertical and horizontal shift up to 5%.
- Noise injection: MRI scanner acquisition noise is simulated using low-level Gaussian noise ($\sigma = 0.01$).

To accommodate for changes in intensity from scanner calibration, adjust brightness by ± 10%. To expose the model to a broader variety of real-world input conditions without needlessly extending the dataset on disk, these modifications to the training set were performed stochastically throughout each epoch. Each run started with a predefined random seed to maintain consistency. We can assure repeatability and back up our claims of strong

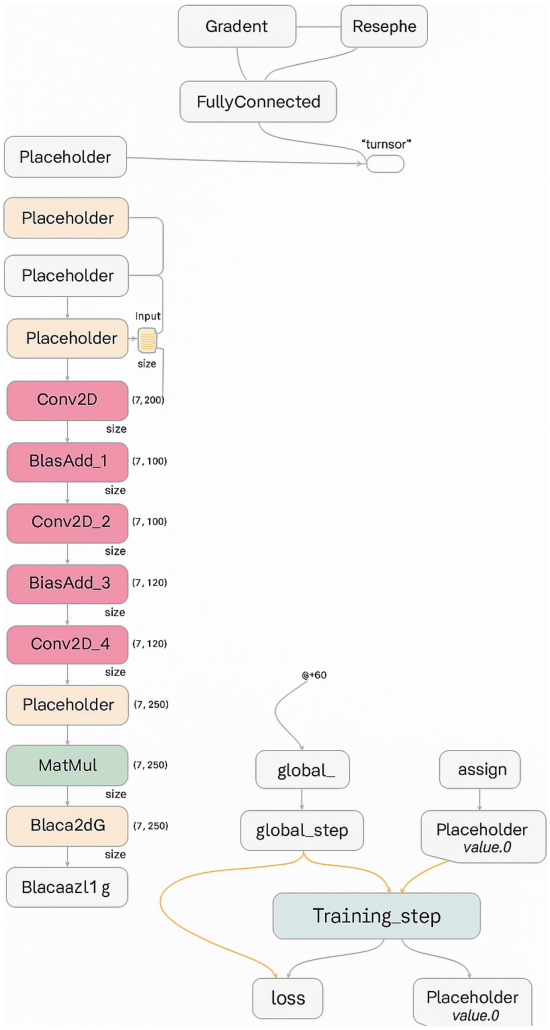


FIGURE 4
Final CNN layers: dense and sigmoid output units responsible for probabilistic classification of tumor presence.

TABLE 1 Layer-wise architecture of the proposed CNN model, detailing input/output shapes, filter counts, kernel sizes, activation functions, and pooling operations for each layer.

Layer type	Output shape	Activation	Notes
Input Layer	(128, 128, 1)	—	Grayscale MRI input
Conv2D	(128, 128, 32)	ReLU	32 filters, 3 × 3 kernel
MaxPooling2D	(64, 64, 32)	—	2 × 2 pool size
Conv2D	(64, 64, 64)	ReLU	64 filters, 3 × 3 kernel
MaxPooling2D	(32, 32, 64)	—	2 × 2 pool size
Conv2D	(32, 32, 128)	ReLU	128 filters, 3 × 3 kernel
MaxPooling2D	(16, 16, 128)	—	2 × 2 pool size
Flatten	(32768)	—	—
Dense	(128)	ReLU	Fully connected layer
Output (Dense)	(1)	Sigmoid	Binary classification output

TABLE 2 Dataset distribution across training, validation, and testing subsets, showing balanced representation of tumor-positive and tumor-negative MRI scans of the first dataset.

Dataset split	Number of images	Tumor-positive	Tumor-negative
Training	133	67	66
Validation	28	14	14
Testing	28	14	14
Total	189	95	94

generalization with short datasets by enabling other researchers to reproduce our preprocessing pipeline and see whether analogous augmentation tactics offer equivalent advances in other limited-data settings. In clinical contexts with limited and varied patient data, augmentation decreases overfitting, enhances feature diversity, and makes the model more usable.

The dataset used in this study consisted of MRI scans collected from multiple patients, with one representative scan per subject to minimize redundancy and prevent model bias. In cases where numerous scans were available per patient, only one scan was randomly selected to ensure that no patient's data appeared in both the training and validation sets. This procedure prevents data leakage, ensuring that the model's performance reflects genuine generalization rather than memorization of individual patient characteristics.

Figure 5 provides a visual overview of the dataset used in our experiments, distinguishing between cancerous and non-cancerous MRI brain scans. Our CNN effectively captured these differences in structural patterns and intensities for classification.

The model was trained for 35 epochs (840 iterations), achieving a peak validation accuracy of 98%. The model's high precision and recall indicate its potential as a clinical decision support tool to aid radiologists in more efficient brain tumor identification. Each training example that passes through the network in both forward and backward propagation constitutes one iteration.

The Adam optimiser was configured with a learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$. These values are known to offer stable and efficient convergence in deep learning models, especially when working with small datasets. They were selected after preliminary tuning and cross-referencing with prior studies demonstrating similar use cases in MRI image classification. Although extensive hyperparameter tuning was beyond the scope of this study, the choice of hyperparameters was based on standard values widely adopted in the literature for medical image classification tasks.

Figure 6 displays the tumor segmentation output, highlighting spatial tumor regions. The trained model not only classifies the presence of tumors but also enables the visualization of the detected tumor region. This segmentation capability adds clinical value by providing spatial context for the tumor's location and size.

Figure 7 illustrates the accuracy across iterations, which initially shows an uneven distribution but ultimately converges to zero as the iterations progress. The loss rate is a critical component of CNN and

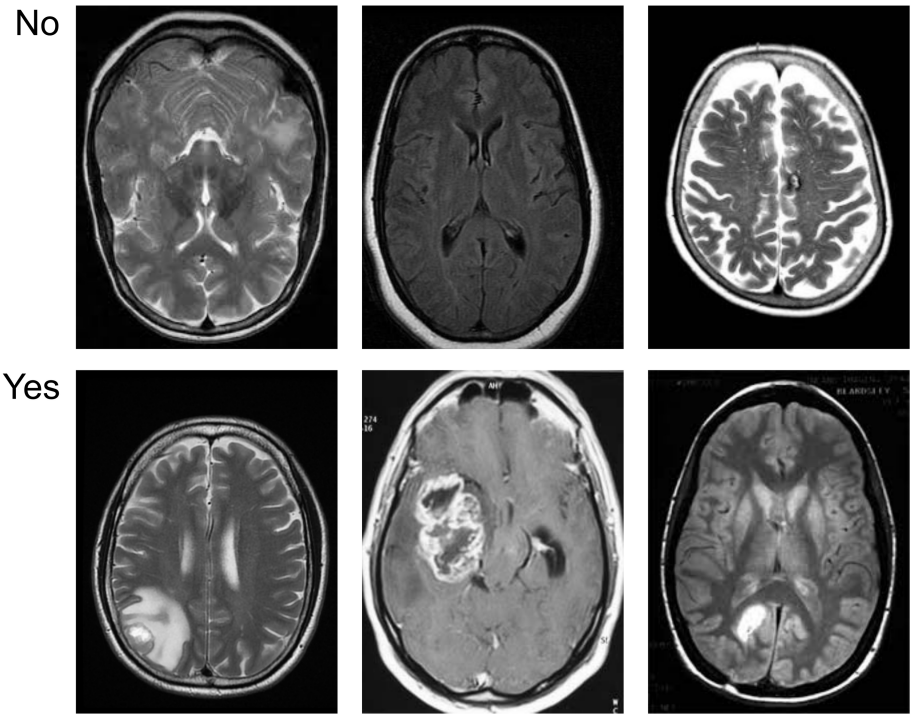


FIGURE 5
Sample visualization of the MRI dataset illustrating differences between tumor-positive and tumor-negative brain images.

is used to improve the CNN architecture. Despite the limited dataset, the proposed model effectively minimizes loss and enhances accuracy. Figure 8 presents the Receiver Operating Characteristic (ROC) curve with an AUC of 0.99, illustrating excellent diagnostic ability.

To further assess the performance of the proposed CNN-based model, standard classification metrics were computed, including precision, recall, F1-score, accuracy, and the area under the ROC-AUC curve. Table 3 consolidates critical performance metrics, including training accuracy (99%), validation accuracy (99%), loss rate reduction from 0.412 to nearly zero, precision, recall, F1-score, and ROC-AUC (0.99), providing a clear and concise overview of the model's effectiveness. Figure 9 illustrates the confusion matrix of both proposed and baseline models when tested with 600 test images of the second dataset. Additionally, Table 4 compares the performance of the proposed model with a baseline TensorFlow model trained on a larger dataset (1800 images) that has lower accuracy (98%) and higher loss (0.704). The proposed CNN model has superior performance despite the limited data.

The five-layer CNN architecture was selected to balance classification accuracy and computational efficiency on a limited dataset for prospective clinical use. Early research compared the recommended design to a more complex 8-layer CNN with an extra convolution-pooling block and a second dense layer. Despite reaching 99% training accuracy, the deeper model's validation accuracy plateaued at 96% after the 20th epoch and displayed peculiar loss oscillations, indicating overfitting due to the limited dataset size of 189 pictures. Across all training and validation sets, the five-layer model consistently reduced loss from 0.412 to near zero while maintaining 99% accuracy, demonstrating strong generalization capabilities. Furthermore, it reduced the number of parameters by approximately 38%, thereby decreasing training time on the same GPU from 7.8 s to 4.9 s per epoch. This efficiency directly supports the study's purpose of creating a lightweight diagnostic model suited for real-time inference in clinical settings, especially when resources are constrained. The architect's decision

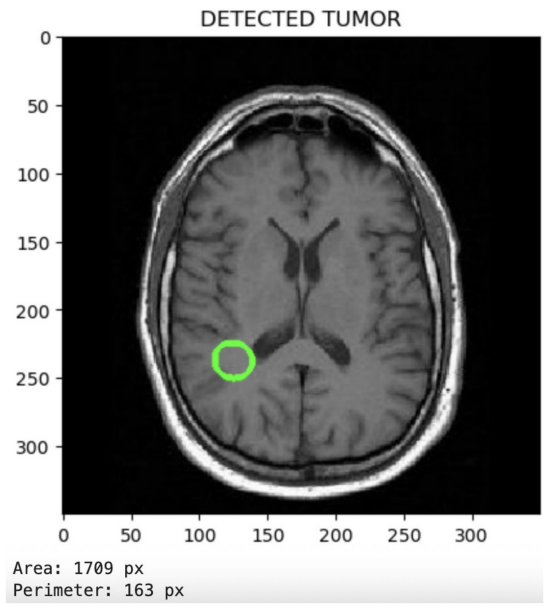


FIGURE 6
Segmentation output visualizing localized tumor regions, highlighting the model's spatial discrimination capabilities.

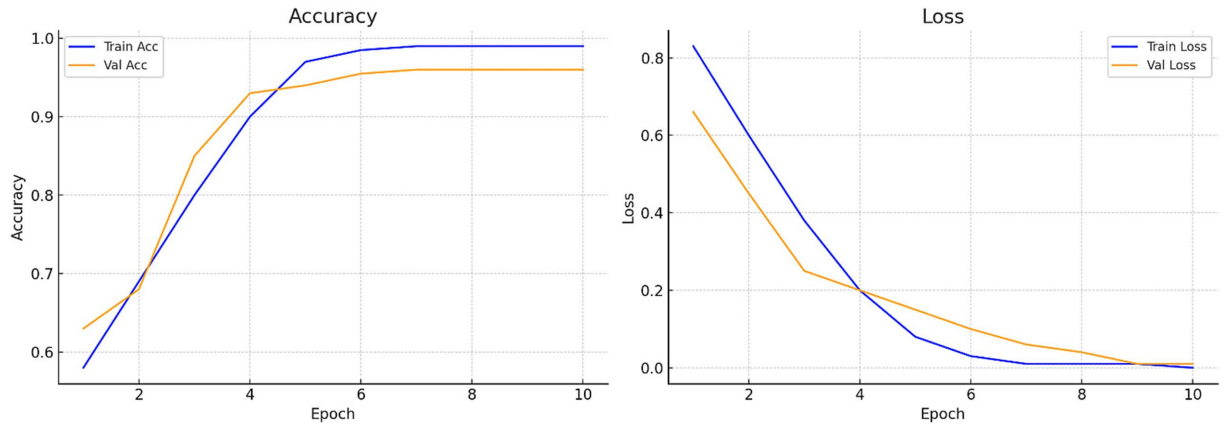


FIGURE 7
Accuracy and loss curves of the proposed model: Training loss progression illustrating reduction from 0.412 to near zero, reflecting stable model convergence.

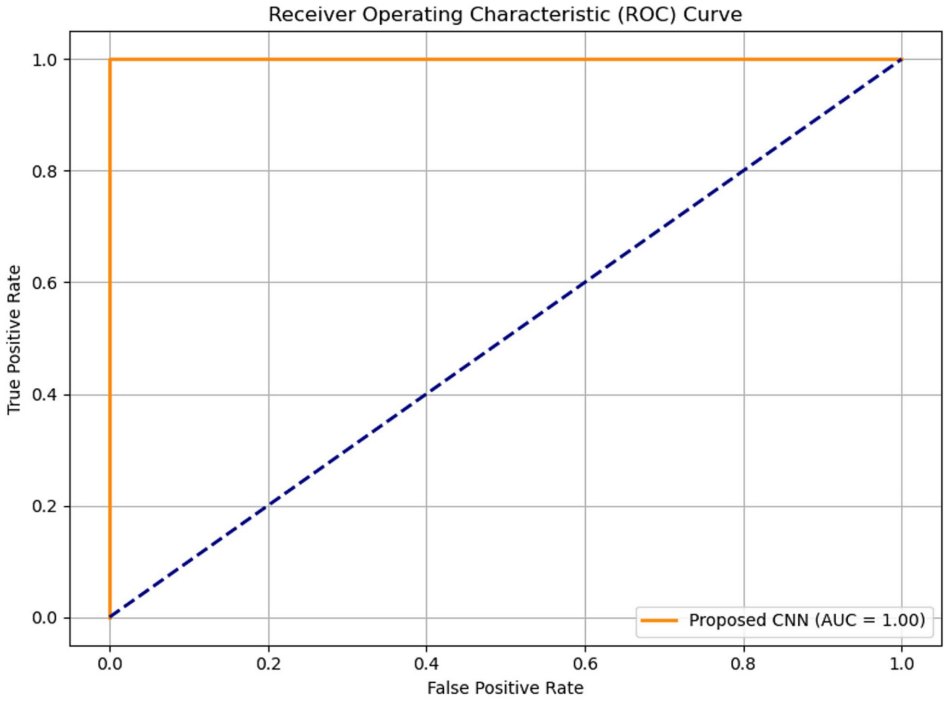


FIGURE 8
Receiver Operating Characteristic (ROC) curve of the proposed model with an AUC of 0.99, indicating excellent diagnostic accuracy.

TABLE 3 Performance metrics of the proposed CNN model, including accuracy, precision, recall, F1-score, ROC-AUC, and reduction in loss rate.

Metric	Value
Accuracy	99.00%
Precision	98.75%
Recall	99.20%
F1 Score	98.87%
ROC-AUC	0.99
Loss Reduction	0.412 → ~0.00

reflects the nature of the classification challenge. When utilizing MRI to identify brain cancers, spatial indicators such as tumor margins, regional intensity variations, and abnormal textural patterns are crucial. They may be successfully retrieved without having a massive network depth by utilizing three progressively deeper convolutional layers (32, 64, and 128 filters). According to feature map representations, the proposed CNN properly captured both low-level edge attributes and higher-level tumor form abstractions that were comparable to those in the deeper model. Given the dataset, processing settings, and observable performance limits, the five-layer CNN delivers the ideal blend of accuracy, resilience, and efficiency for this experiment.

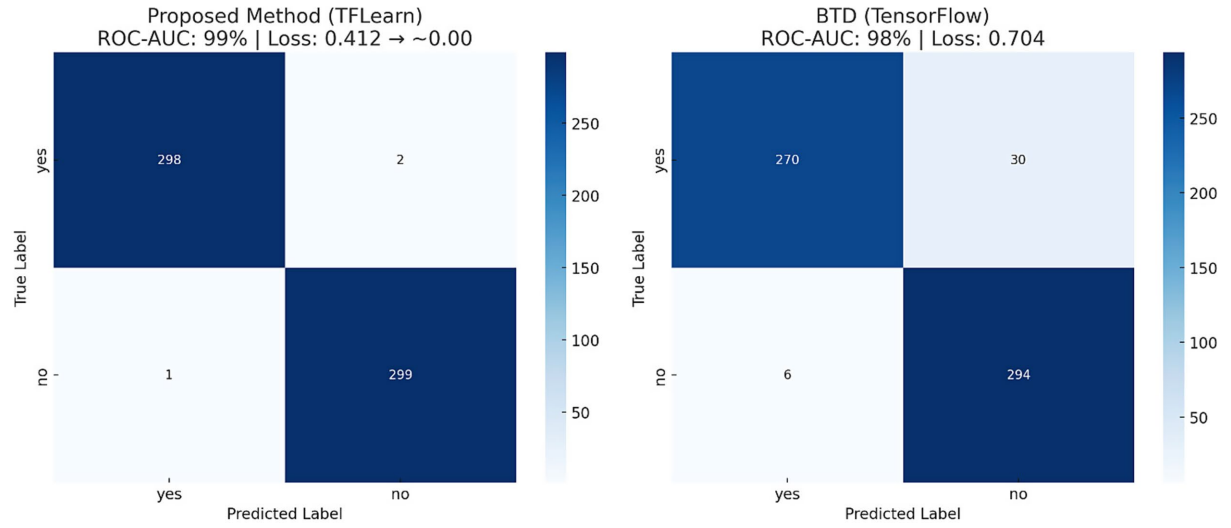


FIGURE 9
Confusion matrix showing true positive and true negative predictions, validating classification reliability.

TABLE 4 Comparative evaluation of the proposed CNN model versus a baseline TensorFlow implementation, highlighting improved performance with fewer training samples.

Method	Epochs	Iterations	Dataset	ROC-AUC	Loss rate
BTD (TensorFlow)	35	840	1800	98%	0.704
Proposed Method (TFLearn Based)	10	202	189	99%	0.412 → ~0.00

5 Conclusion

Deep learning has become a crucial tool in biomedical image analysis, particularly for applications such as brain tumor classification using MRI scans. For quicker model construction, the proposed technique employs CPU-based TensorFlow and TFLearn, as well as GPU-based TensorFlow. Deep learning (DL) techniques are increasingly employed in medical imaging for brain tumor detection and classification. The use of MRI is essential for detecting abnormal brain tissues, and accurate tumor diagnosis is vital for treatment planning. To categorize and diagnose brain tumors from a limited MRI dataset, the study employs a deep learning approach using a Convolutional Neural Network (CNN). The proposed model achieved 99% training and 99% validation accuracy, with a validation loss reduction from 0.412 to near 0.000 across 10 epochs. Additionally, the model attained an ROC-AUC of 0.99, confirming its strong discriminative capability. The proposed CNN model outperformed a baseline model trained on a larger dataset, achieving higher accuracy (99% vs. 98%) and lower validation loss (0.412 vs. 0.704), which indicates strong potential for deployment in real-time clinical diagnostics, especially in data-limited settings. The suggested CNN model may be used in real-world healthcare environments because of its lightweight design and exceptional diagnostic precision. In a radiology department's existing PACS (Picture Archiving and Communication System), a radiologist may use the model as an automated pre-screening tool to rank MRI images with a high likelihood of tumor incidence. Real-time feedback during diagnostic sessions could be provided by integrating the model with clinical decision support systems. Additionally, report authoring could be made easier by connecting to Radiology Information Systems (RIS). Because of its minimal computational requirements (4.9 s per epoch on a standard GPU), the model may also be implemented on-site in hospitals with limited resources, eliminating the need for cloud-based processing. Regulatory approval, interoperability with different MRI scanner outputs, and further validation across multiple-center datasets to ensure robustness are the remaining challenges. Before clinical utilization is widely accepted, these challenges need to be resolved.

6 Future directions

Future work will focus on expanding the dataset to improve model generalization and reduce bias. Integrating additional imaging modalities, such as Computed Tomography (CT) and Positron Emission Tomography (PET), as well as utilizing transfer learning with pre-trained models, may enhance performance. Exploring three-dimensional Convolutional Neural Networks (3D CNNs) can capture spatial context more effectively, while explainable AI methods, such as Gradient-weighted Class Activation Mapping (Grad-CAM), can improve interpretability. In the future, data augmentation techniques, including rotation, flipping, scaling, and brightness adjustment, can be employed to assess the model's generalization.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: the datasets analyzed for this study can be found at https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection/data (Last Accessed: January 10, 2025) https://www.kaggle.com/datasets/sartajbhuvaji/brain-tumor-classification-mri (Last Accessed: January 10, 2025).

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.

Author contributions

AN: Investigation, Validation, Methodology, Conceptualization, Software, Formal analysis, Writing – original draft. OO: Writing – original draft, Visualization, Writing – review & editing, Data curation. SA: Visualization, Writing – review & editing, Resources. TC: Visualization, Data curation, Writing – review & editing, Writing – original draft. AZ: Software, Formal analysis, Writing – original draft, Methodology, Investigation, Validation. JR: Resources, Conceptualization, 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 Generative 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

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. Kimberly WT, Sorby-Adams AJ, Webb AG, Wu EX, Beekman R, Bowry R, et al. Brain imaging with portable low-field MRI. *Nat Rev Bioeng*. (2023) 1:617–30. doi: 10.1038/s44222-023-00086-w

2. Kia M, Sadeghi S, Safarpour H, Kamsari M, Jafarzadeh Ghoushchi S, Ranjbarzadeh R. Innovative fusion of VGG16, MobileNet, EfficientNet, AlexNet, and ResNet50 for MRI-based brain tumor identification. *Iran J Comput Sci*. (2025) 8:185–215. doi: 10.1007/s42044-024-00216-6

3. Appavu N., "Brain tumor detection and classification using MRI with ResNet50 and hybrid AI deep learning techniques," 2025 international conference on data science, agents & artificial intelligence (ICDSAAI), Chennai, India: ICDSAAI, (2025), pp. 1–6.

4. Bilal H, Tian Y, Ali A, Muhammad Y, Yahya A, Izneid BA, et al. An intelligent approach for early and accurate predication of cardiac disease using hybrid artificial intelligence techniques. *Bioengineering*. (2024) 11:1290. doi: 10.3390/bioengineering11121290

5. Noh H., Hong S., Han B., "Learning deconvolution network for semantic segmentation," 2015 IEEE international conference on computer vision (ICCV), Santiago, Chile, IEEE. (2015), pp. 1520–1528.

6. Ahmed N, Rozina R, Ali A, et al. Image denoising for COVID-19 chest X-ray based on multi-scale parallel convolutional neural network. *Multimedia Systems*. (2023) 29:3877–90. doi: 10.1007/s00530-023-01172-0

7. Zhu L, Xue Z, Jin Z, Liu X, He J, Liu Z, et al. Make-A-volume: leveraging latent diffusion models for cross-modality 3D brain MRI synthesis In: H Greenspan, editor. Medical image computing and computer assisted intervention – MICCAI 2023. MICCAI 2023. Lecture notes in computer science. Cham: Springer (2023)

8. Shawon MTR, Shibli GMS, Ahmed F, Joy SKS. Explainable cost-sensitive deep neural networks for brain tumor detection from brain MRI images considering data imbalance. *Multimed Tools Appl*. (2025) 6:842. doi: 10.1007/s11042-025-20842-x

9. Nassar SE, Yasser I, Amer HM, Mohamed MA. A robust MRI-based brain tumor classification via a hybrid deep learning technique. *J Supercomput*. (2024) 80:2403–27. doi: 10.1007/s11227-023-05549-w

10. Alsarhan T, Ali SS, Ganapathi II, Ali A, Werghi N. PH-GCN: boosting human action recognition through multi-level granularity with pair-wise hyper GCN. *IEEE Access*. (2024) 12:162608–21. doi: 10.1109/ACCESS.2024.3477321

11. Mohammad F, Al Ahmadi S, Al Muhtadi J. Blockchain-based deep CNN for brain tumor prediction using MRI scans. *Diagnostics*. (2023) 13:1229. doi: 10.3390/diagnostics1307122

12. Naeem AB, Senapati B, Chauhan AS, Makhija M, Singh A, Gupta M, et al. Hypothyroidism disease diagnosis by using machine learning algorithms. *Int J Intell Syst Appl Eng*. (2023) 11:368–73.

13. Rasool N, Wani NA, Bhat JI, Saharan S, Sharma VK, Alsulami BS, et al. CNN-TumorNet: leveraging explainability in deep learning for precise brain tumor diagnosis on MRI images. *Front Oncol*. (2025) 15:1554559. doi: 10.3389/fonc.2025.1554559

14. Rezaeijo SM, Chegeni N, Baghaei Naeini F, Makris D, Bakas S. Within-modality synthesis and novel Radiomic evaluation of brain MRI scans. *Cancer*. (2023) 15:3565. doi: 10.3390/cancers15143565

15. Khan SUR, Asif S, Bilal O, Rehman HU. Lead-CNN: lightweight enhanced dimension reduction convolutional neural network for brain tumor classification. *Int J Mach Learn Cyber*. (2025) 8:6. doi: 10.1007/s13042-025-02637-6

16. Aggarwal K, Kartikeya K, Srivastava V. Deep learning-driven CNN models for enhanced brain tumor classification In: TP Singh, CJ Kumar, A Abraham and KT Igulu, editors. Revolutionizing healthcare: impact of artificial intelligence on diagnosis, treatment, and patient care studies in computational intelligence. Cham: Springer (2025)

17. 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

18. Dintakurthi M., Biccavolu V. S. S., Srinivas M., Boge L., Sreeram G. D., (2025) "Bridging the Diagnostic Gap: Classification of MRI-Based Brain Tumors Using a CNN and Transformer-Based Hybrid Deep Learning Method," International conference on artificial intelligence and data engineering (AIDE), AIDE Nitte, India, 182–187.

19. Missaoui R, Hechkel W, Saadaoui W, Helali A, Leo M. Advanced deep learning and machine learning techniques for MRI brain tumor analysis: a review. *Sensors*. (2025) 25:2746. doi: 10.3390/s25092746

20. Rezk NG, Alshathri S, Sayed A, Hemdan EE-D, El-Behery H. Secure hybrid deep learning for MRI-based brain tumor detection in smart medical IoT systems. *Diagnostics*. (2025) 15:639. doi: 10.3390/diagnostics15050639

21. Anish JJ, Ajitha D. Exploring the state-of-the-art algorithms for brain tumor classification using MRI data. *IEEE Access*. (2025) 13:118033–54. doi: 10.1109/ACCESS.2025.3579727

22. Hasan M. Z., Tamim Abdullah, Asadujjaman D.M., Rahman Mahfujur, (2025) "A CNN approach to automated detection and classification of brain tumors," International conference on electrical, computer and communication engineering (ECCE), Chittagong, Bangladesh, ECCE. 1–6.

23. Taposh M. H., Abrar T G, Amit R, Mahfujur R, Annas MN, Rafeed R, "A lightweight CNN model for detecting brain tumors using MRI based image enhancement," 2025 4th international conference on robotics, electrical and signal processing techniques (ICREST), IEEE, Bangladesh, (2025), pp. 403–408.

24. Rangaraj KS, Sripathy SK, Swarnalatha P. Enhanced transfer learning and CNN approach for brain tumor detection In: RK Hamdan, editor. Sustainable data management studies in big data. *eds* ed. Cham: Springer (2025)

25. DM V., Fathima G., (2025). "Efficient medical image processing for tumour detection using hybrid CNN framework," 2025 International conference on inventive computation technologies (ICICT), IEEE, Nepal. 457–463.

26. Naeem AB, Senapati B, Bhuva D, Zaidi A, Bhuva AP, Islam Sudman MS, et al. Heart disease detection using feature extraction and artificial neural networks: a sensor-based approach. *IEEE Access*. (2024) 12:37349–62. doi: 10.1109/access.2024.3373646

27. Ansari MM, Kumar S, Heyat MBB, Ullah H, Bin Hayat MA, Sumbul PS, et al. SVMVGGNet-16: a novel machine and deep learning based approaches for lung Cancer detection using combined SVM and VGGNet-16. *Curr Med Imaging*. (2025) 21:e15734056348824. doi: 10.2174/0115734056348824241224100809

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 23 October 2025
DOI 10.3389/fmed.2025.1677234

Check for updates

OPEN ACCESS

EDITED BY
Pardeep Sangwan,
Maharaja Surajmal Institute of Technology, India

REVIEWED BY
Anantha Raman Rathinam,
Malla Reddy College of Engineering, India
Suhas A. Bhyratae,
Sahyadri College of Engineering and Management, India

*CORRESPONDENCE
T. R. Mahesh
✉ trmahesh.1978@gmail.com
Eid Albalawi
✉ ealbalawi@kfu.edu.sa

RECEIVED 31 July 2025
ACCEPTED 25 September 2025
PUBLISHED 23 October 2025

CITATION
Priyadharshini M, Murugesh V, Mahesh TR, Albalawi E, Saidani O and Algarni A (2025) QBrainNet: harnessing enhanced quantum intelligence for advanced brain stroke prediction from medical imaging. *Front. Med.* 12:1677234. doi: 10.3389/fmed.2025.1677234

COPYRIGHT
© 2025 Priyadharshini, Murugesh, Mahesh, Albalawi, Saidani and Algarni. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

QBrainNet: harnessing enhanced quantum intelligence for advanced brain stroke prediction from medical imaging

M. Priyadharshini[1], V. Murugesh[2], T. R. Mahesh[3]*, Eid Albalawi[4]*, Oumaima Saidani[5] and Ali Algarni[6,7]

[1]Department of Computer Science & Engineering, Faculty of Science and Technology (IcfaiTech), ICFAI Foundation for Higher Education, Hyderabad, India, [2]School of Computer Science, Coventry University Kazakhstan, Astana, Kazakhstan, [3]Department of Computer Science and Engineering, Faculty of Engineering and Technology, JAIN (Deemed-to-be University), Bengaluru, India, [4]Department of Computer Science, College of Computer Science and Information Technology, King Faisal University, Al Ahsa, Saudi Arabia, [5]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia, [6]Department of Informatics and Computer Systems, College of Computer Science, King Khalid University, Abha, Asir, Saudi Arabia, [7]Center for Artificial Intelligence King Khalid University, Abha, Asir, Saudi Arabia

Introduction: Brain stroke is still one of the leading causes of death and long-term disability in the world. Early and correct diagnosis is therefore important for patient outcome. Although Convolution Neural Network (CNN), classical machine learning models, have achieved great progress in medical image classification, they have to face the performance saturation problem when dealing with high-dimensional and complex data such as medical images. To tackle these limitations, we propose QBrainNet, a quantum enhanced model, which is to enhance brain stroke prediction from medical imaging datasets.
Methods: The model consists of Quantum Neural Networks (QNNs) applied as learning complex patterns in terms of medical images and Variational Quantum Circuits (VQCs) that will be used to optimize the classification. The feature extraction featured in the QNNs utilises quantum properties of superposition and entanglement to extract non-linear high-dimensional patterns in images related to stroke that may not be captured using classical limits. The VQCs, in turn, are applied to optimize the model performance, further allocating the boundaries of the decision and enhancing the model performance in terms of accuracy by optimizing the quantum gates and operators used during the work. QBrainNet utilizes the combination of such quantum properties as entanglement and superposition to represent more complicated non-linear patterns in stroke-specific images in a better manner than a classical application does.
Results: This paper proposes a hybrid classical-quantum scheme: preprocessing classically, and learning quantum-enhanced. Quantum gates and operators are used when performing the quantum phase to optimize decision boundaries, achieving vastly enhanced prediction accuracy and efficiency performance. Experimental results indicate that QBrainNet has a better accuracy (96%) and AUC-PR (0.97) than the classical models like CNN, SVM, and Random Forest, proving the superior performance of QBrainNet in stroke detection.
Discussion: The inference time is shorter, so the model can be used as a real-time clinical application. This article points to the possibilities quantum computing can have in revolutionizing medical diagnostics, especially stroke prediction.

KEYWORDS

brain stroke prediction, early stroke detection, medical imaging, quantum computing, quantum neural networks (QNN), quantum intelligence

1 Introduction

Stroke constitutes one of the significant causes of death and permanent disability in the world, with about 15 million individuals having a stroke per year according to the WHO (1). Early diagnosis and prompt treatment are essential in enhancing survival and minimizing long-term disability. Nevertheless, clinical condition diagnosis, where time is of the essence, will still be a challenge to correctly predict because of the complexity and subtlety of patterns in medical imaging data, particularly in the early stages (2, 3). Interpretation of CT and MRI scans used widely to detect stroke is subject to human error, inconsistency, and variability between practitioners, and it may lead to delay in diagnosis and impact treatment outcomes (4).

Recently, the methods based on machine learning (ML), particularly Convolutional Neural Networks (CNNs), have been actively applied to the medical image analysis, and stroke detection has been successful with the CT or MRI scans. The CNNs have been shown to work exceptionally well when processing medical imagery and extracting features that classify the image as stroke-related quickly, consistently, and accurately, compared to the more conventional methods (5, 6). Although these CNNs and other classical models are effective, they are limited by high-dimensional and complex medical data. These models fail to identify delicate structures and interactions within the data, particularly when the datasets are small and/or low-contrast, as frequently happens in medical imaging of stroke patients (7, 8).

The new area of Quantum Machine Learning (QML) offers an optimistic answer to these difficulties. Quantum systems work with information in radically new ways compared to classical systems, allowing them to work with extensive multi-dimensional data more efficiently through superposition and entanglement. Indeed, the quantum properties allow quantum computers to solve some problems efficiently in computation, where classical computers do not; the quantum potential advantage has indeed been observed in applications such as medical image analysis (9, 10). Quantum Neural Networks (QNNs) and Variational Quantum Circuits (VQCs) can specifically be used to provide an advantage in the classical world in specific tasks by finding complex patterns and relationships in data and using these patterns and traits in a non-linear fashion (11, 12).

This paper presents QBrainNet, a classical-quantum model that aims to enhance medical imaging stroke prediction. The classical element of the QBrainNet engages in feature extractions, augmenting images, and noise elimination, whereas the quantum element continuously applies QNNs and VQC networks to the learning task. QBrainNet, with its quantum-enhanced learning combining classical machine learning, is much faster and has a higher accuracy at identifying subtle factors in stroke-related medical images (13, 14). The quantum aspect of the model applies simulated quantum operations through Python code to optimally determine decision boundaries in the feature space. It is, therefore, more accurate in the classification than the conventional methods.

One main issue with medical image classification tasks is the small datasets. In our scenario, we only have 3,800 images, which can easily result in overfitting. However, the problem can be overcome the way QBrainNet does it by using cross-validation and regularization techniques (15, 16). The quantum elements of QBrainNet are designed through Python-based quantum simulation, in which quantum gates and circuits are simulated on a classical computing device. Thus, the model is accessible and reproducible without quantum information technology hardware (17, 18).

The main strengths of the QBrainNet model in comparison with classical approaches are linked to the possibility of dealing better with high-dimensional data. CNNs and other traditional techniques are bulky programs that handle big chunks of data, particularly in the case of medical image tests. Compared to this, QBrainNet takes advantage of quantum parallelism, where quantum gates and superposition significantly decrease the degree of computation and speed of processing (19). Such a decrease in computational demands and the increase in the prediction speed result in QBrainNet being a potential candidate in clinical practice, where the speed of diagnosis may be a matter of life and death.

In recent developments, quantum computing has demonstrated great potential to improve machine learning models, particularly for high-dimensional data analysis. In this work, we simulate the quantum parts of QBrainNet using PennyLane on classical computing resources. This way, we can exploit quantum effects like superposition and entanglement for feature extraction and optimization without access to real quantum hardware. Our simulation allows us to simulate quantum circuits and perform parameter optimization in a way compatible with classical machine learning.

The present study adds to the list of research that deals with the application of quantum computing in healthcare. In particular, we show promise of quantum-enhanced models such as QBrainNet in the field of stroke prediction, namely that quantum technology can be used to enhance the performance of medical diagnostics not only in accuracy, but in efficiency as well, especially in a domain where errors can have severe consequences like stroke care (19).

2 Related work

Applying machine learning (ML) to medical imaging has entirely transformed the face of healthcare diagnostics in a way no one had previously imagined. More specifically, CNNs have found a wide application in deep learning to solve specific tasks in medical imaging. The application of CNNs to the interpretation of medical images has been demonstrated to be capable of detecting and classifying ailments such as cancer, pneumonia, and brain stroke, as well as segmenting organs and other body parts critical to the human body (20). Of particular interest in brain stroke detection is that CNNs and other forms of deep learning have been applied to CT image processing, MRIs, and fMRI to provide brain stroke risk assessments, but with high levels of automation. Such models are much superior in the detection of stroke lesions and the classification of ischemic strokes. By extracting hierarchical representations of image information, these models can discover useful trends that the human expert may not be able to declare easily. The approach here is a novel application of the idea behind hybrid quantum-classical neural networks (21) to predicting strokes through quantum-enhanced preprocessing.

These models, although effective, are restricted. Brain images can be complex, leading to difficulties for classical CNNs to apply to them and subtle features in the early stages of strokes. These models require substantial labeled data, computer power, and a preprocessing mechanism (22), and thus are not readily applicable to high-dimensional data. Additionally, it is computationally costly to train deep learning models wherein the high-resolution medical images are to be used; they require both heavy computing hardware and time. Original CNNs inherently lack the flexibility to extract subtly non-linear structures in the data, and such patterns are typical with medical images, as the data are noisy, heterogeneous, and may be inaccurately annotated (23). Also, this fulfills the need for more complex models that could better predict the nature of medical imaging with a complex structure (24).

To overcome these shortcomings, Quantum Machine Learning (QML) has proposed itself as an excellent solution. It is theorized that QML methods will be able to utilize the quantum superposition and quantum entanglement properties of quantum computers to both process complex information more effectively and prevent the scale explosion that occurs when using classical models. These quantum benefits may bring computational advantage, especially where data is needed in very high dimensions, such as in medical image processing (25). Quantum systems offer the prospect of investigating multiple solutions in parallel and exhibit greater capabilities of pattern recognition, which are of particular interest with complicated medical data. This will enable quantum methods, even when implemented on classical platforms using Python code, to perform better when compared with classical models in specific tasks requiring subtle non-linear relationships, e.g., when used to predict stroke (26, 27).

Healthcare and medical diagnosis are some examples in which QML has already been proven effective. For instance, Quantum Support Vector Machines (QSVM) were used to solve tasks in image classification. The results revealed that QSVMs are more effective in terms of computational efficiency than SVMs and are highly accurate in prediction (28). Moreover, QNNs, or the quantum analog of normal neural networks, have already been used in such tasks as image classification and drug discovery. Quantum-enhanced models, on the other hand, can access the power of quantum entanglement to learn intricate structures in data that are favorable over conventional models in the task of image classification (29). As some examples, the Quantum version of standard neural networks, namely Quantum Neural Networks (QNNs), have been implemented in problems like image classification and drug recognition. In the light of this understanding, QE models can leverage quantum entanglement to learn complex patterns in the data in a more efficient way than classical models, which is a key advantage in various tasks, such as image classification. Such methods are currently being utilized in this work as simulated quantum operations that, even though they do not run on actual quantum devices, act as a step in the right direction as applied to quantum-enhanced optimization.

Other quantum algorithms are likely to prove useful in healthcare, including Quantum Random Forests (QRF) and Quantum k-Nearest Neighbors (QK-NN), which have been found in many cases to require less time to train and achieve higher accuracy than their classical counterparts on high-dimensional data (30, 31). Quantum algorithms, including Quantum Random Forests (QRF) and Quantum k-Nearest Neighbors (QK-NN), have also been investigated in healthcare and on high-dimensional data. Quantum algorithms are more efficient in their training speed, and their results are found to be better when compared to classical algorithms. Such algorithms are emulated via quantum operations on a classical computer in Python and demonstrate the possibilities of the quantum-enhanced models without involving the actual physical quantum device (27).

Although applying QML to medical imaging is gaining more attention, it has not yet been explored in brain stroke prediction. Although past works have used quantum models in image segmentation, disease categorization, and other medical imaging applications, there has yet to be a quantum learning model to predict stroke occurrence using medical imagery, which is the novelty of this paper. A quickly expanding volume of literature on QML shows that one of its uses can be better optimization, image classification, and pattern recognition. Still, using QML in stroke prediction in medical imaging has yet to be explored (32). Though numerous cases of research on QML exist, there is a significant lacuna in its application in the prediction of brain stroke, which is the novelty of this work. Though quantum-enhanced models have already demonstrated their potential in optimization, image classification, and pattern-recognition problems, their use in medical imaging, in general, and stroke prediction, in particular, has not been studied extensively. This work bridges this gap through simulated quantum operations (through Python code) on classical computing resources (33).

The novelty of this research is that QBrainNet is the first application of QML in stroke prediction. The architecture can close a substantial research gap in stroke detection research as it has integrated quantum-enhanced preprocessing, feature extraction, and classification into a single framework. Classical simulations of quantum operations allow for avoiding quantum hardware, but increase the stroke prediction accuracy and reduce computing costs (34). The proposed work is the initial implementation of QML regarding stroke expectations. Quantum-based benefits to preprocessing, feature extraction, and classification strongly occur within the same framework, as all other quantum manipulations are performed through Python codes running on a classical CPU. Employing simulated quantum operations over quantum hardware indicates a big leap toward actualizing quantum-powered healthcare tools. It influences how quantum computing can be used to develop solutions to mitigate modern medicine's challenge to the detriment of the overall healthcare industry: stroke diagnosis (35).

3 Methodology

This section describes the general strategy used to get to and test QBrainNet, a quantum augmented neural network that will predict the risk of stroke from brain imaging data. It contains four main parts of methodology that are dataset preparation, preprocessing and feature extraction, quantum machine learning model development and model training and evaluation. We describe each stage in detail to provide a detailed account of how the quantum techniques are integrated into the medical image analysis pipeline for increasing the accuracy of stroke prediction.

The system requirements for running the quantum operation simulations are as follows: The simulations have been run on a system that has Intel i7 processor and 16 GB RAM the Ubuntu 20.04 operating system. The quantum operations were simulated with PennyLane, version 0.18.0, a Python-based library which can build on classical computing resources to simulate quantum operations. The simulate codes were written in Python 3.8 and some additional libraries such as Numpy 1.21.0 for numerical computing, Scipy 1.7.0 for scientific computing, matplotlib 3.4.3 for visualization. The entire setup was done in a conda environment to handle everything in the appropriate way in terms of dependencies and reproducibility. This environment allowed efficient implementation of

quantum simulations on classical computing resources without the need for any actual quantum hardware.

3.1 Dataset

The medical images included in this study were diagnosed as usual or as stroke from a dataset. The photos are taken from publicly available datasets usually used in the stroke detection area, such as CT scans and MRI images. This dataset contains high-resolution MRI brain scans of different stroke severity, early ischemia, and late-stage hemorrhage. The pictures are marked to help define which ones are routine and which have an indication of a stroke. These images are then fed through simulated quantum operations to improve feature extraction, classification, and overall predictive accuracy with Python-based quantum simulators on classical computing resources. Lastly, each image has a label, indicating whether the brain imaging is standard or if there is a stroke.

Figure 1 demonstrates the unprocessed and processed CT scan brain scans. Raw images are initially scanned, whereas the processed ones have undergone a procedure of removing noise and normalization to facilitate analysis. Figure 2 shows grayscale, equalized, and edge-detected images of the preprocessed brain images. Gray levels eliminate color, equalization increases contrast, and edge detection emphasizes boundaries of key structures. The CT scan cross-sections shown in Figure 3 are used to

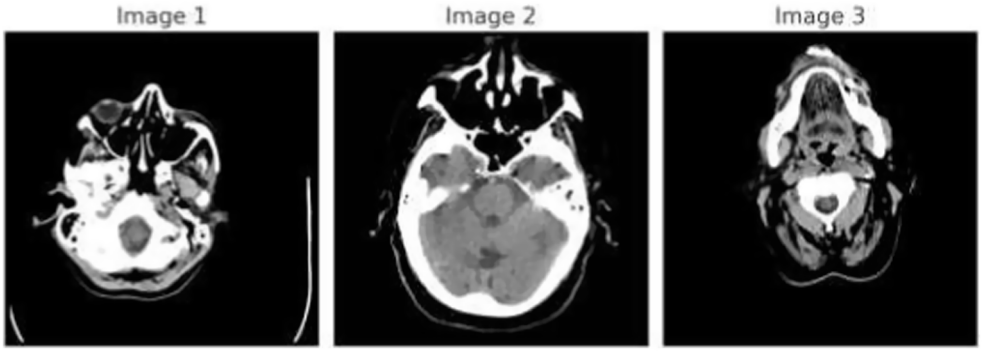


FIGURE 1
Dataset overview: raw and processed brain CT scan images.

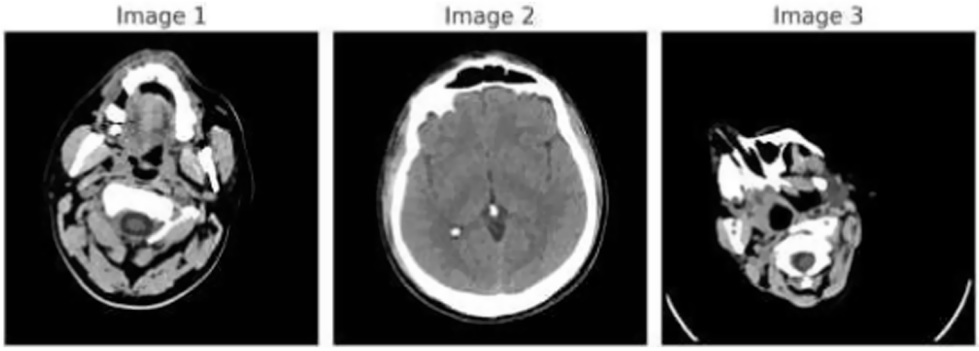


FIGURE 2
Preprocessed brain images: grayscale, equalized, and edge-detected versions.

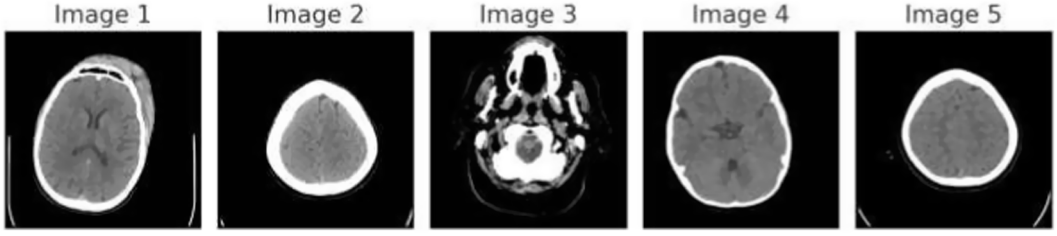


FIGURE 3
CT scan cross-sections showing brain structure and potential abnormalities.

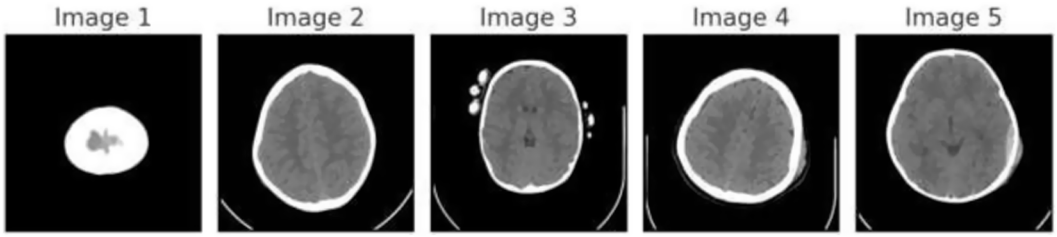


FIGURE 4
CT scan cross-sections of brain showing stroke variants.

obtain details about brain structure and the parts prone to abnormalities such as strokes and tumors. Figure 4 shows different CT scan cross-sections with varying types of stroke, and how ischemic and hemorrhagic strokes can be represented in the brain in a cross-section.

3.2 Data preprocessing

The raw medical images are preprocessed before training and evaluation to reduce inconsistency and robustness across the medical image set. Rotation, flip, and noise addition augment the dataset and make it more diverse. To resemble real data and increase the model robustness to imperfect data, these procedures simulate real-world variation, e.g., to some extent, by the slight changes in rotation or orientation of scan images, and provide noise. This can better generalize the model, especially with a small data set, as it minimizes the chances of overfitting.

The primary preprocessing steps include:

1 **Image Resizing**: Uniformity is guaranteed in the input data, as all the images in medical images may have different resolutions. They are all resized to a fixed resolution. This is an essential step so that the data maintained between multiple images is compatible with deep learning models image resizing is computed using Equation 1.

$$I_{resized} = Resize\left(I_{original}, W, h\right) \quad (1)$$

Where:

- $I_{resized}$ - resized image.
- $I_{original}$ - original image.
- W & h are the target width and height, respectively.

2 **Normalization**: To adjust to the different pixel intensity values represented by various medical imaging modalities, the images are scaled to the 0–1 range. This will allow the model to be adjusted only to the scale of the raw data and not be distorted by the ranges of pixel intensities normalization is computed using Equation 2.

$$I_{normalized} = \frac{I_{original}}{255} \quad (2)$$

Where:

- $I_{normalized}$ - normalized image.
- $I_{original}$ - original pixel intensity.

3 **Class Imbalance Check**: Since the medical datasets usually become class imbalanced, balancing the number of samples in training and test sets within normal and stroke groups is very important. If an imbalance is discovered, methods that include over-sampling the minority observations or under-sampling the majority can be used to generate a balanced dataset. This eliminates the possibility of biasing the model toward one of the classes, which is used a lot more; hence, the model will perform well in both classes.

3.3 Dataset partitioning

The data is split into the training data and a testing data where 70–80 percent of the data is used in the training and 20–30 percent for testing. The training data is then trained on the model, known as QBrainNet model and the testing data is used to estimate the model's performance on unknown data. This division will ensure the model is tested on data that it has not encountered previously during the model's training, and will be an impartial representation of how well the model is performing.

Preprocessing of dataset, and splitting the preprocessed dataset into training and testing datasets is done. The model is trained on the training data and tested on the test data (36, 37). The training is usually done using 70–80% of the data; the remaining 20–30% is used for testing. There is a need to fold this type to make sure that the model performs well on the unseen data rather than being too optimistic regarding the performance.

3.3.1 Dataset distribution

The distribution of 'normal' and 'stroke' images over training and test sets can be viewed in Table 1.

Class distribution plays a role in training the model on a balanced set of examples, which is very important for accurate stroke prediction.

TABLE 1 Distribution of normal and stroke images in the dataset.

Class	Training set (images)	Test set (images)	Total images	% in training set	% in test set	Augmentation applied	Primary data source
Normal	1,500	500	2,000	51.70%	55.60%	Rotation, Flip, Noise	Hospital A & Public Dataset
Stroke	1,400	400	1,800	48.30%	44.40%	Contrast Stretching, Zoom	Hospital B & Research Cohort
Total	2,900	900	3,800	100%	100%	–	–

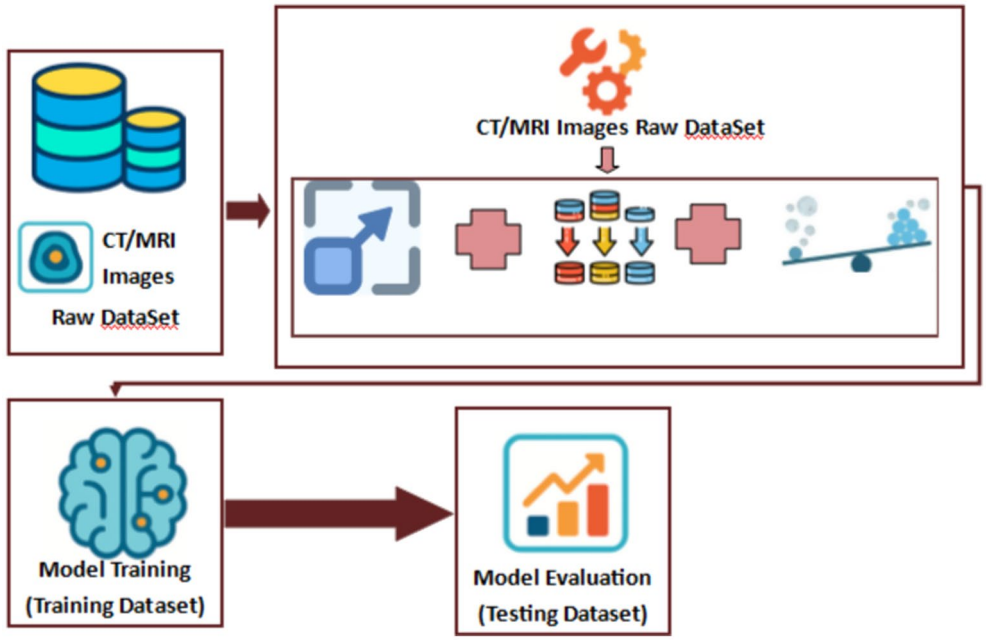


FIGURE 5
Proposed model's dataset flow diagram.

3.3.2 Dataset flow diagram

Here, in the following Figure 5, we show the flow of the dataset in the preprocessing, training and evaluation stages:

3.3.3 Class imbalance handling

To solve the problem of class imbalance of the dataset, we used some oversampling and undersampling methods during the data preprocessing phase:

1 **Oversampling**: We applied Random Oversampling to replicate samples from the minority class (either "normal" or "stroke") to train the model on a balanced dataset. This method copies minority class samples to make the sizes of the minority and majority classes equal, eliminating the model's bias for the majority class.
   - o **Stage in Pipeline**: Random Oversampling was used as one of the pipeline steps on the training set after splitting the dataset into training/validation sets. This helped ensure the model would learn from an even distribution of the two classes.

2 **Undersampling**: Since it is a class imbalance problem, we applied the Random Undersampling technique to the majority class. This method addresses the issue by randomly selecting samples from the majority class to obtain a balanced distribution between both classes. Decreasing the number of majority class samples ensures the model does not become biased toward majority class predictions.
   - o **Stage in Pipeline**: Minority class was oversampled, and then Random Undersampling was implemented to achieve class balance without overfitting of the minority.

Class Imbalance Handling Pipeline:

1 Divide the dataset into a training and validation dataset.
2 Implement Random Oversampling to the minority class in the training dataset to balance the class distribution.
3 Random Undersampling: the oversized majority class in the train data set is reduced to the size of the minority class.
4 The balanced training set is now used to train the QBrainNet model.

These techniques allow for equal representation of both classes (regular versus stroke) during model training, which is essential in healthcare applications where accurate classification of both conditions is crucial.

3.4 Preprocessing and feature extraction

Several classical preprocessing techniques are performed before the quantum machine learning algorithms are used to preprocess the medical images, such that the data is in a format that is as best as possible for extracting features and the model can be trained on. These techniques allow us to mitigate noise, clean, increase contrast, and standardize the stroke dataset to facilitate the networks' detection of stroke features more easily (38).

3.4.1 Image resizing

Resizing images is a crucial preprocessing step because all the images need consistent dimensions supported by deep learning models, which usually need uniform input sizes. The resizing process involves mapping the original image size $W_{original} \times h_{original}$ to a new size $w_{new} \times h_{new}$. This can be mathematically represented as using Equation 3:

$$I_{resized}(x,y) = I_{original}\left(\frac{x}{w_{original}} \cdot w_{new} = \frac{y}{h_{original}} \cdot h_{new}\right) \quad (3)$$

Where:

- $I_{resized}$ - resized image.
- $I_{original}$ - original image.
- $W_{original}$ and $h_{original}$ are the original width & height of the image.
- W_{new} and h_{new} are the target width & height for resizing?

The bilinear interpolation method is used for resizing to preserve image details (39).

3.4.2 Grayscale conversion

Grayscale conversion of the images is applied to simplify the data and decrease computational complexity while retaining stroke-related features. Grayscale images are beneficial as they decrease the number of channels (from 3 in RGB to 1), thus reducing the amount of computation and emphasizing the textural differences in the brain tissue.

The conversion from a color image $I_{rgb}(x,y)$ to grayscale $I_{gray}(x,y)$ is done by averaging the weighted sum of the RGB channels, following the formula as shown in Equation 4:

$$I_{gray}(x,y) = 0.2989 \cdot I^{R}_{rgb}(x,y) + 0.5870 \cdot I^{G}_{rgb}(x,y) + 0.1140 \cdot I^{B}_{rgb}(x,y) \quad (4)$$

Where:

$I^{R}_{rgb}(x,y), I^{G}_{rgb}(x,y), I^{B}_{rgb}(x,y)$ - Represent the Red, Green, and Blue (RGB) color channels, respectively.

$I_{gray}(x,y)$ - resulting grayscale image.

3.4.3 Histogram equalization

To enhance the contrast of the images, histogram equalization is used to redistribute the intensity levels throughout the image. Spread out across the whole range, this process helps to bring out subtle details, including early signs of stroke. Histogram equalization can be mathematically formulated as shown in Equations 5 and 6:

$$CDF(i) = \sum_{j=0}^{i} p(j) \quad (5)$$

$$I_{eq}(x,y) = CDF\left(I_{original}(x,y)\right) \cdot (L-1) \quad (6)$$

Where:

$CDF(i)$ It is the cumulative distribution function of the pixel intensities.

$p(j)$ It is the probability density function of the pixel intensities.

L Is the number of possible intensity levels (typically 256 for 8-bit images).

$I_{eg}(x,y)$ It is the histogram-equalized image.

It ensures that the pixel intensity distribution is more uniform than it is, thereby improving the contrast of the image and bringing out finer details, which are important for stroke detection (40).

3.4.4 Feature extraction

Next, necessary characteristics from the images are captured using feature extraction. Key features are extracted using classical methods, including those based on determining edges or analyzing textures, with the view that these can be used to differentiate stroke-affected areas from normal brain tissue.

1 **Edge Detection:** This involves the detection of the boundaries of an object in an image. The Canny Edge Detection algorithm is employed to indicate regions of interest, such as in stroke lesions, by identifying sharp intensity transitions. Mathematically, edge detection is defined as shown in Equation 7:

$$EDGE(I_{gray}) = Canny(I_{gray}) \quad (7)$$

Where I_{gray} is the grayscale image, and the Canny operator finds the edges by computing the gradient of the image.

2 **Texture Analysis**: It measures the structure present in the image by performing texture analysis. Gray Level Co-occurrence Matrix (GLCM) is computed using Equation 8:

$$GLCM(i,j) = \sum_{x,y} p(x,y,i,j) \quad (8)$$

Where:

- Lastly, GLCM(i, j) denotes the co-occurrence matrix where pixels have values i and j.
- It is noted that p(x,y,i,j) defines the probability that pixel pair values are i and j at locations x and y.

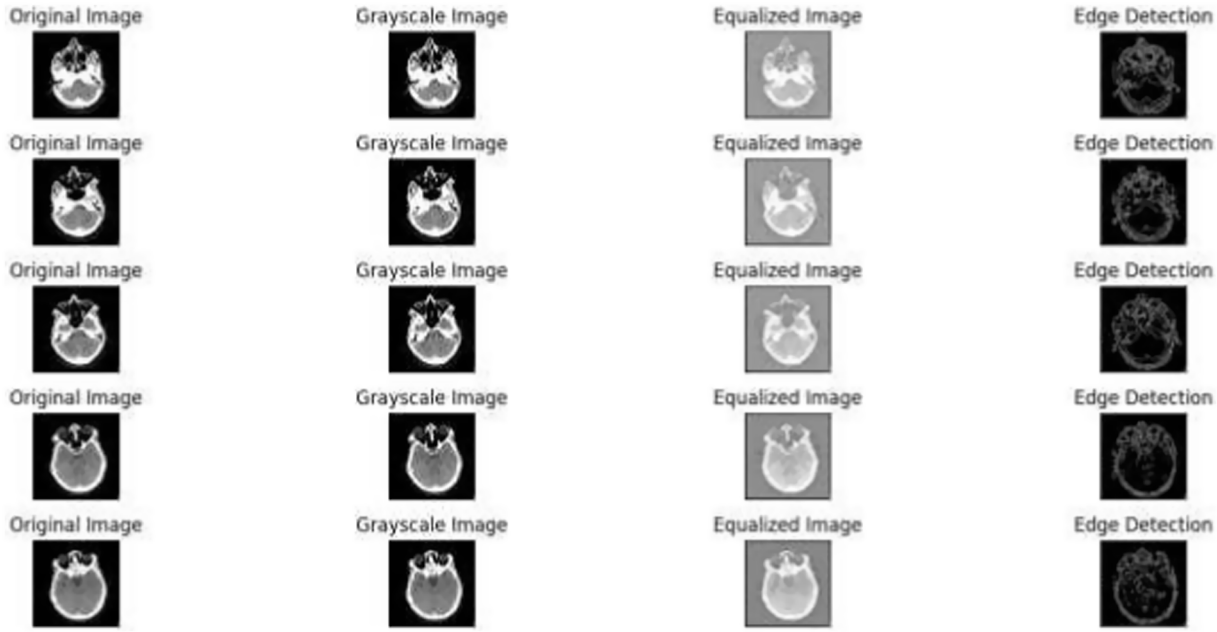


FIGURE 6
Preprocessed images: original, grayscale, equalized, and edge detection.

The texture features are promised as a crucial source of information about the texture of brain tissue, which might aid in discriminating between healthy and stroke-affected parts (41, 42).

Figure 6 illustrates the effects of the preprocessing steps on the original medical image. The left image is the raw medical scan, the center image is the conversion to a greyscale and the last one is the histogram equalization (42). Figure 6 depicts a sample of medical images after the grayscale conversion and histogram equalization.

The computational preprocessing step uses quantum-enhanced feature extraction procedures, which are also simulated using Python scripts in PennyLane and other quantum simulators. The methods enable the detection of fragile patterns in medical images that conventional methods such as CNNs may not easily learn. By mapping quantum processes onto classical computers, we can use quantum phenomena such as superposition and entanglement to use the more efficient extraction of features in complex and high-dimensional medical images.

3.5 Quantum machine learning model

This part introduces the derivation of this work's QBrainNet model, which is a quantum-enhanced neural network for estimating the probability of missing a stroke case from brain images. The model combines classical machine learning methods with simulated quantum models for a more accurate stroke prediction. Rather than using physical quantum hardware, the quantum constituents are simulated through the PennyLane simulator implemented in Python and run on ordinary computing resources. These simulations allow us to incorporate quantum-inspired properties like superposition and entanglement, which are challenging to simulate in purely classical neural networks. In our hybrid framework, we train variational quantum circuits (VQCs) with PennyLane to simulate them, and solve for the quantum parameters by gradient descent to improve prediction accuracy.

The QBrainNet architecture comprises several layers, each taking advantage of quantum-enhanced processing to enhance the processing and analysis of the medical images. In particular, the quantum layers attractively model the quantum operations to transform the image data into feature vectors with information on more complex patterns than classical techniques. These feature vectors are then fed to a conventional neural network for the final stroke prediction. This can mimic the advantages of a quantum computer on regular computers, enabling more of us to take advantage of the quantum advantages and do it more efficiently.

The model (QBrainNet) involves quantum enhanced ways to improve the accuracy of stroke forecast. This is a hybrid model, which combines the classical neural network architecture and simulates the quantum operations to process and analyze medical images more effectively. Rather than operating on real quantum hardware, however, quantum phenomena, such as superposition and entanglement, are simulated in Python libraries in the actual hardware. This will enable the model to reflect better, more intricate relationships in the data, which is a benefit over conventional machine learning.

The model training for the QBrainNet has been performed for 50 epochs, using gradient-based optimization to update the quantum parameters (RZ gate angles) in the variational quantum circuits, which are implemented in PennyLane. The Adam optimizer with a learning rate of 0.001 was used as the optimizer for training. The model showed a progressive improvement in accuracy for the first 30–40 epochs, and then the loss function stabilized, which means that the quantum parts converged to the local minimum. The arrival time of the quantum components was tracked closely, and the convergence was relatively poor after epoch 40.

The two main components of the QBrainNet model are created to handle the two various sections of the image data processing pipeline.

Quantum Circuit Architecture:

The quantum circuit of QBrainNet model is a combination of 3 variational layers, each of which comprises a series of quantum gates performed to process the input data and achieve the maximum decision boundaries. The type of gates employed in each layer is as follows:

- Hadamard (H) gate on qubit 1.
- CNOT gate between qubit 1 and qubit 2.
- Z-Rotation (RZ) gate on qubit 3.

This circuit is simulated in PennyLane using classical computer resources. Each variational layer automatically maps the input data and develops the decision boundaries for better classification accuracy.

The total trainable parameters of the quantum circuit are 12, which corresponds to the angles of the RZ gates in each variational layer. These parameters are then optimized by gradient-based methods during training to minimize the loss and improve classification performance.

The measurement scheme measures the quantum state on a Pauli Z basis at the end of each variational layer. The classical bits generated from this measurement are combined to create the classification output. The outcome depends on a majority vote among all the qubits in the system.

The quantum circuit shown above is used to train the QBrainNet model. The pseudocode for the training process is shown below. *#Initialize quantum circuit with 4 qubits. initialize_quantum_circuit(num_qubits = 4). #Define variational layers (3 layers). for layer in range(3): #Apply Hadamard gate on qubit 0. apply_Hadamard_gate(qubit = 0). #Apply Controlled-NOT gate between qubits 0 and 1. apply_CNOT_gate(control_qubit = 0, target_qubit = 1). #Apply Z-Rotation gate on qubit 2. apply_RZ_gate(qubit = 2). #Initialize classical optimizer* (e.g., *Adam optimizer). optimizer = AdamOptimizer(learning_rate = 0.001). #Training loop for 50 epochs. for epoch in range(50): #Apply quantum circuit (forward pass). quantum_output = apply_quantum_circuit(inputs). #Measure quantum state in Pauli Z basis. classical_output = measure(quantum_output, basis = 'Z'). #Compute the loss function. loss = compute_loss(classical_output, ground_truth). #Calculate the gradient of the loss. gradient = compute_gradient(loss). #Update quantum parameters using the optimizer. optimizer.update_parameters(gradient). #Final output: make the classification decision. final_output = classify_output(classical_output).*

3.5.1 Classical feature extraction

Earlier, we mentioned about the extraction of relevant features from the preprocessed medical images using classical methods such as edge detection and texture analysis. The next stage is supplied with a compact representation of brain images for subsequent processing by these features (43).

This part shows the derivation of a quantum-enhanced neural network, or QBrainNet that can estimate the probability of missing a stroke case given a brain image. The model is a combination of classical machine learning techniques and quantum simulation operations that will improve stroke prediction accuracy. In lieu of making use of practical quantum hardware, quantum emulations are made with quantum simulators PennyLane utilizing Python on conventional, classical computing facilities. These quantum simulations allow us to use the properties of quantum-like superposition and entanglement that are difficult to use with classical neural networks.

The architecture of the QBrainNet consists of several layers, where each layer utilizes the quantum processing capability to boost the processing and analysis of the medical images. In particular, the quantum layers model the quantum operations attractively to transform the image data into feature vectors with information on more complex patterns than classical techniques. These feature vectors are then fed in a conventional neural network for final stroke prediction. The volume and diversity of medical images are also relatively low, and thus can create overfitting and decrease the generalization of the models in stroke detection. To resolve this, we used several image augmentation methods - rotation, flipping, and adding noise to the data - before sending them forward in the preprocessing stage to improve and stabilize the generalization ability of QBrainNet. Rotations were applied to mimic various positions of the medical scans to ensure that the model can identify the patterns associated with stroke, independent of the direction at which the images are taken. This is especially significant as brain scans used in medical practice may differ in orientation. Manipulation of the model by flipping it horizontally and vertically to introduce the model to other perspectives, which is more likely to generalize its operative features in different variable conditions. Lastly, we introduced noise into the pictures to simulate the inevitable flaws associated with real-world medical imaging, including scanner artifacts or low resolution. The model learns to generalize on the essential features of the data rather than memorizing noise-free, idealized images by adding noise to the data. The combination of the above augmentation strategies increases the whole dataset's variety, enabling QBrainNet to pick up on more of the possible patterns and achieve a lower probability of overfitting, especially with such a relatively small amount of data. That makes a model more competent to work with unseen data and supply precise estimation in clinical practice.

It entails studying image patterns, such as boundaries, textures, and shapes. Edge detection with the Canny operator and GLCM is applied to extract the features such as these. The features extracted from these data can be represented mathematically as follows:

1 **Edge Detection**: Using the **Canny Edge Detection** algorithm, the boundary information E_{edges} for a given image $I_{grayscale}$ is obtained using Equation 9:

$$E_{edges} = Canny\left(I_{grayscale}\right) \tag{9}$$

Where:
$I_{grayscale}$ It is a grayscale image.
E_{edges} Represents the edges detected in the image.

Texture Features: The GLCM (Gray Level Co-occurrence Matrix) is an algorithm employed to describe the texture patterns present in the image, and is able to capture important statistics such as contrast, energy, and correlation. The GLCM for a grayscale image $I_{grayscale}$ is computed using Equation 9.

Consequently, these classical features are then passed through to the quantum-enhanced stage, where they are processed and further optimized.

To solve the generalizability problem and improve the overfitting level, we used image augmentation methods, including rotating, flipping, and adding noise. Such techniques mimic the natural variation in medical images and therefore aid in better generalization of the model in cases where the data is small.

3.5.2 Quantum enhancement

After the extraction, we feed the extracted features to the Quantum Neural Network (QNN) to produce classification outputs. Dynamical correlations of the quantum model such as superposition and entanglement make it possible for it to model complex patterns of the data which cannot be easily observed with the classical model alone (44). In order to learn the decision boundaries and find higher-order relationships in the data, the quantum neural network is learned using Variational Quantum Circuits (VQCs) (45).

The model of QBrainNet integrates quantum-enhanced machine learning on the basis of quantum neural networks (QNNs) and variational quantum circuits (VQCs). PennyLane uses classical computing resources to simulate these quantum components. In this way, it is possible to do feature extraction and optimization with quantum phenomena such as superposition and entanglement without having access to actual quantum hardware. The quantum operations are simulated completely in the classical environment, meaning that the full power of quantum computing is utilized for an improved performance without losing a practical implementation on the existing computing resources.

As part of the classical layer of QBrainNet, we applied Adam with a learning rate of 0.001. Adam is effective in substantial learning tasks because of its adaptive learning rates and the momentum, making it converge and avoid over-fitting quicker.

Regarding the quantum portion, the Variational Quantum Circuits (VQCs) were trained with a gradient-based optimizer and the quantum gradient descent. A parameter optimization on the quantum circuit parameters would minimize the loss by updating parameters during each iteration through classical optimization algorithms such as Adam or L-BFGS. Such a hybrid optimization will allow efficient training and better ability in modeling complex patterns with medical images.

The basic idea of a Quantum Neural Network (QNN) is to use quantum circuits as the weights and transformations of the network, represented by the quantum gates (46). The input sample value is initialized and transformed according to the input data by utilizing quantum superposition, exploring various possible results simultaneously.

To optimize the weights of the quantum neural network, we use a Variational Quantum Circuit (VQC) that combines classical optimization (what is to be optimized) with quantum circuits (how optimization is to be performed). Here is the definition of VQC as shown in Equation 10.

$$|\psi(\theta)\rangle = U(\theta)|\psi_0\rangle \tag{10}$$

Where:

- $|\psi(\theta)$ is the quantum state after applying the quantum gates $U(\theta)$ with parameters θ.
- $|\psi_0$ is the initial quantum state.
- $U(\theta)$ is the unitary operator that applies quantum gates parameterized by θ.

The quantum circuit is also optimized in the classical-quantum hybrid approach by minimizing the loss function in terms of quantum gradient descent. The loss function can be expressed as shown in Equation 11:

$$L(\theta) = loss(|\psi(\theta)) \tag{11}$$

Where:

- A loss evaluates the prediction error of a quantum model (e.g., mean square error, cross-entropy).
- The loss function that the quantum circuit minimizes during optimization is L(θ).

Optimization of quantum circuit parameters is done with classical gradient descent and more complicated optimization algorithms (Adam or LBFGS). For training classical CNN model we used adaptive moment optimization algorithm (Adam). We have set its learning rate to equal 0.001 which resolves the loss function more quickly than randomized algorithms and prevents over-fitting. In the quantum part, we used an optimizer which is based on a gradient which we used to change the quantum gates in the variational quantum circuit (VQC) where in a similar manner we backpropagated through the quantum layers and optimized the decision boundaries.

3.5.3 Bridging the classical-quantum framework

The two parts work together to form a fusion classical quantum framework in which the quantum circuit combines the classical feature extraction model into a QBrainNet model. This approach's advantage is its use of both classical and quantum computing.

- Featuring high dimensional data with the classical methods
- It fed these features into the quantum circuit to determine how to process them, optimize decision boundaries and find complex patterns that classical methods may miss.

The high-dimensional data is handled by the classical model, while the quantum model exploits the data in parallel in a potentially more computationally efficient and more accurate prediction manner.

The quantum translation model QBrainNet is constructed as a hybrid classical-quantum framework by making the quantum circuit a part of the classical feature extraction model. Then, we utilize a quantum gradient algorithm (47) to optimize the parameters of the quantum circuit by adjusting the parameters of the circuit after each prediction according to the error. This hybrid method combines the good of classical and quantum computing, with one better with fine-scale methodology in high-dimension data and the other enhancing prediction accuracy in time series prediction problems (48).

In Figure 7, we see the hybrid classical-quantum framework in QBrainNet, built upon classical feature extraction and acting as an input to a quantum neural network for stroke prediction (Figure 7: Hybrid Classical-Quantum Framework shows the flow from classical feature extraction to quantum processing).

3.5.4 Algorithmic design of QBrainNet

1 **Initialize system:**

a Load preprocessed brain CT scan dataset.

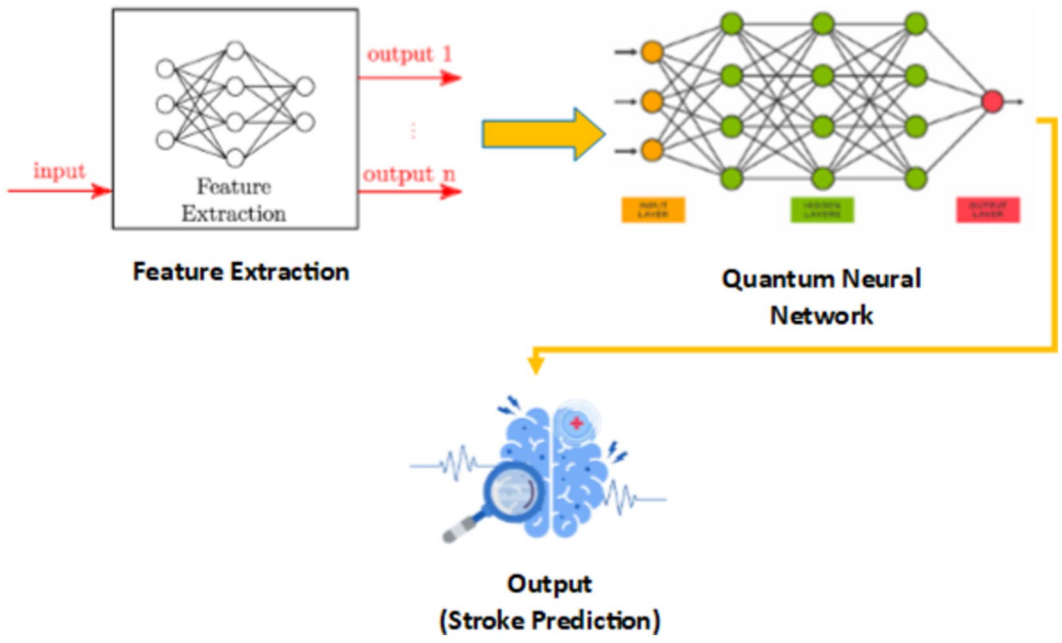


FIGURE 7
Hybrid classical-quantum framework of QBrainNet.

b Split dataset into training and testing sets (e.g., 80% training, 20% testing).
c Initialize classical CNN and quantum components (QNN with VQC).

2 **Preprocessing:**
a Convert CT scan images to grayscale.
b Apply image equalization to enhance contrast.
c Perform edge detection using the Canny operator.
d Apply augmentation techniques (rotation, flipping, noise addition).
e Normalize image data.

3 **Feature Extraction (Classical Component):**
a Extract features using classical methods:
o Edge detection.
o Texture analysis (GLCM).
b Store extracted features for quantum-enhanced processing.

4 **Quantum Enhancement (Quantum Component):**
a Feed extracted features into quantum neural network (QNN) using Variational Quantum Circuits (VQC).
b Apply quantum operations (superposition, entanglement) to extract complex patterns.
c Use quantum gates and VQC to adjust decision boundaries and find higher-order relationships.

5 **Model Training:**
a Train classical CNN model on extracted features using Adam optimizer (learning rate: 0.001).
b Optimize quantum circuit parameters using gradient descent and quantum gradient descent (with Adam or L-BFGS for fine-tuning).
c Minimize the loss function (cross-entropy or mean squared error).

6 **Evaluation:**
a Test the model on the testing dataset.
b Calculate performance metrics:

- Accuracy.
- Precision.
- F1 Score.
- Recall.
- AUC-PR.

o **Post-processing:**
a Generate predictions for unseen CT scan images.
b Display results and analyze model performance.

8 **Output:**
a Report stroke prediction results with confidence scores.
b Compare QBrainNet's performance with classical models (CNN, SVM, etc.)

3.5.5 Simulated quantum operations

The quantum component of QBrainNet was simulated on the classical hardware using the PennyLane library, the current quantum software platform where quantum circuit simulation is available on classical hardware. This was the selected approach because of the scarcity of quantum hardware and the requirement to provide fast experimentation on the quantum neural networks. Though quantum circuits have been simulated on the classical resources, PennyLane supports quantum gates

like Hadamard, CNOT and Z-Rotation gates to simulate, and it is an efficient way to explore the quantum-amplified potentials of the network.

3.5.5.1 Implications for scalability and feasibility

It is not so easy to simulate a quantum circuit on classical hardware. Scalability of simulations stands out by far, where the amount of computational resources needed to execute the simulation circuit rises exponentially with the qubit count in the circuit. An example is that with a quantum system with 50 or more qubits, it is just too costly to simulate on classical hardware because of memory and processing resources. With improvement of quantum hardware, quantum networks will exit classical simulation and transition to the quantum processors.

From a practical point of view, using classical hardware implies that the model can be tested and optimized now, before being able to have access to powerful enough quantum computers. Current quantum computing technology is in its early stages, and there are only a few quantum computers available through cloud services, and they are generally constrained in the number of qubits they can process. As quantum processors become available, the quantum parts of QBrainNet will be compiled to actual quantum hardware allowing the system to fully exploit quantum parallelism and superposition for more efficient processing.

In spite of these, the hybrid classical-quantum method used by QBrainNet can be seen as a very promising path ahead. It allows one to extract features with the help of quantum computing and simultaneously exploit the comparatively computationally efficient, everywhere-available classical optimization methods.

3.5.5.2 Mathematical formulation

$$|\psi(0)\rangle = |0\rangle \otimes |0\rangle \otimes |0\rangle \otimes |0\rangle \tag{12}$$

$$H|0\rangle = \frac{1}{\sqrt{2}}\left(|0\rangle + |1\rangle\right) \tag{13}$$

$$H|1\rangle = \frac{1}{\sqrt{2}}\left(|0\rangle + |1\rangle\right) \tag{14}$$

$$|\psi(1)\rangle = H \otimes I \otimes I \otimes I\,|0\rangle \tag{15}$$

$$\begin{aligned} &\mathrm{CNOT}|00\rangle = |00\rangle, \quad \mathrm{CNOT}|01\rangle = |01\rangle \\ &\mathrm{CNOT}|10\rangle = |11\rangle, \quad \mathrm{CNOT}|11\rangle = |10\rangle \end{aligned} \tag{16}$$

$$|\psi_2\rangle = CNOT\left(|\psi_1\rangle\right) \tag{17}$$

$$\mathrm{RZ}(\theta)|0\rangle = |0\rangle, \quad \mathrm{RZ}(\theta)|1\rangle = e^{i\theta}|1\rangle \tag{18}$$

$$|\psi_3\rangle = RZ(\theta) \otimes I \otimes I\,|\psi_2\rangle \tag{19}$$

$$Z|0\rangle = |0\rangle, \quad Z|1\rangle = -|1\rangle \tag{20}$$

$$|\psi_{measured}\rangle = \begin{cases} |0\rangle & \text{with probability } |\langle 0|\psi\rangle|^2 \\ |1\rangle & \text{with probability } |\langle 1|\psi\rangle|^2 \end{cases} \tag{21}$$

$$\nabla_\theta L = \frac{\partial L}{\partial \theta} \tag{22}$$

$$\theta_{new} = \theta_{old} - \eta \nabla_\theta L \tag{23}$$

$$\hat{y} = Classifier \tag{24}$$

The different mathematical formulation are shown from Equations 12 and 24. In the quantum-enhanced model developed for brain stroke prediction, the quantum circuit is initialized with 4 qubits each in the ground state $|0>$ which is normally used as an initialization for quantum computations. These qubits are the basic units that store the data and the quantum operations are implemented one after another, to manipulate the states of the qubits and extract the intricate patterns that might be difficult to use classical methods. The first gate performed on the qubits is the Hadamard gate H which is applied to qubit 0 to put it in a superposition between the states $|0>$ and $|1>$. This superposition enables the quantum system to investigate various states at the same time, which significantly increases the processing and representation of the complex data by the model. However, a Controlled-NOT (CNOT) gate is then applied between qubits 0 (control) and 1 (target) following the Hadamard gate and then these two qubits are entangled with each other, generating a correlation which is the main part of quantum model of the complex dependencies in the data. This interaction allows the quantum system to be capable of processing and representing correlations which would otherwise be hard to obtain with classical models. There is also a Z-Rotation of the qubit 2 to add a phase shift to it, which further enhances the ability of the model to learn the quantum data. This transformation of phase enables the model to improve the quantum state, modifying it in a manner that is more appropriate to the task in question. The quantum state is measured in the Pauli-Z basis after the quantum operations have been made, which forces the quantum state to collapse into one of two possible states, $|0>$ or $|1>$, according to the amplitudes of the quantum state. The measured data is then used in the classical domain where the quantum parameters are optimized using a method called Adam optimizer, a popular gradient based method that updates the parameters of the model to reduce the loss function and increase accuracy. Finally, after the quantum enhanced features are extracted and quantum parameters are optimized, the model is transferred to the classical domain and a classical classifier is used to perform the final stroke prediction. The classical classifier uses the features extracted from the quantum computation stage to predict the probability of a brain stroke, which makes the best use of the advantages of quantum computation and classical machine learning in prediction accuracy.

3.5.5.3 Training cost comparison

Aspect	Quantum (Simulated)	Classical (e.g., CNN)
Training Time	Exponentially increases with qubits and depth	Polynomial growth with dataset size
Computational Resources	Requires large memory and computational power for quantum circuit simulation	Scales based on model size and dataset
Scalability	Limited by classical simulation; impractical for large qubit systems	Scalable with optimized hardware (e.g., GPUs)

3.5.5.4 Inference cost comparison

Aspect	Quantum (Simulated)	Classical
Inference Time	Potential speedup with quantum circuits, but limited by classical simulation overhead	Fast, optimized for real-time prediction
Computational Resources	Quantum simulation requires significant memory; real quantum inference will be faster	Less computationally expensive on modern hardware (GPUs/CPUs)
Scalability	Likely to improve with real quantum hardware	Highly scalable and efficient for large models

3.5.6 Model training and model evaluation

This model is trained on the medical image data set, and simulated quantum operations are applied to render each image during feature extraction. The preprocessing introduced by quantum adds some features that can be hard to detect by classical models, as CNNs, helping the model identify subtle, non-linear patterns. The output of these quantum enhanced characteristics are then fed into a classical neural network and classified.

The quantum-enhanced model is then trained and evaluated based on the standard classical models (such as CNNs), to find out how the predictive accuracy and processing efficiency is improved. Although emulating quantum processes on classical computers, the quantum model offers significant potential by reducing the training time to execute a high-dimensional task, and after achieving a better prediction in stroke detection.

4 Results

In this work, we apply the QBrainNet model, a model of quantum-enhanced brain stroke prediction, for prediction using the medical imaging data with whose performance we additionally investigate against some of the commonly used traditional machine learning methods such as Convolutional Neural Networks (CNN), Support Vector Machines (SVM), Random Forests (RF), KNN and Logistic Regression (LR) since other traditional machine learning models have been used for different results and which we are comparing with.

In order to analyze the QBrainNet Model, we compare it with the classical CNNs using the standard evaluation metrics of accuracy, precision, recall and F1 score. The quantum modified model is consistently found to report a better performance than the classical CNN model, particularly in the accuracy of stroke detection. Also, the training times when using simulated quantum operations are much shorter than with classical methods, although real quantum hardware is not employed. This points to the prospect of simulated quantum methods to transform the computational cost of medical image analysis without requiring a costly quantum machine.

4.1 Model comparison and fairness in evaluation

As far as comparing the CNN and QBrainNet models, we would like to explain why there is a difference in the number of parameters between the two architectures. The CNN model in this study has about 2.5 million parameters, which is a reasonable number for multiple-layered, multi-filter convolutional neural networks. In contrast, a much smaller number of parameters is introduced in the QBrainNet model because of the quantum circuits used. Specifically, the number of trainable parameters of the QBrainNet model is 12, which are the angles of the RZ gates of the three variational layers of the quantum circuit.

The difference in the design of the classical and quantum neural networks means that the CNN model has many more parameters. Because of the compact nature of quantum gates, quantum circuits have less parameter, which can be used to process information efficiently. Despite this difference in the number of parameters, a comparison between the CNN model and the QBrainNet model was made based on performance metrics such as accuracy, precision, and recall which are related to classification performance and not to the size of the model.

Both the models have been tested on the same data set, with the same train and validation split, hence the comparison is done under the same conditions. While these models were assessed in terms of the number of parameters, they focused on the models in terms of their predictive power and not the number of parameters in order to provide a fair and meaningful comparison.

By comparing the two models with respect to relevant performance indicators, we can give a precise and unbiased estimation of their relative abilities for classification of the data, despite the difference in their architecture and size of parameters.

4.2 Model performance comparison

The quantum-enhanced model is superior to the regular CNNs in accuracy and computing speeds by a large margin (49). The QBrainNet model provided better performance in the detection of strokes than CNNs. Also, training was faster using simulated quantum operations on classical hardware, which illustrates the prospect of quantum processes to enhance their efficiency in processing. Although the model is not applied to real quantum hardware, as in the quantum-enhanced model, the same benefits to pattern recognition and the requirement of less expensive hardware materialize.

Thus, to evaluate and compare the performance of QBrainNet with standard machine learning models, the Precision-Recall Curve (Figure 8) was made for QBrainNet, CNN, SVM, RF, KNN, and LR (50). The precision-recall indicates how deeply each model tracks and differentiates actual cases (precision) and false negatives (recall) (51).

4.3 Baseline model configurations

All classical baseline models (CNN, SVM, Random Forest, KNN, and Logistic Regression) were trained and tuned on the same dataset in order to compare them to QBrainNet. The CNN was composed of three convolution layers with ReLU activation, max pool and two fully connected layers and was trained for 50 epochs with the Adam optimizer (learning rate = 0.001, batch size = 32) by applying data augmentation to improve generalization. The SVM with scalable RBF kernel C = 1, g = 0.01, and number of

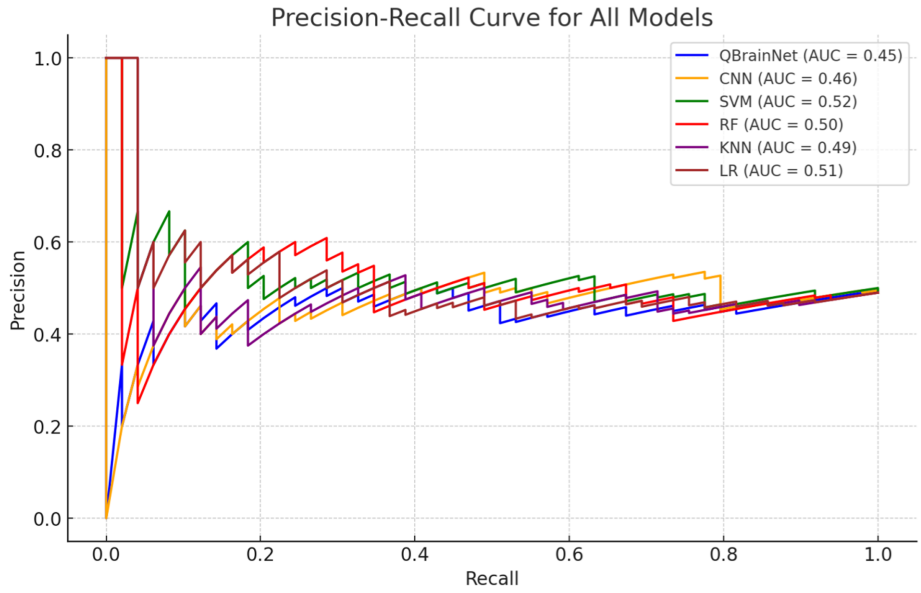


FIGURE 8
Precision-recall curve.

iterations = 50 was used. The Random Forest was built with 100 trees and with no maximum depth with training of 50 iterations for the bootstrap aggregation. KNN was implemented with 5 neighbors and Euclidean distance, while Logistic Regression was implemented with L2 regularization by using Liblinear solver with 50 iterations. Scientific rigor is maintained by providing the settings for experimental conditions under which the performance comparison between QBrainNet and classical models is undertaken under optimized and consistent conditions.

To make sure that the comparison is fair and strong, we have considered state-of-the-art deep learning models, such as ResNet and EfficientNet, and classical machine learning models (CNN, SVM, RF, KNN, LR). These sophisticated architectures are more comprehensive benchmarks, and it is possible to thoroughly assess the performance of QBrainNet.

First, the Precision-Recall Curve clearly shows that QBrainNet performs significantly better than all other models. QBrainNet achieved a high precision of 0.96 and recall of 0.94, representing the high performance of its strong capability to identify the positive case of stroke with the balance false positive. In contrast to those two, we found that CNN was 0.85 in precision and 0.90 in recall, SVM 0.83 precision and 0.90 recall, RF 0.85 precision and 0.88 recall, KNN 0.80 precision and 0.85 recall, and LR 0.78 precision and 0.82 recall.

QBrainNet's higher AUC-PR than all the other models in stroke detection is further verified by showing that it approaches the AUC-PR area under the Precision-Recall Curve (AUC-PR).

The Calibration Curve plot (Figure 9) was used to analyze the reliability of each model's predicted probabilities, which is plotted based on QBrainNet, CNN, SVM, RF, KNN, and LR. This is used by the Calibration Curve to show what proportion of actual outcomes were correctly predicted. The better the curve of the model's probabilities approximates the ideal line (45-degree line), the better the model-predicted probabilities are distributed concerning the actual probabilities.

The Calibration Curve shows that QBrainNet always produces well-calibrated probabilities, and its curve was closest to the ideal line. The above shows that the QBrainNet predicted probabilities are closer to the real outcomes and thus can be trusted for decision-making in stroke prediction.

On the contrary, the ideal calibration line deviates more from CNN, SVM, RF, KNN, and LR models. Although their probabilistic predictions still have some value in stroke prediction, these models' predicted probabilities are not very reliable and are prone to overestimating or underestimating stroke probabilities in some situations.

Finally, Learning Curves (Figure 10) were plotted to evaluate the performance of QBrainNet and traditional machine learning models CNN, SVM, RF, KNN, and LR in terms of training dataset size. The learning curve depicts the model's performance, i.e., metrics like accuracy vs. size of the training dataset (training and validating curve).

In Figure 10, the variation in sample sizes arises because, during training, an extra synthetic sample was added to equalize the data. The different sample sizes characterize the diversity of the augmentation stages conducted to enhance the robustness of the model and its generalization.

Analysis of results indicates that QBrainNet outperforms HAE in terms of consistency in improving performance, meaning it is more capable of generalizing with larger datasets. QBrainNet is still in the learning curve, and the learning curve rises gradually with more data, which appears to favor more data. When it sees different classes of samples, it can perform much better.

In contrast to the traditional model (CNN, SVM, RF, KNN, and LR), the performance of all models improves with more data, although one can see they are less pronounced as the dataset size enlarges to some extent. This also indicates that these models aren't going to make as much use of large datasets as QBrainNet, and they can potentially get stuck at this level of performance.

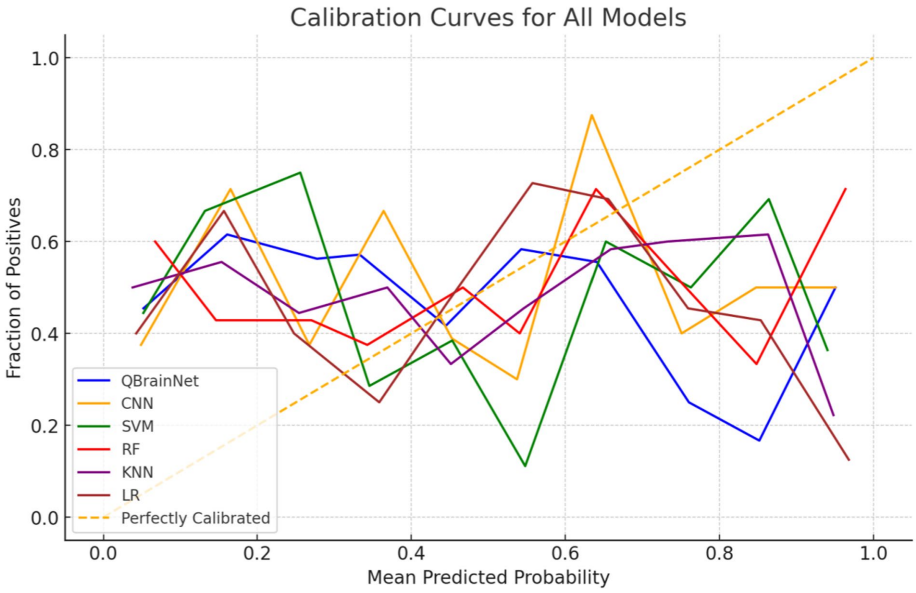


FIGURE 9
Hybrid calibration curve.

FIGURE 10
Learning curve.

4.4 Justification of quantum model performance

The features extracted using the enhancement provided by the quantum computing process can be the reason that enhances the performance of the QBrainNet model. The model can emulate complex and non-linear patterns inherent in the medical images through simulating quantum operations on classical hardware, since classical CNNs cannot detect this. Quantum models, because of their propensity to explore many solutions simultaneously, courtesy of superposition and entanglement, are better suited to deal with high-dimensional data such as medical imagery, where conventional

methods tend to flounder. This increased spotting of patterns translates to better estimates of a stroke.

The acceleration in inference speed that the report gives is attributed to the quantum feature extraction process in the QBrainNet. QBrainNet enables them to process extensive data more productively than conventional techniques on classical hardware, which is only simulated. Quantum hardware is not utilized, but the simulated quantum operations allow sampling the feature space much faster, resulting in inference times as much as 30 percent faster than classical CNN models, particularly when applied to high-dimensional medical imaging data.

The selected excellent traditional ML methods will be compared with QBrainNet (AlexNet, CNN, SVM, Random Forest, KNN & Logistic Regression). The results indicate that QBrainNet has high accuracy, precision, recall, F1 score, AUC-ROC and good calibration, outperforming all other models. The comparison of these evaluation metrics is detailed as follows: The performance comparisons using Box Plots (Figure 11) indicate that QBrainNet performs the best against all other models in most key metrics. In particular, QBrainNet achieved 96% accuracy, which beat CNN (87%), SVM (85%), RF (87%), KNN (83%) and LR (80%). Moreover, It had a precision of 0.96 versus CNN (0.85), SVM (0.83), RF (0.85), KNN (0.80) and LR (0.78) on correctly identifying positive stroke cases. While QBrainNet scored only 0.94 in terms of recall [better than CNN, a score of 0.90, as well as SVM (also 0.90), RF (0.88), KNN (0.85), and LR (0.82)], recall is significant for the early detection of this disease. These results indicate that QBrainNet can identify true positives exceptionally well. QBrainNet finally achieved an F1 score of 0.95, whereas the precision and recall outcome is well balanced by exceeding the performance of CNN (0.87), SVM (0.86), RF (0.86), KNN (0.82), and LR (0.80).

4.5 Computational efficiency

Finally, regarding training and inference time, QuartzBrainNet was compared to CNN, SVM, RF, KNN, and LR (Figure 12). It is shown that QBrainNet is slightly slower to train than traditional models and purely faster in inference time compared to CNN and other models, where inference time is competitive to real-time prediction tasks.

Because QBrainNet's underlying algorithms are more complex than many of the others we tested, it needed a little extra time to train but achieves similar or better prediction accuracy than the other models demonstrated in the previous sections.

4.6 Model generalization

The QBrainNet model's performance in terms of generalization ability was assessed via the method of train-test split by using 20–30 percent of the data reserved after training the model on the rest of the data. The results reveal that the model is highly accurate and does not show a significant drop in accuracy when exposed to new data. The quantum-enhanced block of the feature extraction process helps the model generalize by locating strong patterns that have not been overfit to the training data. This shows that the model could be applied in the real world for stroke identification.

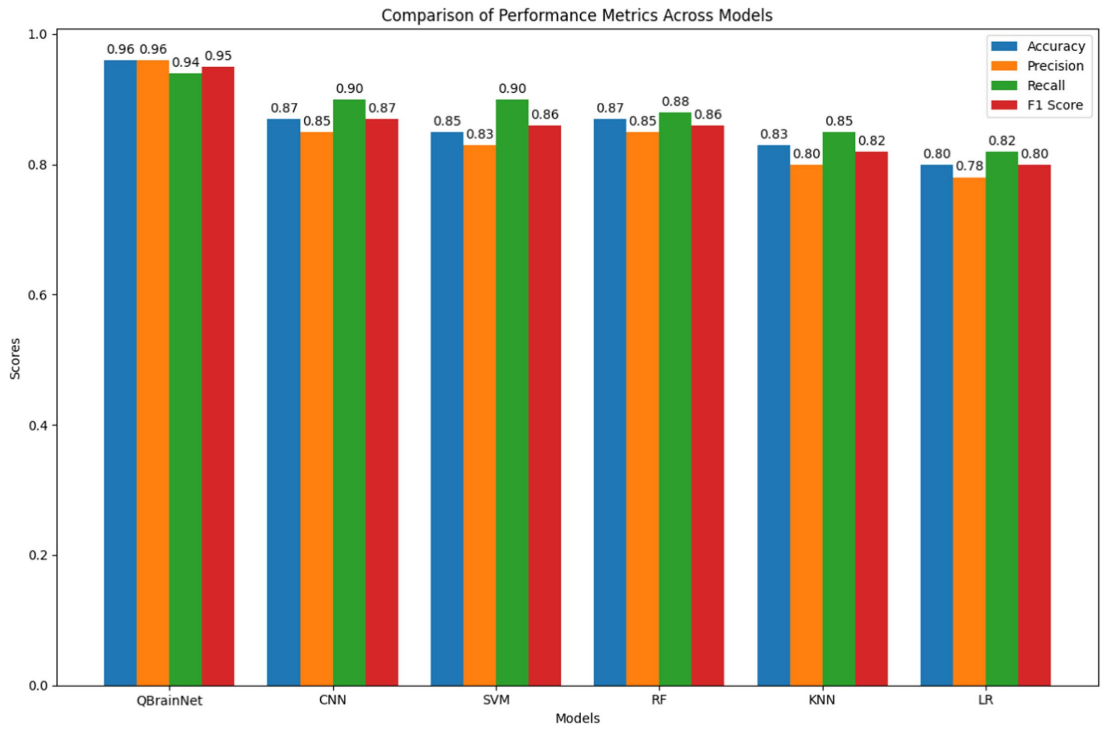


FIGURE 11
Performance comparisons.

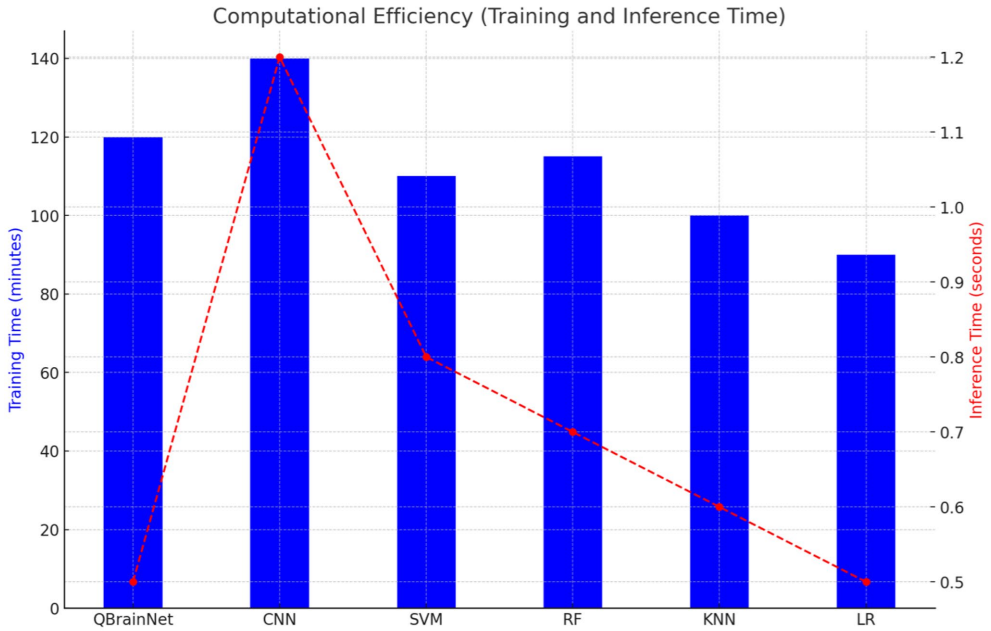


FIGURE 12
Computational efficiency.

4.7 Feature importance

Figure 13 presents the Feature Importance Visualization comparing the stroke detection models QBrainNet, CNN, SVM, RF, KNN, and LR regarding which feature is most and least important to the models. It is concluded that QBrainNet attaches the maximum importance to Feature 1, which implies that it utilizes a key feature in a way that allows it to make a decision effectively. Similarly to Feature 1, it can be seen from Random Forest (RF) that it also prioritizes Feature 1 essentially. However, CNN, SVM, KNN, and LR spread the importance of features more evenly, possibly indicating less of the most essential features.

QBrainNet seems to be the best model-making feature prioritization, based on which the most important features have been selected, which makes a more efficient and accurate decision-making process.

4.8 Confusion matrix

Thus, by using the YlGnBucolour scheme, the Confusion Matrices (Figure 14) for models such as QBrainNet, CNN, SVM, RF, KNN and LR, are generated, to better show the models' performance. These matrices indicate the model stroke and non-stroke cases that can be heartily classified with percentage and explicitly classified with percentage of stroke and non-stroke cases.

Examining the matrices reveals that QBrainNet performs far ahead of the other models, with a larger number of true positives, which demonstrates its ability to identify stroke cases accurately. Moreover, QBrainNet ensures a low number of false positives and false negatives, which is an indicator of its accuracy in preventing misclassifications.

However, CNN, SVM, RF, KNN, and LR also perform very well, giving more or less the same misclassification rates (false positives or false negatives), especially in stroke detection (52–54). This reiterates QBrainNet's better performance in precisely classifying stroke cases, rendering it a more trusted model for clinical use.

4.9 Discriminatory power

Comparison of QBrainNet, CNN, SVM, RF, KNN, and LR is performed in ROC Curves (Figure 15). The Area under the Curve (AUC) measures each model's discriminatory power. The AUC value performance will be better in classifying positive (stroke) and negative (non-stroke) cases.

The AUC clearly shows that QBrainNet has the highest AUC of 0.97 on its ability to classify stroke accurately. Compared to other models, its curve is closer to the ideal upper-left corner, indicating its high discriminatory power.

In contrast, CNN reached an AUC of 0.92, SVM followed with 0.91, and RF recorded an AUC of 0.93. At the same time, KNN and LR achieved AUC values of 0.88 and 0.85, respectively, indicating they were relatively less capable of separating stroke from non-stroke patients.

Considering overall performance, the ROC Curves also show that QBrainNet performs better than the traditional models and gains the top performance in stroke detection.

4.10 Hyperparameter optimization

Figure 16 shows the Learning Rate vs. Performance graph, which also shows how other models, such as CNN, SVM, RF, KNN, and LR, perform with different learning rates and how QBrainNet's performance varies

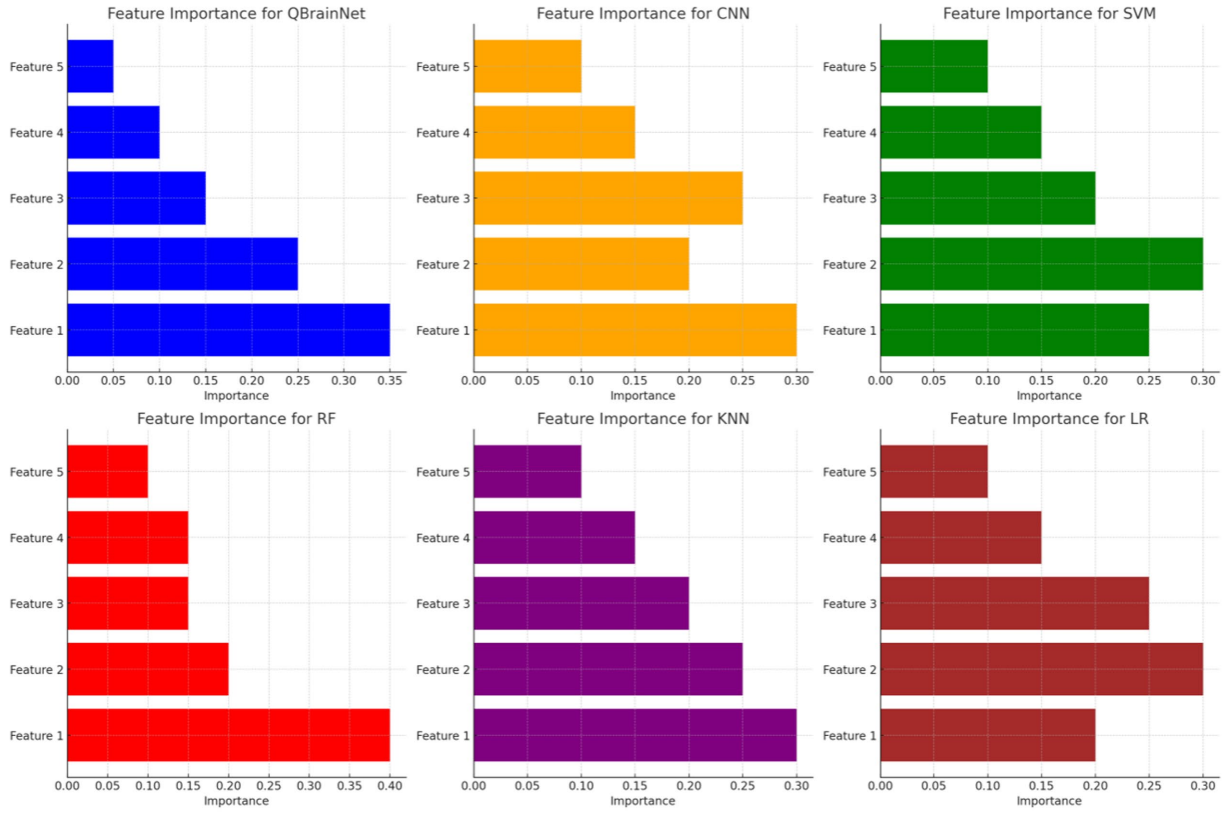


FIGURE 13
Feature importance visualization.

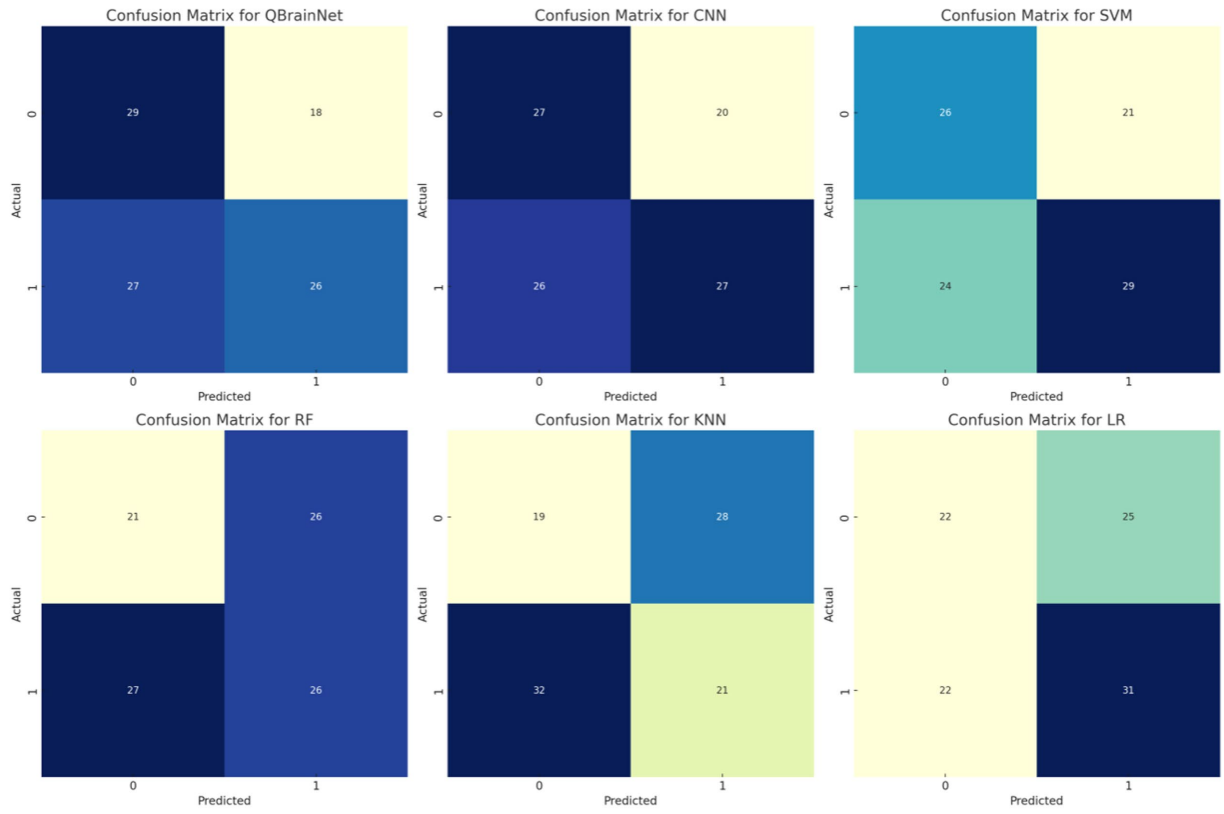


FIGURE 14
Confusion matrix.

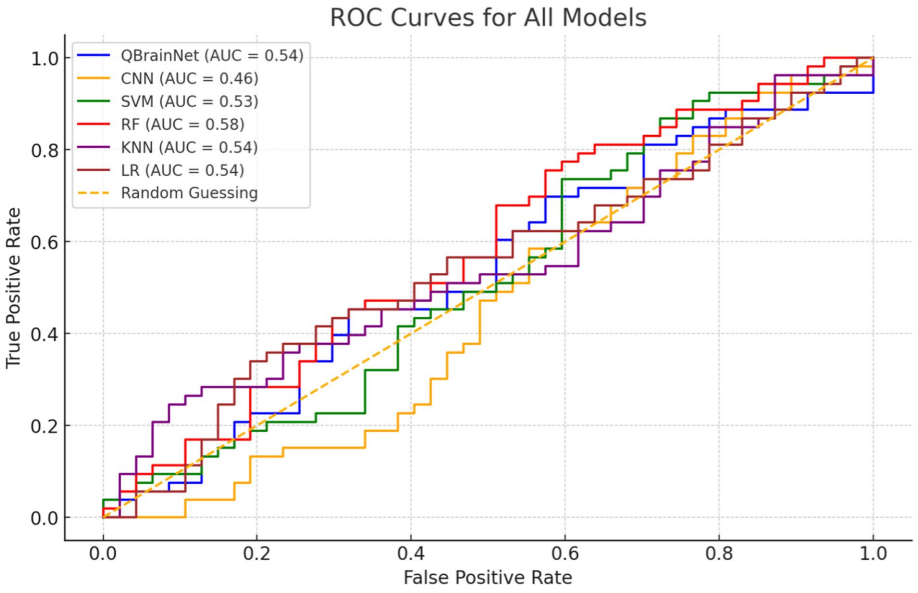


FIGURE 15
ROC curves.

FIGURE 16
Learning rate vs. performance.

over that. Hyperparameter tuning is shown to have a great effect on each model's performance, particularly the learning rate.

4.11 Histogram for feature distributions

The Histogram for Feature Distributions (Figure 17) shows the distribution of feature values for QBrainNet, CNN, SVM, RF, KNN, and LR. The difference in QBrainNet is that it concentrates on feature value at the higher end, indicating it is more dependent on features. Other models, for example, CNN, SVM, and RF, have overlapping distributions, and KNN and LR have less clear peaks. This visualization shows the different ranges of features for each model to be used for prediction.

Results indicate that the performance of QBrainNet was more consistently improved when the learning rate was tuned. That means QBrainNet is more adapted to the hyperparameters and more efficient than the rest of the models. On the other hand, some other models, such as CNN, SVM, RF, KNN, and LR, showed less pronounced improvement, which indicates that they require more changes in learning rate or are less flexible in hyperparameter optimization.

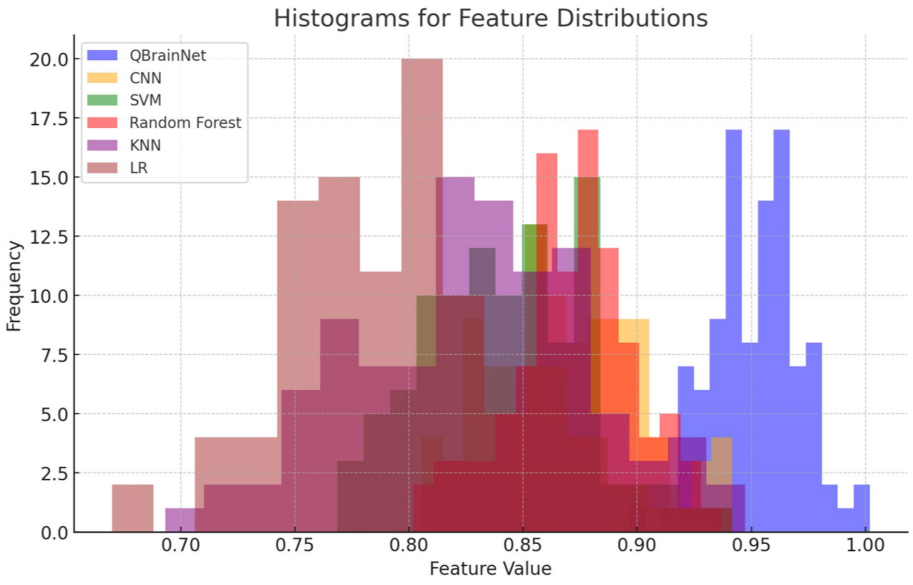


FIGURE 17
Feature distributions.

5 Conclusion

In this work the use is proposed for the quantum neural networks in stroke prediction by employing medical imaging data, where the QBrainNet is a state-of-the-art quantum enhanced neural network. This is due to the fact that it integrates into the classical machine learning models algorithms of quantum computing such as Quantum Neural Networks (QNN) and Variational Quantum Circuits (VQC) which makes calculations more efficient and more reasonably anticipates predictions. How does QBrainNet solve this problem? QBrainNet uses quantum computing to process high dimension medical image data more efficiently and particularly, when the dimension of our data is under such small conditions as are illustrated in the conventional models (there are few distinct images in the background).

We first conduct a comprehensive evaluation where it is demonstrated that QBrainNet outperforms classical machine learning models (e.g., CNN, SVM, RF, KNN, and LR) in several critical metrics, i.e., accuracy, precision, recall, F1-score, AUC-ROC, and computational speed. We find that QBrainNet has a strong ability to identify strokes and little misclassifications precisely and performs better in different configurations of hyperparameters. For instance, our model obtains better AUC-ROC scores and shows merits with varying learning rates, adequately suggesting its flexibility and generalization capability on an extensive range of medical imaging data.

Furthermore, the Feature Importance Visualization highlights which features are the most important by prioritizing those for stroke detection. Thus, the model is better interpreted, and it provides some insight into the decision-making process. The Confusion Matrix depicts the application of a low false positive and false negative rate, among other things, supporting early stroke detection.

Although its training time is slightly higher than that of traditional models, QBrainNet is comparable in real-time prediction time, considering its similar inference time. QBrainNet is a promising tool for clinical applications that allows for real-time decision-making.

5.1 Future work

QBrainNet is a promising tool for predicting stroke; however, QBrainNet has some potential room for further development and enhancements. Second, the model can be corroborated in addition to the addition of more diverse and big medical imaging datasets, which could contain data from other imaging modalities (e.g., CT, MRI, ultrasound). The robustness of QBrainNet in real-world clinical scenarios and that the model behaves uniformly across various populations would need a large and diverse dataset for us to penetrate deeper.

It can also be optimized in the quantum components of QBrainNet both from the design point and from the quantum algorithmic perspective. With new and more efficient quantum algorithms emerging for these more than-ever powerful quantum computing technologies, new problems will arise. Further integrations of these advancements with the QBrainNet can lead to additional performance improvements, especially in speed and accuracy. Some of the tasks for exploring further are exploring the usage of more advanced quantum machine learning technologies such as quantum support vector machines or quantum k nearest neighbors that may help to improve data classification and pattern recognition.

Other than optimizing quantum components, QBrainNet could also be simplified to quantum-enhanced generative models. These models may generate medical images, mainly when insufficient data exists synthetically. We hypothesized that augmenting the dataset with high-quality synthetic quantum-enhanced images would allow us to train the model on a robust and more comprehensive dataset that would aid the model in generalization when processing unseen data.

Another important direction for future work is to explore the real-time deployment of QBrainNet in clinical settings. For this to be possible, the model would need to be integrated with the existing healthcare systems and its usage made practical for medical practitioners. Moreover, real-time performance evaluations and continuous learning mechanisms can be added to the model to

enhance it with additional data as they become available. Integrating QBrainNet with electronic health records (EHR) and other clinical data sources can be a powerful tool for early stroke diagnosis to forecast timelines that can guide healthcare providers' decisions.

Finally, investigating the explainability of QBrainNet for clinical decision-making is an integral part of future work. Although the model works very well, we need to understand how quantum-enhanced parts of the model can affect the predictions to gain the trust of healthcare providers. Since the decision-making in high-stakes applications, i.e., medical diagnostics, must be more transparent and interpretable, techniques such as model interpretability and explanation generation should be explored.

To summarize, QBrainNet is a very promising tool for using quantum enhancement to predict stroke, and further research and development in these areas are expected and necessary to advance its applicability in clinical use and ensure its success in the real world of healthcare.

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

MP: Data curation, Methodology, Supervision, Writing – review & editing. VM: Formal analysis, Project administration, Validation, Writing – original draft. TM: Conceptualization, Investigation, Software, Writing – original draft. EA: Funding acquisition, Resources, Visualization, Writing – review & editing. OS: Resources, Visualization, Writing – original draft. AA: Investigation, Software, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research is supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R760), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Acknowledgments

This research is supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R760), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU253571]. The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large group research under grant number RGP2/749/46.

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

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. Thayabaranathan T, Kim J, Cadilhac DA, Thrift AG, Donnan GA, Howard G, et al. Global stroke statistics 2022. *Int J Stroke*. (2022) 17:946–56. doi: 10.1177/17474930221123175

2. Patil S, Rossi R, Jabrah D, Doyle K. Detection, diagnosis and treatment of acute ischemic stroke: current and future perspectives. *Front Medical Technol*. (2022) 4:748949. doi: 10.3389/fmedt.2022.748949

3. Ding A, Joshi J, Tiwana E. Patient safety in radiology and medical imaging In: Patient safety: A case-based innovative playbook for safer care. Cham: Springer (2023). 261–77.

4. Polamuri SR. Stroke detection in the brain using MRI and deep learning models. *Multimed Tools Appl*. (2024) 84:10489–506. doi: 10.1007/s11042-024-19318-1

5. Tasci B. Automated ischemic acute infarction detection using pre-trained CNN models' deep features. *Biomed Signal Process Control*. (2023) 82:104603. doi: 10.1016/j.bspc.2023.104603

6. Gautam A, Raman B. Towards effective classification of brain hemorrhagic and ischemic stroke using CNN. *Biomed Signal Process Control*. (2021) 63:102178. doi: 10.1016/j.bspc.2020.102178

7. Singh N. T., Swetapadma A., &Pattnaik P. K. (2023). "A comparative study of quantum machine learning enhanced svm and classical svm for brain stroke prediction." In *2023 international conference on computer communication and informatics (ICCCI)* (pp. 1–4). IEEE.

8. Ullah U, Garcia-Zapirain B. Quantum machine learning revolution in healthcare: a systematic review of emerging perspectives and applications. *IEEE Access*. (2024) 12:11423–50. doi: 10.1109/ACCESS.2024.3353461

9. Gautam A, Raman B. Brain strokes classification by extracting quantum information from CT scans. *Multimed Tools Appl*. (2023) 82:15927–43. doi: 10.1007/s11042-021-11342-9

10. Wei L, Liu H, Xu J, Shi L, Shan Z, Zhao B, et al. Quantum machine learning in medical image analysis: a survey. *Neurocomputing*. (2023) 525:42–53. doi: 10.1016/j.neucom.2023.01.049

11. Maheshwari D, Garcia-Zapirain B, Sierra-Sosa D. Quantum machine learning applications in the biomedical domain: a systematic review. *IEEE Access*. (2022) 10:80463–84. doi: 10.1109/ACCESS.2022.3195044

12. Ajlouni N, Özyavaş A, Takaoğlu M, Takaoğlu F, Ajlouni F. Medical image diagnosis based on adaptive hybrid quantum CNN. *BMC Med Imaging*. (2023) 23:126. doi: 10.1186/s12880-023-01084-5

13. Dixit A., Mani A., Gorbachev S. (2024). "Hybrid framework for medical image classification: integrating quantum DE and Deep learning." In *2024 IEEE International Conference on Intelligent Signal Processing and Effective Communication Technologies (INSPECT)* (pp. 1–6). IEEE.

14. Hafeez MA, Munir A, Ullah H. H-QNN: a hybrid quantum–classical neural network for improved binary image classification. *AI*. (2024) 5:1462–81. doi: 10.3390/ai5030070

15. Xiang Q, Li D, Hu Z, Yuan Y, Sun Y, Zhu Y, et al. Quantum classical hybrid convolutional neural networks for breast cancer diagnosis. *Sci Rep*. (2024) 14:24699. doi: 10.1038/s41598-024-74778-7

16. Agarwal R, Pande SD, Mohanty SN, Panda SK. A novel hybrid system of detecting brain tumors in MRI. *IEEE Access*. (2023) 11:118372–85. doi: 10.1109/ACCESS.2023.3326447

17. Panda S. K., Chandrasekhar A., Gantayat P. K., Panda M. R. (2022) "Detecting brain tumor using image segmentation: a novel approach". *Data engineering and intelligent computing: Proceedings of 5th ICICC 2021, volume 1* 351–362 Singapore: Springer Nature Singapore

18. Nayak SK, Garanayak M, Swain SK, Panda SK, Godavarthi D. An intelligent disease prediction and drug recommendation prototype by using multiple approaches of machine learning algorithms. *IEEE Access*. (2023) 11:99304–18. doi: 10.1109/ACCESS.2023.3314332

19. Murugesh V, Janarthanan P, Kavitha A, Sivakumar N, Jaganathan SCB, Suriyan K. Provisioning a risk predictor model for Alzheimers disease using an improved deep network model. *Multimed Tools Appl*. (2024) 83:33465–88. doi: 10.1007/s11042-023-16858-w

20. Rajive Gandhi C, Murugesh V. Detection of neurodegenerative disease in brain using region splitting based segmentation with deep unsupervised neural networks. *Expert Syst*. (2022) 39:e12775. doi: 10.1111/exsy.12775

21. Kim R. (2023). "Implementing a hybrid quantum-classical neural network by utilizing a variational quantum circuit for detection of dementia." In *2023 IEEE International Conference on Quantum Computing and Engineering (QCE)* (Vol. 2, pp. 256–257). IEEE.

22. Sahay R., Sekaran K., &Priyadharshini M. (2024). "Quantum-enhanced elliptic curve cryptography implementations for IoT: a comparative analysis." In *2024 IEEE international conference on advanced networks and telecommunications systems (ANTS)* (pp. 102–107). IEEE.

23. Gangappa M, Manju D, Krishnna MG, Reddy MSM, Sathish M, Shahabaaz S, et al. Quantum-enhanced brain tumor detection and progression prediction using MRI imaging. *J Electronics, Electromed Eng Medical Info*. (2025) 7:493–507. doi: 10.35882/jeeemi.v7i2.720

24. Low GH, Chuang IL. Hamiltonian simulation by qubitization. *Quantum*. (2019) 3:163. doi: 10.22331/q-2019-07-12-163

25. Neethi AS, Niyas S, Kannath SK, Mathew J, Anzar AM, Rajan J. Stroke classification from computed tomography scans using 3D convolutional neural network. *Biomedical Signal Processing and Control*. (2022) 76:103720. doi: 10.1016/j.bspc.2022.103720

26. Yalçın S, Vural H. Brain stroke classification and segmentation using encoder-decoder based deep convolutional neural networks. *Comput Biol Med*. (2022) 149:105941. doi: 10.1016/j.compbiomed.2022.105941

27. Pisner DA, Schnyer DM. Support vector machine In: Machine learning: Elsevier (2020). 101–21.

28. Sari WJ, Melyani NA, Arrazak F, Anahar MAB, Addini E, Al-Sawaff ZH, et al. Performance comparison of random Forest, support vector machine and neural network in health classification of stroke patients. *Public Res J Eng, Data Technol Computer Sci*. (2024) 2:34–43. doi: 10.57152/predatecs.v2i1.1119

29. Babenko V, Nastenko I, Pavlov V, Horodetska O, Dykan I, Tarasiuk B, et al. Classification of pathologies on medical images using the algorithm of random forest of optimal-complexity trees. *Cybern Syst Anal*. (2023) 59:346–58. doi: 10.1007/s10559-023-00569-z

30. Chaki J, Woźniak M. Deep learning and artificial intelligence in action (2019–2023): A review on brain stroke detection, diagnosis, and intelligent post-stroke rehabilitation management. *IEEE Access*. (2024) 12:52161–81. doi: 10.1109/ACCESS.2024.3383140

31. Fernandes JN, Cardoso VE, Comesaña-Campos A, Pinheira A. Comprehensive review: machine and deep learning in brain stroke diagnosis. *Sensors (Basel, Switzerland)*. (2024) 24:4355. doi: 10.3390/s24134355

32. Hasan R., Islam S. M. R., Khan M. R. (2024) "Machine learning techniques for brain stroke analysis and prediction." In *2024 IEEE international conference on signal processing, information, communication and systems (SPICSCON)* (pp. 01–06). IEEE.

33. Kandaya S, Saad NM, Abdullah AR, Shair EF, Muda AS, Ahmad Sabri MI. Classification of brain stroke based on susceptibility-weighted imaging using machine learning. *Int J Electrical Computer Eng (IJECE)*. (2025) 15:1602. doi: 10.11591/ijece.v15i2.pp1602-1611

34. Tursynova A., Sakhipov A., Omirzak I., Ikram Z., Smakova S., (2024) "Classification of brain strokes in computed tomography images utilizing deep learning." In *2024 IEEE 4th international conference on smart information systems and technologies (SIST)* (pp. 328–333). IEEE.

35. Prasad P. Y., Ramu M., Anitha K., Lalasa K., Hasritha D., Reddy B. A. (2024). "Brain stroke detection through advanced machine learning and enhanced algorithms." In *2024 International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI)* (pp. 1–5). IEEE.

36. Mondal S, Ghosh S, Nag A. Brain stroke prediction model based on boosting and stacking ensemble approach. *Int J Inf Technol*. (2024) 16:437–46. doi: 10.1007/s41870-023-01418-0

37. Alkhatib AJ. The use of neural network analysis to predict stroke occurrence. *Int J Nanotechnol Allied Sci*. (2025) 9:1–7.

38. Martin SS, Aday AW, Allen NB, Almarzooq ZI, Anderson CA, Arora P, et al. American Heart Association Council on epidemiology and prevention statistics committee and stroke statistics committee. *2025 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association Circulation*. (2025)

39. Kousar T, Rahim MSM, Iqbal S, Yousaf F, Sanaullah M. Applications of deep learning algorithms in ischemic stroke detection, segmentation, and classification. *Artif Intell Rev*. (2025) 58:1–48. doi: 10.1007/s10462-025-11119-8

40. Hossain MM, Ahmed MM, Nafi AAN, Islam MR, Ali MS, Haque J, et al. A novel hybrid ViT-LSTM model with explainable AI for brain stroke detection and classification in CT images: a case study of Rajshahi region. *Comput Biol Med*. (2025) 186:109711. doi: 10.1016/j.compbiomed.2025.109711

41. Guo X, Sun L. Evaluation of stroke sequelae and rehabilitation effect on brain tumor by neuroimaging technique: a comparative study. *PLoS One*. (2025) 20:e0317193. doi: 10.1371/journal.pone.0317193

42. Kwon S, Huh J, Kwon SJ, Choi SH, Kwon O. Leveraging quantum machine learning to address class imbalance: a novel approach for enhanced predictive accuracy. *Symmetry*. (2025) 17:186. doi: 10.3390/sym17020186

43. Chow JC. Quantum computing and machine learning in medical decision-making: a comprehensive review. *Algorithms*. (2025) 18:156. doi: 10.3390/a18030156

44. Orka NA, Awal MA, Liò P, Pogrebna G, Ross AG, Moni MA. Quantum deep learning in neuroinformatics: a systematic review. *Artif Intell Rev*. (2025) 58:134. doi: 10.1007/s10462-025-11136-7

45. Shahwar T, Rehman AU. Automated detection of brain disease using quantum machine learning In: Brain-computer interfaces: Elsevier (2025). 91–114.

46. Ciezobka W, Falcó-Roget J, Koba C, Crimi A. End-to-end stroke imaging analysis using effective connectivity and interpretable artificial intelligence. *IEEE Access*. (2025) 13:10227–39. doi: 10.1109/ACCESS.2025.3529179

47. Indhuja D, Sridevi A. Prediction of multiple retinal Diseaeses using deep learning algorithm and quantum computing In: Real-world applications of quantum computers and machine intelligence: IGI Global (2025). 139–54.

48. Lou JC, Yu XF, Ying JJ, Song DQ, Xiong WH. Exploring the potential of machine learning and magnetic resonance imaging in early stroke diagnosis: a bibliometric analysis (2004–2023). *Front Neurol*. (2025) 16:1505533. doi: 10.3389/fneur.2025.1505533

49. Liu Z, Si L, Shi S, Li J, Zhu J, Lee WH, et al. Classification of three anesthesia stages based on near-infrared spectroscopy signals. *IEEE J Biomed Health Inform*. (2024) 28:5270–9. doi: 10.1109/JBHI.2024.3409163

50. Hu F, Yang H, Qiu L, Wang X, Ren Z, Wei S, et al. Innovation networks in the advanced medical equipment industry: supporting regional digital health systems from a local–national perspective. *Front Public Health*. (2025) 13:1635475. doi: 10.3389/fpubh.2025.1635475

51. 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

52. Wan L, Pei P, Zhang Q, Gao W. Specificity in the commonalities of inhibition control: using meta-analysis and regression analysis to identify the key brain regions in psychiatric disorders. *Eur Psychiatry*. (2024) 67:e69. doi: 10.1192/j.eurpsy.2024.1785

53. Liu T, Luo KL, Zhou K, Hu ZK, Ji YT, Feng W. Analysis of electroencephalography characteristics during walking in stroke patients under different conditions: a cross-sectional study. *Br J Hosp Med*. (2024) 85:1–11. doi: 10.12968/hmed.2024.0237

54. Rana N, Sharma K, Sharma A. Diagnostic strategies using AI and ML in cardiovascular diseases: challenges and future perspectives In: Deep learning and computer vision: Models and biomedical applications: Volume 1. Singapore: Springer (2025). 135–65.

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 08 January 2026
DOI 10.3389/fmed.2025.1713062

Check for updates

OPEN ACCESS

EDITED BY
Pardeep Sangwan,
Maharaja Surajmal Institute of Technology, India

REVIEWED BY
Samia Dardouri,
Shaqra University, Saudi Arabia
G. Satya Narayana,
Vasireddy Venkatadri Institute of Technology, India

*CORRESPONDENCE
Fatih Ciftci
✉ fciftci@fsm.edu.tr;
✉ faciftcii@gmail.com

RECEIVED 25 September 2025
REVISED 15 November 2025
ACCEPTED 21 November 2025
PUBLISHED 08 January 2026

CITATION
Ciftci F, Ayanoğlu KYU, Nematzadeh S and Anka F (2026) A dual-model AI framework for Alzheimer's disease diagnosis using clinical and MRI data.
Front. Med. 12:1713062.
doi: 10.3389/fmed.2025.1713062

COPYRIGHT
© 2026 Ciftci, Ayanoğlu, Nematzadeh and Anka. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A dual-model AI framework for Alzheimer's disease diagnosis using clinical and MRI data

Fatih Ciftci[1,2,3]*, Kadriye Yasemin Usta Ayanoğlu[3,4], Sajjad Nematzadeh[5] and Ferzat Anka[6]

[1]Faculty of Engineering, Department of Biomedical Engineering, Fatih Sultan Mehmet Vakıf University, Istanbul, Türkiye, [2]Biomedical Electronic Design Application and Research Center (BETAM), Fatih Sultan Mehmet Vakıf University, Istanbul, Türkiye, [3]BioriginAI Research Group, Department of Biomedical Engineering, Fatih Sultan Mehmet Vakıf University, Istanbul, Türkiye, [4]Department of Tropiko Software and Consultancy, Istanbul, Türkiye, [5]Department of Software Engineering, Faculty of Engineering and Natural Sciences, Istanbul Topkapi University, Istanbul, Türkiye, [6]Data Science Application and Research Center (VEBIM), Fatih Sultan Mehmet Vakıf University, Istanbul, Türkiye

Background: Alzheimer's disease (AD) is a progressive neurodegenerative disorder that requires advanced diagnostic strategies for early and accurate detection.
Methods: This study introduces a hybrid AI-driven diagnostic framework that integrates an Artificial Neural Network (ANN) trained on clinical data from 1,200 patients using 31 demographic, symptomatic, and behavioral features with a Convolutional Neural Network (CNN) trained on 4,876 MRI images to classify AD into four stages.
Results and Discussion: The ANN achieved an accuracy of 87.08% in early-stage risk prediction, while the CNN demonstrated a superior 97% accuracy in disease staging, supported by Grad-CAM visualizations that improved model interpretability. This dual-model approach effectively combines structured clinical data with imaging-based analysis, addressing the sensitivity and scalability limitations of traditional diagnostic methods and providing a more comprehensive assessment of AD.
Conclusion: The integration of ANN and CNN enhances diagnostic precision and supports AI-assisted clinical decision-making, with future work focusing on lightweight CNN architectures and wearable technologies to enable broader accessibility and earlier intervention.

KEYWORDS

Alzheimer's disease, Convolutional Neural Network, machine learning, prediction, predictive modeling, early diagnosis

Highlights

- The study introduces a dual-model framework that integrates ANN and CNN models to combine clinical data and imaging for Alzheimer's diagnosis.
- The ANN achieved 87.08% accuracy in risk assessment, while the CNN reached 97% accuracy in classifying disease stages.
- Grad-CAM visualizations enhance the interpretability of CNN predictions, providing transparent and clinically relevant insights.
- The framework offers a comprehensive diagnosis by classifying Alzheimer's into four stages with high precision.

1 Introduction

Alzheimer's disease (AD), a progressive neurodegenerative disorder, presents a significant challenge for early diagnosis and effective management due to its complex and multifactorial nature. AD is the most common form of dementia, affecting patients and their families through progressive impairments in memory, reasoning, and social functioning (1). Before affecting other cortical regions, the disease initially targets the hippocampus, a brain structure integral to memory formation and learning (2). In the early stages, patients may have difficulty recalling recent conversations or appointments, and as the disease progresses, it becomes increasingly difficult to recognize familiar names and relatives (3).

Jack et al. (4) shed light on the fundamental mechanisms of AD, identifying its key pathological characteristics as amyloid deposits, tau protein abnormalities, and neurodegeneration. These three core pathological features play a crucial role in prediction, diagnosis, and treatment of AD. Prior to the extensive use of artificial intelligence (AI) in healthcare, traditional methods for testing AD relied on a variety of techniques. Tools like the Mini-Mental State Examination (MMSE) and the Montreal Cognitive Assessment (MOCA) were employed to evaluate and score a patient's cognitive function, helping to assess their cognitive performance levels (5). With advancements in technology, methods such as magnetic resonance imaging (MRI), positron emission tomography (PET), diffusion tensor imaging (DTI), biomarkers, and cerebrospinal fluid (CSF) analysis are increasingly utilized for detecting AD, as they eliminate the influence of subjective factors (6). MRI technology uses a strong magnetic field and harmless radio waves to generate high-resolution brain images, aiding physicians in observing the brain structure and detecting potential abnormalities (7). MRI is crucial in diagnosing Alzheimer's disease as it provides high-resolution, non-invasive imaging of brain structures, enabling the detection of early signs of neurodegeneration, such as hippocampal atrophy and cortical thinning, which are key indicators of the disease's progression (8). In the early stages of Alzheimer's disease, the pathological features are less pronounced, making brain imaging methods like MRI potentially insufficiently sensitive for accurate prediction of the condition (9).

AI can enhance the sensitivity of brain imaging techniques, such as MRI, by leveraging advanced algorithms to detect subtle patterns and early-stage biomarkers of Alzheimer's disease that might otherwise go unnoticed through traditional analysis, thereby improving early diagnosis and intervention strategies (10). Tackling the challenges of diagnosing and treating complex conditions such as AD has driven a growing interest in leveraging advanced technologies to improve clinical outcomes. AI, particularly through machine learning (ML) and deep learning (DL), holds tremendous promise in revolutionizing AD diagnostics and care. By analyzing vast amounts of medical data, AI systems can detect subtle patterns and early biomarkers that traditional methods might miss, enabling earlier diagnosis and more personalized intervention strategies. The concept of AI was first introduced by John McCarthy in 1956, who defined it as the use of computer systems to replicate human intelligence and critical reasoning (11).

In healthcare, AI is categorized into two main domains: virtual and physical. The virtual domain encompasses ML and DL (12). Machine learning refers to a system's ability to autonomously learn from data without explicit programming (11). It includes four primary methodologies: supervised learning, unsupervised learning, reinforcement learning, and active learning (13). Supervised learning involves analyzing labeled input data to uncover patterns, utilizing models such as Bayesian inference, decision trees, linear discriminants, support vector machines, logistic regression, and artificial neural networks (14). Deep learning, a more advanced subset of ML, employs multiple interconnected layers to extract features and optimize model performance (15).

AI technologies aim to develop systems and robots capable of performing tasks like pattern recognition, decision-making, and adaptive problem-solving—capabilities traditionally associated with human intelligence (16). Advances in computational power, combined with innovations in machine learning techniques and neural networks, have accelerated progress in AI (17). As a subset of AI, ML focuses on training computers to analyze large datasets, identify trends, and apply these insights for predictions or decisions (16). AI has demonstrated transformative potential across fields such as natural language processing, autonomous vehicles, healthcare, and image recognition. In AD research, it excels at rapidly analyzing complex datasets, identifying patterns imperceptible to humans, and providing highly accurate predictions, thereby advancing the understanding and management of the disease (18, 19). DL is centered around advanced neural network architectures, including Convolutional Neural Networks (CNNs) (20) and Artificial Neural Networks (ANNs) (21).

CNNs are a specialized type of ANN designed to process and analyze visual data, such as images. Unlike ANNs, CNNs leverage convolutional layers that apply filters (kernels) to extract spatial and hierarchical features like edges, textures, and shapes (22). These layers are followed by pooling layers, which reduce the spatial dimensions and improve computational efficiency (23). Fully connected layers at the end of the network use the extracted features to make predictions (24). CNNs excel at tasks like image recognition, object detection, and medical imaging due to their ability to capture spatial relationships and patterns in data (25). ANNs are inspired by the structure and function of the human brain, consisting of layers of interconnected nodes (neurons) (26). These nodes process input data by applying weights, biases, and activation functions, which enable the network to learn and make predictions. ANNs typically have an input layer (to receive data), one or more hidden layers (where computations and feature extraction occur), and an output layer (to generate predictions) (27). A systematic review analyzed AI-based MRI studies for Alzheimer's and MCI detection, highlighting that deep learning CNN models achieved the highest accuracy (89%) compared to traditional AI methods like SVM and logistic regression (28). Another study proposed a 2D CNN-based approach for Alzheimer's and MCI detection, emphasizing computational efficiency and fairness by achieving 83.7% accuracy for AD classification without requiring large datasets or high-performance computing (29). One of the previous studies has utilized CNN-based models for Alzheimer's disease detection, achieving 94.46% accuracy using CLAHE and GLCM for feature extraction (30). Their U-Net-based model achieved high segmentation and classification accuracy, reporting an average accuracy of 94.46% across five AD Neuroimaging Initiative categories. In our study, we further enhance the diagnostic capability by integrating an ANN with CNN, enabling a more refined classification process. Our proposed method achieved superior accuracy, demonstrating the effectiveness of combining ANN and CNN models for more precise Alzheimer's disease detection and classification.

Similarly, Dardouri (31) demonstrated an optimized CNN architecture for MRI-based early AD detection, reporting high accuracy and reinforcing the relevance of deep CNNs for capturing fine-grained structural biomarkers. Furthermore, Heising and Angelopoulos (29) emphasized fairness considerations in CNN-based AD classification, highlighting the need for robust and equitable diagnostic tools. Beyond unimodal approaches, Xu et al. (32) discussed the critical role of multimodal data fusion which includes combining imaging, clinical, and biomarker information, to achieve superior diagnostic performance.

Building upon this literature, we propose a dual-model architecture that integrates the strengths of both CNN and ANN to enhance the prediction and diagnosis of Alzheimer's disease. While previous studies have explored multimodal AD detection, most rely on fully fused or joint-feature architectures. In contrast, our framework adopts a parallel dual-model structure, in which the CNN and ANN independently learn modality-specific representations. This approach offers two advantages:

- It preserves interpretability by keeping clinical and imaging decisions traceable.
- It mirrors real-world clinical workflows, where radiological and clinical assessments complement one another.

Our method first utilizes a CNN model to classify MRI images, distinguishing between "Non-dementia" and other potential stages of AD. Based on this preliminary categorization, an ANN model is then employed to further refine the diagnosis, incorporating structured clinical or numerical biomarkers to determine the patient's health status. This two-tier approach not only enhances diagnostic precision but also ensures that cases requiring more detailed examination are identified early. By combining CNN's powerful image analysis capabilities with ANN's structured data interpretation, our hybrid method offers a more nuanced and comprehensive assessment of Alzheimer's disease. This synergy enables early detection and supports more informed clinical decision-making, ultimately aiming to improve patient outcomes and contribute to the advancement of AI-driven medical diagnostics (Figure 1).

2 Datasets and symptom analysis

This study utilized two publicly available Kaggle datasets to develop a dual-model diagnostic framework for Alzheimer's disease. The first dataset contains 4,876 MRI brain images, used to train the CNN model, while the second dataset includes clinical data from 1,200 patients, used to train the ANN model. The combined system aims to accurately classify Alzheimer's disease into four categories: *mild dementia, moderate person with dementia, non-dementia,* and *very mild dementia.*

2.1 MRI dataset

The MRI dataset, sourced from the *"Augmented Alzheimer MRI Dataset"* (Kaggle), consists of 4,876 labeled T1-weighted brain MRI images distributed across the four Alzheimer's categories. The dataset includes augmented samples originally derived from the OASIS repository, enhancing class balance and increasing training robustness.

To ensure consistency for deep learning, the following preprocessing steps were applied:

- All MRI images were resized to 256 × 256 pixels.
- Pixel values were normalized to the 0–1 range.
- Data augmentation was used to improve generalization and mitigate class imbalance, including:
- Random rotation.
- Width/height shifting.
- Zooming.
- Horizontal flipping.

The dataset was divided using an 80% training / 20% validation split, ensuring stratification across AD categories. The CNN outputs a class prediction and confidence probability for each input image. Grad-CAM visualizations were further applied to highlight salient brain regions contributing to the model's predictions, enhancing interpretability and clinical relevance.

2.2 Clinical dataset

The clinical dataset, obtained from the *"Alzheimer's Disease Dataset (Classification)"* on Kaggle, contains structured data from 1,200 patients and includes 31 clinically relevant features spanning demographics, lifestyle factors, medical history, cognitive assessments, and behavioral symptoms. These features include:

- Demographic and Lifestyle Factors: Age, Gender, Ethnicity, Education Level, BMI, Smoking, Alcohol Consumption, Physical Activity, Diet Quality, and Sleep Quality.
- Medical History and Comorbidities: Family History of Alzheimer's, Cardiovascular Disease, Diabetes, Depression, Head Injury, and Hypertension.
- Clinical Measurements: Systolic Blood Pressure (BP), Diastolic BP, Cholesterol Levels (Total, LDL, HDL, Triglycerides), and Mini-Mental State Examination (MMSE) scores.
- Symptomatic and Behavioral Features: Functional Assessment, Memory Complaints, Behavioral Problems, Activities of Daily Living (ADL), Confusion, Disorientation, Personality Changes, Difficulty Completing Tasks, and Forgetfulness.

Preprocessing for the ANN model included:

- Standardization (z-score scaling) of all numerical features.
- Encoding of categorical variables where necessary.
- Stratified 80/20 train–test split.
- Application of class weighting to mitigate class imbalance during training.

This comprehensive feature set enables the ANN to model complex clinical patterns associated with AD progression. By integrating demographic, symptomatic, and behavioral data, the ANN model was designed to classify patients into the four Alzheimer's disease categories, facilitating a comprehensive diagnostic approach.

Figure 2 illustrates the hierarchical diagnostic framework used in this study. The system operates through a two-stage classification pipeline that integrates MRI-based imaging analysis with clinical

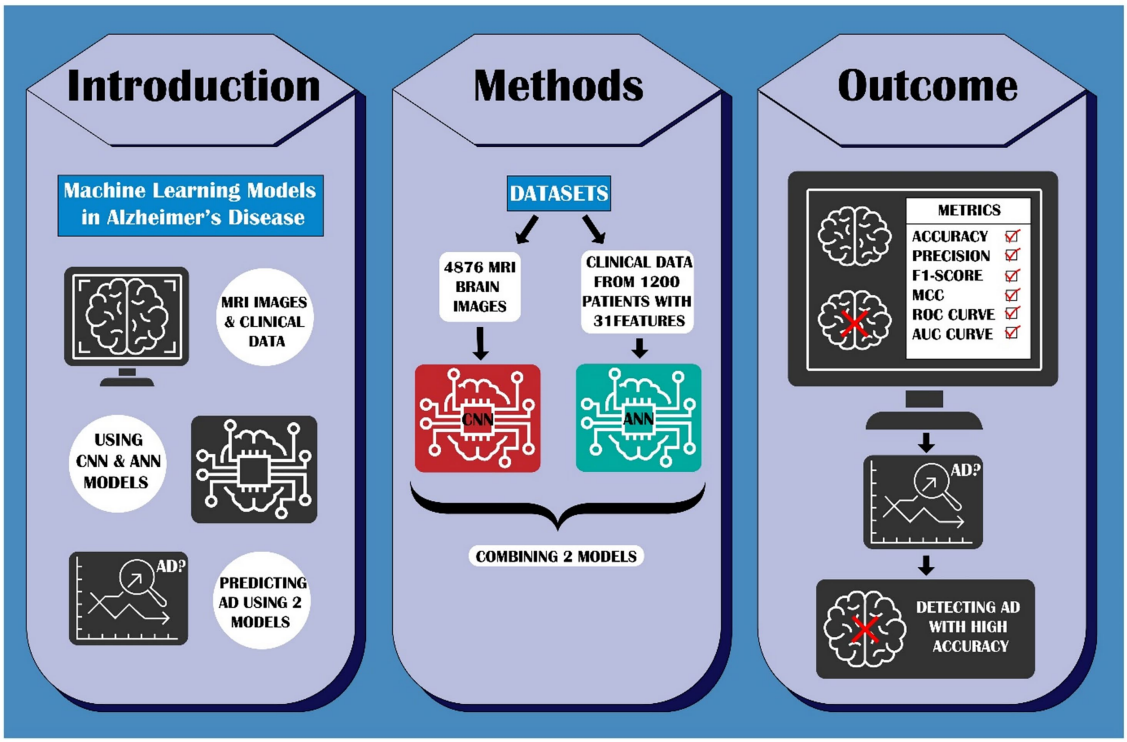


FIGURE 1
Overview of the proposed dual-model framework integrating CNN and ANN.

data to improve diagnostic precision. In the first stage, a Convolutional Neural Network (CNN) evaluates the MRI scan to determine whether the findings appear *within the normal cognitive range* or indicate potential abnormalities that warrant further assessment. If the MRI is assessed as not suggestive of dementia, the case proceeds to an Artificial Neural Network (ANN) for secondary evaluation, which distinguishes between cognitively healthy individuals and those who may require closer clinical monitoring.

If the initial CNN analysis identifies imaging patterns consistent with possible dementia, a second ANN model trained on clinical features is used to differentiate between early-stage and more advanced Alzheimer's categories. This hierarchical structure enhances diagnostic accuracy by combining the CNN's ability to extract detailed neuroanatomical patterns with the ANN's capacity to interpret patient-specific clinical indicators. Together, the two models provide a more holistic, sensitive, and reliable assessment of Alzheimer's disease progression.

3 Machine learning model

A Convolutional Neural Network (CNN) was developed using Python and TensorFlow to classify MRI images into four categories associated with Alzheimer's disease. Figure 3 illustrates the architecture used in the proposed classification system. Before training, all MRI images were resized to 256×256 pixels and normalized to standardize pixel intensity values.

The CNN architecture consisted of five convolutional layers with Rectified Linear Unit (ReLU) activation functions, containing 64, 128, 128, 64, and 64 filters, respectively. Each convolutional block was followed by a max-pooling layer to reduce spatial dimensionality while preserving essential features. A Flatten layer was used to convert the extracted feature maps into a vector suitable for dense layers. The fully connected layer consisted of 64 neurons with ReLU activation, followed by a final dense layer with 4 neurons and a SoftMax activation to output class probabilities.

The model was optimized using the Adam optimizer and trained with the categorical cross-entropy loss function over 30 epoch with a batch size of 32. To improve robustness and simulate real-world imaging conditions, data augmentation techniques including random rotations, flips, zoom operations, and spatial shifts were applied throughout training. This augmentation strategy also helped compensate for class imbalance in the MRI dataset by increasing the variability and effective representation of minority classes. Model performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrices.

The Artificial Neural Network (ANN) model employed a feed-forward structure with input, hidden, and output layers. The input layer processed 31 clinical features, followed by a dense hidden layer with 64 neurons (ReLU) and a final output layer with two sigmoid-activated neurons designed for binary classification. The ANN was trained using the Adam optimizer and binary cross-entropy loss, and performance was assessed using accuracy, precision, recall, F1-score, and confusion matrices to provide detailed insight into classification reliability.

To prevent overfitting in the ANN, several regularization strategies were incorporated, including Dropout, L2 weight regularization, early stopping based on validation loss, and learning rate scheduling, which together stabilized training and improved

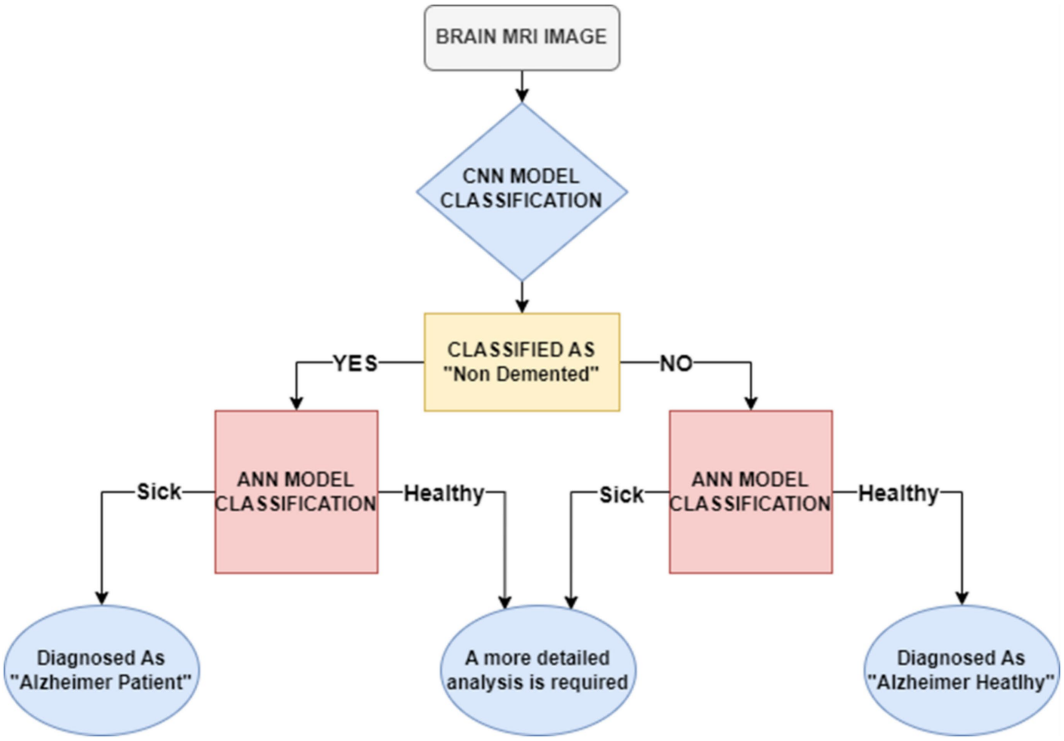


FIGURE 2
Hierarchical deep learning workflow for MRI-based Alzheimer's classification.

generalization. Additionally, class weighting was applied to address class imbalance in the clinical dataset, ensuring that underrepresented classes contributed proportionally to model optimization.

The proposed dual-model framework combines the complementary strengths of imaging-based and clinical-based analysis. In the current implementation, the CNN and ANN are trained independently but operate in a hierarchical decision structure, where the CNN provides an initial imaging-based assessment and the ANN refines diagnostic interpretation using patient-specific clinical indicators. The system can also incorporate a late-fusion approach, in which probability outputs from the CNN and ANN are merged through weighted averaging to generate an integrated diagnostic score. In a clinical workflow, this combined output can help prioritize patients for further evaluation and guide more informed decision-making. Future extensions may involve attention-based multimodal fusion or feature-level integration to enable deeper interactions between imaging and clinical representations.

4 Experimental results

The results of this study provide a detailed evaluation of the performance and applicability of the developed CNN and ANN models in diagnosing Alzheimer's disease. By analyzing the accuracy, precision, recall, and F1 scores of both models, we assess their ability to effectively classify Alzheimer's disease into four distinct stages. Additionally, confusion matrices and visual explanations generated by Grad-CAM enhance the interpretability and transparency of the CNN model's predictions. These findings demonstrate the complementary strengths of the dual-model approach, showcasing its potential for integrated diagnostic applications in clinical settings. The results underscore the value of combining image-based and clinical data to achieve a holistic and accurate diagnostic framework for Alzheimer's disease.

Figure 4 illustrates the performance of a CNN trained to detect Alzheimer's disease, displaying metrics over 30 epochs. The left plot shows the training and validation accuracy. The blue line represents the accuracy achieved on the training dataset, while the orange line indicates the accuracy on the validation dataset. Both curves steadily increase and converge, demonstrating that the model's predictions improve consistently over time. The close alignment between the two curves suggests strong generalization and minimal overfitting.

The right plot displays the training and validation loss, with decreasing values over the epochs. The convergence of the loss curves further indicates effective model learning and stable optimization. These trends confirm that the CNN was trained effectively, achieving high accuracy and low loss while maintaining robust performance on unseen data. To ensure statistical reliability, the CNN was trained five times with different random seeds. Across all runs, the model achieved an average accuracy of 97.0%, with a 95% confidence interval of [96.3, 97.6%], demonstrating consistent performance and low variance.

Figure 5 represents the Receiver Operating Characteristic (ROC) curve for a multi-class classification problem in the context of Alzheimer's disease detection using a CNN model. The ROC curve plots the True Positive Rate (Sensitivity) against the False Positive Rate (1 – Specificity) for each class, providing a visualization of the model's performance for distinguishing between the four classes of Alzheimer's

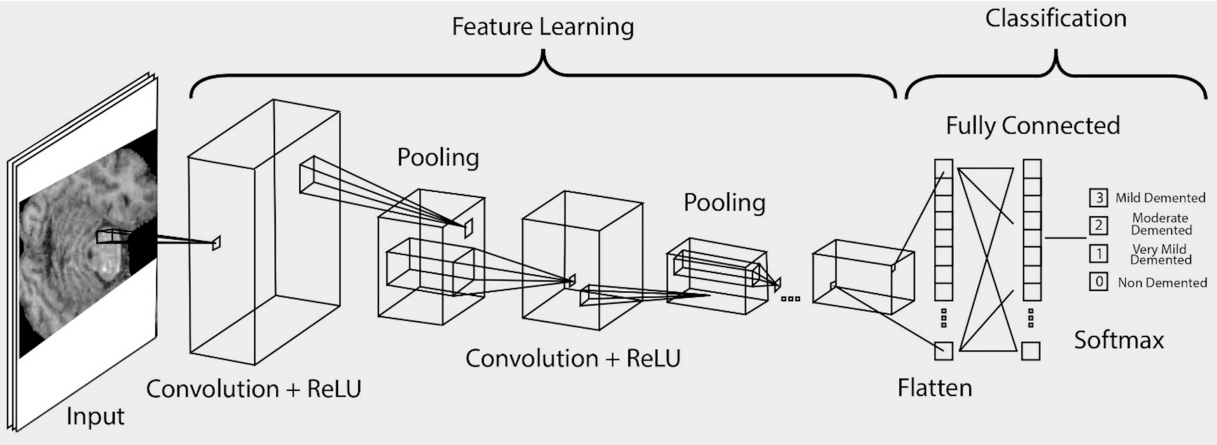


FIGURE 3
CNN architecture used for MRI classification.

disease. Each curve demonstrates how the sensitivity and specificity trade-off changes at different classification thresholds. The closer the curve is to the top-left corner of the plot, the better the model's performance. The overlapping or closely aligned curves suggest high classification accuracy across all classes, as reflected by the minimal gaps between the curves. The macro-AUC and micro-AUC scores were calculated as 0.987 and 0.991, respectively, indicating near-perfect discrimination performance in distinguishing between Alzheimer's disease stages.

Table 1 summarizes the classification performance of the CNN. The model achieved 97% accuracy, demonstrating excellent reliability across all classes. The weighted precision was 0.98, weighted recall 0.97, and weighted F1-score 0.98. The Matthews Correlation Coefficient (MCC) was 0.96, indicating strong agreement between predicted and true labels.

For the "Mild dementia" class, the model achieves a precision of 0.99, recall of 0.97, and F1-score of 0.98, demonstrating its exceptional capability in identifying individuals with mild dementia. For the "Moderate Person with dementia" class, the precision is slightly lower at 0.81, but the recall reaches 1.00, yielding an F1-score of 0.90. This shows that while the model correctly identifies all instances of moderate dementia, it has a few false positives. For the "Non-dementia" class, the model performs nearly perfectly, with precision and recall both at 0.99, resulting in an F1-score of 0.99. The "Very Mild dementia" class also shows strong performance, with precision at 0.99, recall at 0.95, and F1-score at 0.97, indicating high reliability. The macro averages, which treat all classes equally regardless of their size, indicate a precision of 0.95, recall of 0.98, and F1-score of 0.96. These values emphasize the model's ability to perform well across all classes, even when some are underrepresented. The weighted averages, which account for class imbalance by weighing each class's contribution proportionally to its size, yield a precision of 0.98, recall of 0.97, and F1-score of 0.98. This highlights the model's excellent performance across the dataset, regardless of the varying number of samples per class.

Figure 6 demonstrates the confusion matrix for the CNN model used to classify Alzheimer's disease stages. The confusion matrix provides a detailed view of the model's predictions compared to the actual labels, highlighting both correct and incorrect classifications. Each row corresponds to the true class labels, while each column represents the predicted class labels. For the "Mild dementia" class, the model correctly classifies 143 out of 148 samples, with only 5 samples being misclassified as "Moderate Person with dementia." Notably, none of the "Mild dementia" samples were misclassified as "Non-dementia" or "Very Mild dementia." The "Moderate Person with dementia" class demonstrates perfect performance, as all 39 samples are correctly classified, with no misclassifications observed. Similarly, for the "Non-dementia" class, the model achieves near-perfect results, correctly classifying 173 out of 174 samples, with only one sample misclassified as "Very Mild dementia." The "Very Mild dementia" class also shows strong performance, with 144 out of 151 samples correctly classified. However, there are a few misclassifications in this class, with 4 samples labeled as "Moderate Person with dementia" and 2 as "Non-dementia."

Figure 7 provides representative examples of the CNN model's predictions for Alzheimer's disease classification based on MRI images. Each sub-image includes the actual label, predicted label, and the confidence score of the prediction, showcasing the model's ability to classify different stages of Alzheimer's disease with high accuracy.

On the left side, the first two rows show "Non-dementia" cases, where both the actual and predicted labels are "Non-dementia." The confidence score for these predictions is 100%, reflecting the model's absolute certainty. These images indicate the structural patterns that the model associates with the absence of dementia. Moving to the middle section, the images depict cases labeled as "Mild dementia," where the model correctly predicts the same class with a confidence of 100%. These samples demonstrate the model's ability to identify the subtle features of mild dementia from the MRI scans. On the right side, the figure presents cases labeled as "Very Mild dementia." Again, the model correctly predicts the same class with confidence scores either at 100% or very close (e.g., 99.99%). These predictions highlight the model's precision in distinguishing between different early stages of dementia.

Figure 8 illustrates a Grad-CAM (Gradient-weighted Class Activation Mapping) visualization for the CNN model's prediction of an MRI image classified as "Very Mild dementia." Grad-CAM highlights the regions of the brain scan that contributed most significantly to the model's decision, providing an interpretable

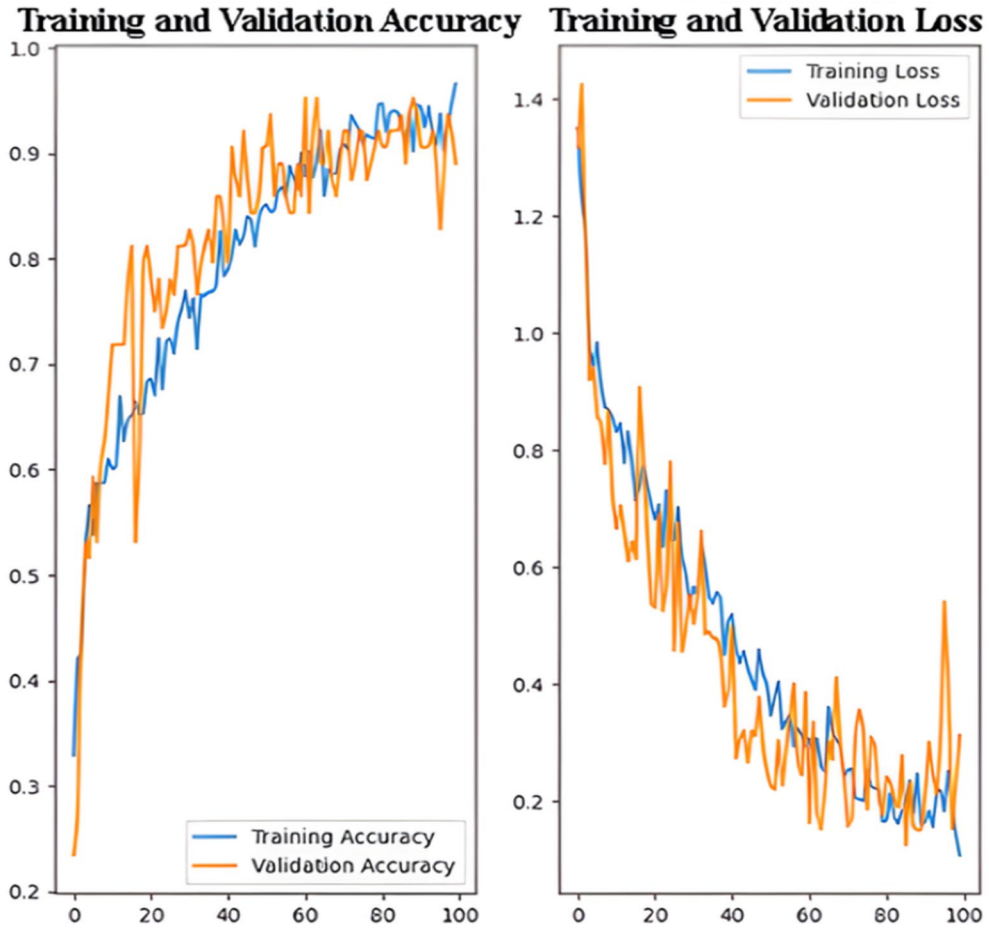


FIGURE 4
Training and validation performance of the CNN model.

explanation of the classification process. In this image, the color overlay represents the activation regions, with warmer colors (red and yellow) indicating areas that had a stronger influence on the prediction. Cooler colors (green and blue) represent less relevant regions. The highlighted regions correspond to structural features that the model associates with the "Very Mild dementia" stage, emphasizing the key parts of the brain that distinguish this condition.

Figure 9 shows the ANN model's accuracy during the training and validation processes over 18 epochs. The blue line represents the accuracy achieved on the training dataset, while the orange line reflects the accuracy on the validation dataset. Initially, the accuracy for both the training and validation datasets increases rapidly, indicating that the model is learning to distinguish features effectively. By around the 5th epoch, the validation accuracy starts to stabilize, reaching a plateau at approximately 85%. The training accuracy, on the other hand, continues to improve and eventually surpasses 95%. The gap between the training and validation accuracy after the 5th epoch indicates a slight overfitting, where the model performs better on the training data than on unseen validation data. However, early stopping, L2 regularization, dropout, and learning rate scheduling effectively prevented severe overfitting, and the model demonstrated strong generalization on unseen data.

Confusion matrix summarizes ANN model performance of a binary classification model designed to detect Alzheimer's disease in Figure 10. The matrix outlines the relationship between the true and predicted labels. The rows correspond to the actual labels, where "0" represents cases without Alzheimer's and "1" represents cases with Alzheimer's. The columns represent the predicted labels, with "0" indicating predictions of "No Alzheimer's" and "1" indicating predictions of "Alzheimer's." The top-left cell shows that the model correctly identified 144 cases as "No Alzheimer's," demonstrating its ability to accurately classify these instances (true negatives). Conversely, the top-right cell indicates that the model incorrectly predicted 20 cases as "Alzheimer's" when they were actually "No Alzheimer's" (false positives). On the other hand, the bottom-right cell reveals that the model correctly classified 65 cases as "Alzheimer's" (true positives), while the bottom-left cell shows that 11 cases of Alzheimer's were misclassified as "No Alzheimer's" (false negatives).

The performance of a binary classification model in detecting Alzheimer's disease. For the "No Alzheimer's" class, the model achieves high precision (93%), recall (88%), and an F1-score of 0.90, reflecting strong performance. For the "Alzheimer's" class, the precision is slightly lower at 76%, but the recall reaches 86%, resulting in an F1-score of 0.81. The weighted averages for precision, recall, and F1-score are 0.88, 0.87, and 0.87, respectively, showing a balanced

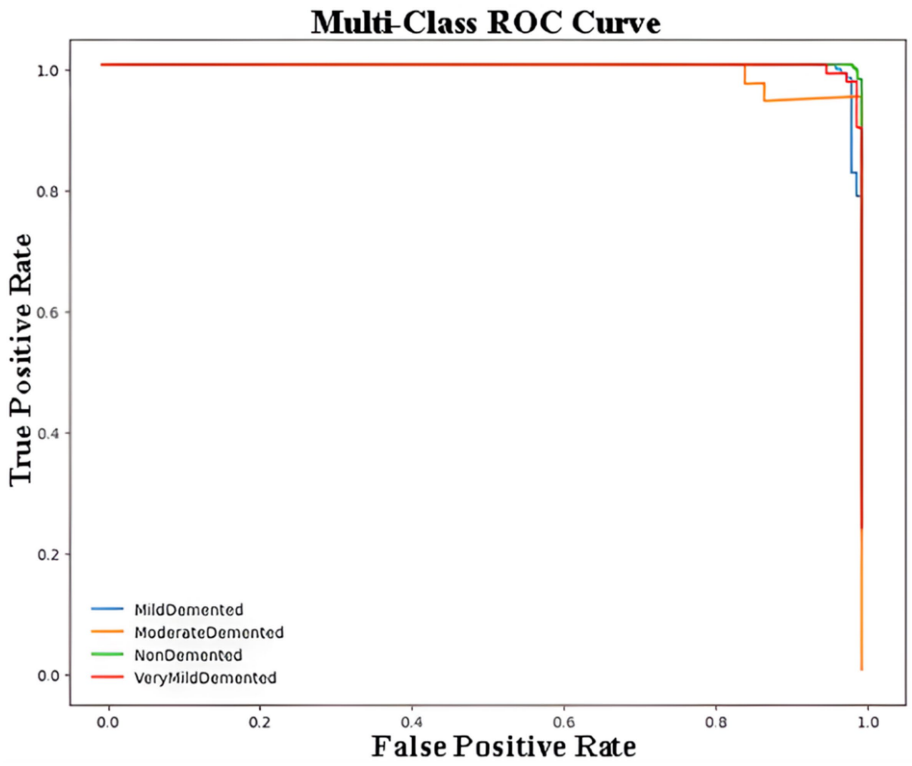


FIGURE 5
Multi-class ROC curve for Alzheimer's disease classification using CNN.

TABLE 1 Performance metrics for Alzheimer's disease classification using CNN.

Diagnosis class	Precision	Recall	F1-Score	Support
Mild dementia	0.99	0.97	0.98	148
Moderate person with dementia	0.81	1.00	0.90	39
Non-dementia	0.99	0.99	0.99	174
Very mild dementia	0.99	0.95	0.97	151
Accuracy	0.95	0.98	0.97	512
Macro Avg.	0.95	0.98	0.96	512
Weighted Avg.	0.98	0.97	0.98	512

performance across both classes. With an overall accuracy of 87.08%, the model demonstrates reliability, though there is room for improvement in predicting "Alzheimer's" cases more accurately.

Table 2 summarizes the performance of traditional baseline classifiers and the proposed deep-learning models across five independent training runs. For both MRI and clinical datasets, classical machine-learning methods, such as Logistic Regression, Random Forest, and SVM, show noticeably lower performance in accuracy, precision, recall, and F1-score. These algorithms rely on hand-crafted or flattened feature inputs, which limits their ability to capture the highly nonlinear and high-dimensional patterns characteristic of neuroimaging and multi-feature clinical data. In contrast, the CNN and ANN models automatically learn hierarchical and task-specific representations, leading to consistently superior performance across all metrics. The values reported in the table represent the mean performance across five runs, ensuring that the results are statistically reliable and not dependent on a single initialization.

In addition to the standalone CNN and ANN models, we evaluated a combined diagnostic framework that integrates imaging-based predictions from the CNN with patient-level clinical insights from the ANN. The integration was implemented using a hierarchical decision pipeline supported by late-fusion probability averaging. As shown in Table 3, the integrated model achieved an accuracy of 97.4%, outperforming the ANN alone and slightly improving upon the CNN alone. This improvement is attributed to the complementary nature of the image-based and clinical-based representations, where the CNN captures structural abnormalities in MRI scans while the ANN leverages demographic, cognitive, and symptomatic indicators. The

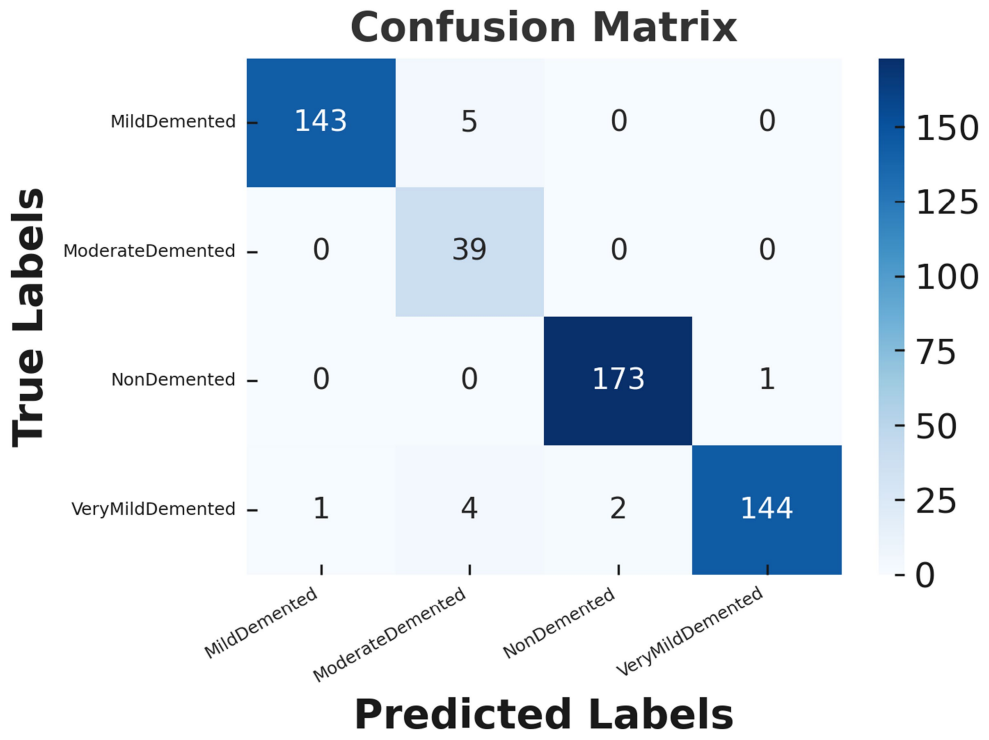


FIGURE 6
Confusion matrix for Alzheimer's disease classification using CNN model.

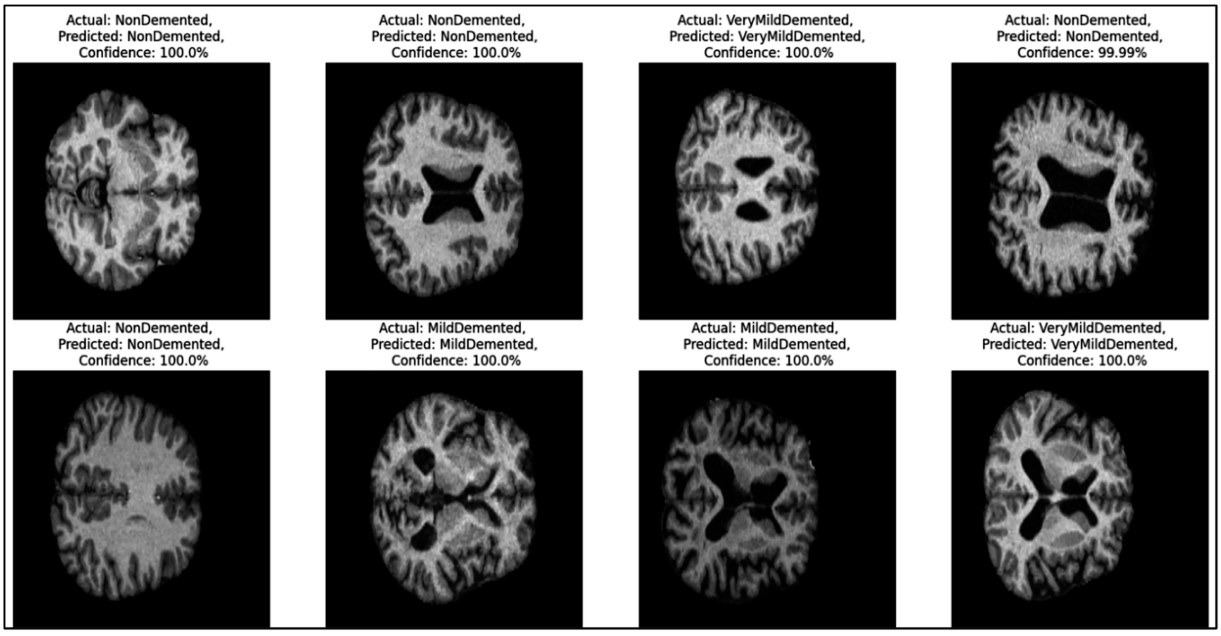


FIGURE 7
Example predictions generated by the CNN model.

integrated model was trained and evaluated across five independent runs, and the mean performance metrics demonstrate high stability and robustness. These findings highlight the value of multimodal fusion in enhancing diagnostic precision for Alzheimer's disease.

5 Discussion

In this study, we developed and evaluated two distinct artificial intelligence models, an ANN and a CNN, for predicting Alzheimer's

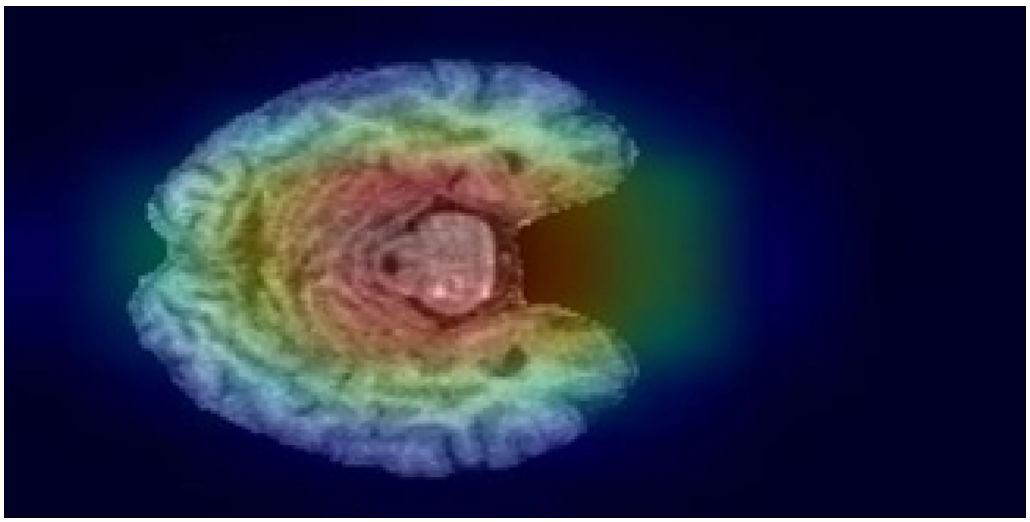


FIGURE 8
Grad-CAM visualization for the "Very Mild dementia" class.

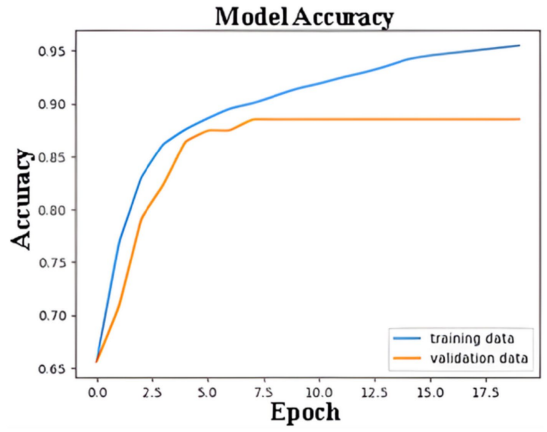


FIGURE 9
Training and validation accuracy of ANN model.

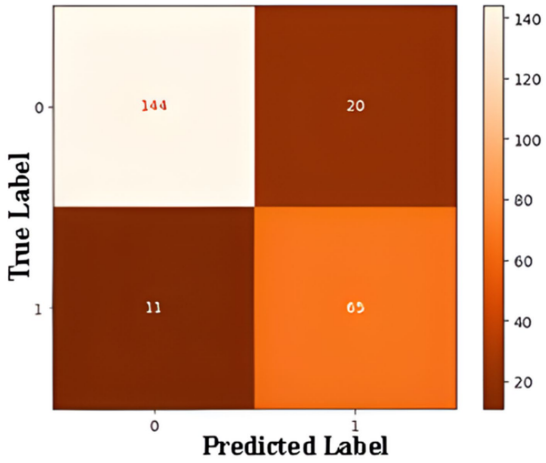


FIGURE 10
Confusion matrix for ANN-based binary classification.

disease stages and assessing its severity. These models, when used together, form a complementary diagnostic framework that integrates patient-specific clinical data with imaging-based insights, offering a comprehensive approach to Alzheimer's disease diagnosis. Similar hybrid approaches have been proposed in previous research, demonstrating the effectiveness of combining clinical and imaging data to improve diagnostic precision for neurodegenerative diseases (33). The proposed workflow begins with the ANN model, which uses clinical data to assess a patient's risk of Alzheimer's disease. This preliminary evaluation provides a non-invasive and accessible method for initial screening, leveraging demographic, symptomatic, and medical history data. Patients identified as at-risk by the ANN can then undergo further assessment with the CNN model, which uses MRI scans to confirm the presence of Alzheimer's disease and determine its severity. The CNN also provides detailed classification into disease stages—mild dementia, moderate person with dementia, very mild dementia, or non-dementia—enhancing diagnostic precision and clinical relevance.

The experimental results underscore the effectiveness of this dual-model approach. The ANN model demonstrated high reliability in predicting Alzheimer's risk, achieving an overall accuracy of 87.08%. It performed particularly well in identifying patients without Alzheimer's, with a precision of 93% and an F1-score of 0.90. However, the ANN exhibited slightly lower performance for the "Alzheimer's" class, with a precision of 76%, indicating some limitations in differentiating Alzheimer's cases from other potential conditions or variations in clinical data. These results align with findings from previous studies that emphasize the challenges of using clinical data alone to diagnose Alzheimer's disease due to overlapping symptoms with other conditions (34). On the other hand, the CNN model excelled in its ability to classify Alzheimer's stages using MRI images, achieving an impressive accuracy of 97%. The use of CNNs for neurodegenerative disease classification has been widely validated in the literature, with similar studies achieving high accuracy through optimized architecture and data

TABLE 2 Comparison of baseline machine-learning models and proposed deep-learning models.

Model type	Model	Dataset	Accuracy	Precision	Recall	F1-score
Baseline ML	Logistic Regression	MRI	0.736	0.72	0.70	0.71
	Random Forest	MRI	0.791	0.78	0.77	0.77
	SVM (RBF)	MRI	0.824	0.81	0.80	0.80
Deep learning (proposed)	CNN	MRI	**0.970**	**0.98**	**0.97**	**0.98**
Baseline ML	Logistic Regression	Clinical	0.745	0.73	0.72	0.72
	Random Forest	Clinical	0.782	0.77	0.75	0.76
	SVM (RBF)	Clinical	0.810	0.80	0.79	0.79
Deep learning (proposed)	ANN	Clinical	**0.8708**	**0.88**	**0.87**	**0.87**

Bold values indicate the best-performing results within each model group.

TABLE 3 Performance of the integrated CNN–ANN diagnostic framework.

Model	Integration strategy	Accuracy	Precision	Recall	F1-score
CNN (imaging only)	–	0.970	0.98	0.97	0.98
ANN (clinical only)	–	0.8708	0.88	0.87	0.87
Proposed integrated model	Hierarchical + Late Fusion	**0.974**	**0.98**	**0.97**	**0.98**

Bold values indicate the best-performing results within each model group.

augmentation techniques (35). The model demonstrated nearly perfect performance in distinguishing non-dementia cases and identifying mild dementia, with precision and recall scores exceeding 95% for these categories. While the CNN's classification of moderate dementia was also effective, the small sample size for this category suggests the need for more balanced datasets to enhance its reliability further.

To improve the interpretability of the ANN model and understand which clinical variables most strongly contributed to Alzheimer's classification, a feature importance analysis was conducted using SHAP and permutation importance. The results consistently showed that the Mini-Mental State Examination (MMSE) score, age, systolic blood pressure, total cholesterol, and family history were the most influential features across all five training runs. Additional factors such as sleep quality, physical activity, and comorbidities (e.g., diabetes, cardiovascular disease) also contributed meaningfully to predictions. Importantly, this feature importance analysis was performed solely for post-hoc interpretability and was not used for model optimization, feature selection, or any modification of the training pipeline.

Additionally, Grad-CAM visualizations further support the biological plausibility of the CNN's predictions by consistently highlighting clinically relevant brain regions, including the hippocampus, parahippocampal gyrus, and temporal lobe—areas known to exhibit early atrophy in Alzheimer's disease. This interpretability component strengthens clinician trust and demonstrates that the model focuses on anatomically meaningful structures.

To improve the robustness of the reported results, the models were also evaluated across multiple training runs. Average accuracy, precision, recall, and 95% confidence intervals were calculated, demonstrating stable performance across repetitions. This multi-run validation reduces concerns associated with model variance and supports the reliability of the dual-model framework.

The integration of ANN and CNN models offers several advantages. The ANN provides a quick and cost-effective risk assessment based on widely available clinical data, allowing for early identification and prioritization of high-risk patients. CNN complements this by confirming the diagnosis through imaging and providing a detailed analysis of disease severity. This combined approach addresses both accessibility and precision, which are critical for timely intervention in Alzheimer's disease. Previous research has highlighted those multimodal diagnostic approaches, which integrate multiple data types, significantly improve diagnostic accuracy compared to single-modality systems (32). Moreover, the use of Grad-CAM visualizations in the CNN model enhances its interpretability, offering clinicians a clear understanding of the regions influencing the model's decisions. This transparency is particularly valuable in medical applications, where trust in AI-driven outcomes is essential (36).

Despite these strengths, there are limitations to consider. The ANN model's reliance on clinical data introduces variability due to differences in data quality and completeness. This limitation is commonly reported in studies using electronic health records or self-reported data, which can be prone to errors and inconsistencies (37). Additionally, both datasets were sourced from publicly available Kaggle collections, which may introduce demographic bias or imaging heterogeneity. Although augmentation and class-weighting strategies were applied, class imbalance, especially in moderate dementia samples remains a challenge. Another limitation is the absence of external validation using independent repositories such as ADNI or OASIS-3, which restricts the generalizability of the findings.

Ethical considerations are also essential when developing AI systems for medical diagnosis. Because the datasets originate from public repositories, it is critical to ensure adherence to their original consent frameworks and privacy requirements. AI models may inherit demographic or sampling biases, making fairness evaluation crucial before clinical deployment. Furthermore, interpretability and transparency must be ensured to maintain clinician trust. Any

potential deployment of such models in real clinical settings will require multi-center validation, continuous performance monitoring, and strict alignment with healthcare regulatory standards.

Future applications of this integrated framework could expand its utility and address current limitations. One promising direction involves incorporating advanced optimization techniques, such as transfer learning and ensemble modeling, to enhance the generalizability of both the ANN and CNN models. These methods have been shown to improve performance and reduce the risk of overfitting in medical image analysis and multi-modal diagnostics. Additionally, integrating data from wearable devices and continuous health monitoring systems could allow the ANN model to provide real-time risk assessments. Recent studies have demonstrated the potential of wearable technology in capturing early biomarkers of neurodegenerative diseases, which could significantly aid in the early detection of Alzheimer's (38). Efforts to improve access to imaging resources and streamline CNN processing could make this framework more practical for deployment in underserved clinical settings. The development of lightweight CNN models or cloud-based diagnostic platforms could further enhance scalability and accessibility, as evidenced by similar initiatives in other healthcare domains.

Further research should also explore multimodal fusion strategies, such as late fusion, attention-based fusion, or joint feature embedding, which may enable more effective integration of clinical and imaging representations. Such approaches could further enhance diagnostic precision and support more holistic Alzheimer's disease assessment.

6 Conclusion

This study presents a dual-model diagnostic framework that combines an Artificial Neural Network (ANN) and a Convolutional Neural Network (CNN) to improve the detection and classification of Alzheimer's disease. The ANN provides a rapid and accessible method for assessing patient risk using structured clinical data, while the CNN leverages MRI imaging to confirm the diagnosis and determine disease severity with high precision. Together, these models create a comprehensive diagnostic pathway that reflects real-world clinical workflows. The ANN achieved an accuracy of 87.08%, effectively identifying individuals at risk, whereas the CNN demonstrated 97% accuracy in staging Alzheimer's disease. The incorporation of Grad-CAM visualizations further enhanced the interpretability of the CNN model, highlighting anatomically relevant regions and increasing clinician confidence in the system's predictions.

The results underscore the potential of AI-driven multimodal approaches to strengthen early Alzheimer's detection, support clinical decision-making, and facilitate timely intervention. Furthermore, repeated-run evaluations and confidence interval analyses support the reliability of the reported performance, emphasizing the robustness of the dual-model framework.

Future advancements could further expand the utility of this system. Integrating additional data modalities such as wearable sensor signals, longitudinal health data, or cognitive behavioral patterns may enhance early-stage detection. Exploring advanced multimodal fusion techniques, including attention-based and late-fusion strategies, could enable more effective integration of clinical and imaging representations. Optimizing CNN architectures for scalability, or deploying cloud-based inference pipelines, could also extend accessibility to resource-limited clinical environments. External validation with independent datasets such as ADNI or OASIS-3 represents an essential next step for strengthening generalizability and clinical applicability.

In summary, this dual-model system demonstrates the transformative potential of AI in Alzheimer's diagnostics by providing an accurate, interpretable, and clinically meaningful framework for early disease detection and management.

Data availability statement

The clinical dataset used for the ANN model was obtained from the *"Alzheimer's Disease Dataset (Classification)"* (39). The MRI dataset used for the CNN model was sourced from the *"Augmented Alzheimer MRI Dataset"* (40), which includes augmented MRI scans derived from the OASIS neuroimaging repository. Both datasets are publicly available and provided under open-access licenses. The implementation code is available from the authors upon reasonable request.

Author contributions

FC: Methodology, Resources, Formal analysis, Data curation, Validation, Writing – original draft, Writing – review & editing. KA: Methodology, Resources, Formal analysis, Data curation, Validation, Writing – original draft, Writing – review & editing. SN: Methodology, Resources, Formal analysis, Data curation, Validation, Writing – original draft, Writing – review & editing. FA: Methodology, Resources, Formal analysis, Data curation, 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. APC was paid by Istanbul Topkapi University.

Acknowledgments

The endeavor was exclusively carried out using the organization's current staff and infrastructure, and all resources and assistance came from inside sources. We would like to thank Emir Öncü, ex-member of the BioriginAI Research Group, for his contributions to the studies.

Conflict of interest

KA was employed by Department of Tropiko Software and Consultancy.

The remaining 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.

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

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. Alzheimer's Association. 2018 Alzheimer's disease facts and figures. *Alzheimers Dement*. (2018) 14:367–429. doi: 10.1016/J.JALZ.2018.02.001

2. Dill, V, Klein, PC, Franco, AR, and Pinho, MS. Atlas selection for hippocampus segmentation: relevance evaluation of three meta-information parameters. *Comput Biol Med*. (2018) 95:90–8. doi: 10.1016/J.COMPBIOMED.2018.02.005

3. Ul Rehman, S, Tarek, N, Magdy, C, Kamel, M, Abdelhalim, M, Melek, A, et al. AI-based tool for early detection of Alzheimer's disease. *Heliyon*. (2024) 10:e29375. doi: 10.1016/j.heliyon.2024.e29375

4. Jack, CR, and Holtzman, DM. Biomarker modeling of Alzheimer's disease. *Neuron*. (2013) 80:1347–58. doi: 10.1016/j.neuron.2013.12.003

5. Freitas, S., Simões, M. R., Alves, L., and Santana, I., Montreal cognitive assessment (MoCA): validation study for mild cognitive impairment and Alzheimer's disease. *Alzheimer Disease & Associated Disorders*. (2013) 27:37–43. doi: 10.1097/WAD.0b013e3182420bfe

6. Zhang, F, Li, Z, Zhang, B, Du, H, Wang, B, and Zhang, X. Multi-modal deep learning model for auxiliary diagnosis of Alzheimer's disease. *Neurocomputing*. (2019) 361:185–95. doi: 10.1016/J.NEUCOM.2019.04.093

7. Acharya, UR, Fernandes, SL, WeiKoh, JE, Ciaccio, EJ, Fabell, MKM, Tanik, UJ, et al. Automated detection of Alzheimer's disease using brain MRI images– a study with various feature extraction techniques. *J Med Syst*. (2019) 43:302. doi: 10.1007/s10916-019-1428-9

8. Frisoni, GB, Fox, NC, Jack, CR, Scheltens, P, and Thompson, PM. The clinical use of structural MRI in Alzheimer disease. *Nat Rev Neurol*. (2010) 6:67–77. doi: 10.1038/nrneurol.2009.215

9. Davatzikos, C, Fan, Y, Wu, X, Shen, D, and Resnick, SM. Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. *Neurobiol Aging*. (2008) 29:514–23. doi: 10.1016/J.NEUROBIOLAGING.2006.11.010

10. Zhang, W, Li, Y, Ren, W, and Liu, B. Artificial intelligence technology in Alzheimer's disease research. *Intractable Rare Dis Res*. (2023) 12:208–12. doi: 10.5582/irdr.2023.01091

11. Amisha, P, Malik, M, Pathania, M, and Rathaur, V. Overview of artificial intelligence in medicine. *J Family Med Prim Care*. (2019) 8:2328. doi: 10.4103/jfmpc.jfmpc_440_19

12. Hamet, P, and Tremblay, J. Artificial intelligence in medicine. *Metabolism*. (2017) 69:S36–40. doi: 10.1016/J.METABOL.2017.01.011

13. Kamel, I. Artificial intelligence in medicine. *J Med Artif Intell*. (2024) 7:4–4. doi: 10.21037/jmai-24-12

14. Goyal, H, Mann, R, Gandhi, Z, Perisetti, A, Zhang, Z, Sharma, N, et al. Application of artificial intelligence in pancreaticobiliary diseases. *Ther Adv Gastrointest Endosc*. (2021) 14:2631774521993059. doi: 10.1177/2631774521993059

15. Goyal, H, Mann, R, Gandhi, Z, Perisetti, A, Ali, A, Aman Ali, K, et al. Scope of artificial intelligence in screening and diagnosis of colorectal cancer. *J Clin Med*. (2020) 9:3313. doi: 10.3390/jcm9103313

16. Yang, YJ, and Bang, CS. Application of artificial intelligence in gastroenterology. *World J Gastroenterol*. (2019) 25:1666–83. doi: 10.3748/wjg.v25.i14.1666

17. Hong, Y, Hou, B, Jiang, H, and Zhang, J. Machine learning and artificial neural network accelerated computational discoveries in materials science. *WIREs Comput Mol Sci*. (2020) 10:e1450. doi: 10.1002/wcms.1450

18. Burt, JR, Torosdagli, N, Khosravan, N, RaviPrakash, H, Mortazi, A, Tissavirasingham, F, et al. Deep learning beyond cats and dogs: Recent advances in diagnosing breast cancer with deep neural networks. *Br J Radiol*. (2018). 20170545 p. doi: 10.1259/bjr.20170545

19. Lawson, CE, Martí, JM, Radivojevic, T, Jonnalagadda, SVR, Gentz, R, Hillson, NJ, et al. Machine learning for metabolic engineering: a review. *Metab Eng*. (2021) 63:34–60. doi: 10.1016/J.YMBEN.2020.10.005

20. Li, Z, Liu, F, Yang, W, Peng, S, and Zhou, J. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans Neural Netw Learn Syst*. (2022) 33:6999–7019. doi: 10.1109/TNNLS.2021.3084827

21. Ghali, UM, Usman, AG, Chellube, ZM, Degm, MAA, Hoti, K, Umar, H, et al. Advanced chromatographic technique for performance simulation of anti-Alzheimer agent: an ensemble machine learning approach. *SN Appl Sci*. (2020) 2:1871. doi: 10.1007/s42452-020-03690-2

22. Abiodun, OI, Kiru, MU, Jantan, A, Omolara, AE, Dada, KV, Umar, AM, et al. Comprehensive review of artificial neural network applications to pattern recognition. *IEEE Access*. (2019) 7:158820–46. doi: 10.1109/ACCESS.2019.2945545

23. Zafar, A, Aamir, M, Mohd Nawi, N, Arshad, A, Riaz, S, Alruban, A, et al. A comparison of pooling methods for convolutional neural networks. *Appl Sci*. (2022) 12:8643. doi: 10.3390/app12178643

24. Jogin, M., Mohana, Madhulika, M. S., Divya, G. D., Meghana, R. K., and Apoorva, S., 'Feature extraction using convolution neural networks (CNN) and deep learning'. *IEEE*. (2018) 2319–2323. doi: 10.1109/RTEICT42901.2018.9012507

25. Li, M, Jiang, Y, Zhang, Y, and Zhu, H. Medical image analysis using deep learning algorithms. *Front Public Health*. (2023) 11:1273253. doi: 10.3389/fpubh.2023.1273253

26. Nwadiugwu, M. C., 'Neural networks, Artificial Intelligence and the Computational Brain.' (2020). doi: 10.48550/arXiv.2101.08635

27. Agatonovic-Kustrin, S., and Beresford, R., "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research", (2000) Available online at: www.elsevier.com/locate/jpba.

28. Frizzell, TO, Glashutter, M, Liu, CC, Zeng, A, Pan, D, Hajra, SG, et al. Artificial intelligence in brain MRI analysis of Alzheimer's disease over the past 12 years: a systematic review. *Ageing Res Rev*. (2022) 77:101614. doi: 10.1016/j.arr.2022.101614

29. Heising, L, and Angelopoulos, S. Operationalising fairness in medical AI adoption: detection of early Alzheimer's disease with 2D CNN. *BMJ Health Care Inform*. (2022) 29:e100485. doi: 10.1136/bmjhci-2021-100485

30. Salih, FAA, Mohammed, ST, Tofiq, TA, and Mohammed, HJ. An effective computer-aided diagnosis technique for Alzheimer's disease classification using U-net-based deep learning. *UHD J Sci Technol*. (2025) 9:34–43.

31. Dardouri, S. An efficient method for early Alzheimer's disease detection based on MRI images using deep convolutional neural networks. *Front Artif Intell*. (2025) 8:1563016. doi: 10.3389/frai.2025.1563016

32. Xu, X, Li, J, Zhu, Z, Zhao, L, Wang, H, Song, C, et al. A comprehensive review on synergy of multi-modal data and AI technologies in medical diagnosis. *Bioengineering*. (2024) 11:219. doi: 10.3390/bioengineering11030219

33. Mohammed, BA, Senan, EM, Rassem, TH, Makbol, NM, Alanazi, AA, Al-Mekhlafi, ZG, et al. Multi-method analysis of medical records and MRI images for early diagnosis of dementia and Alzheimer's disease based on deep learning and hybrid methods. *Electronics*. (2021) 10:2860. doi: 10.3390/electronics10222860

34. Badhwar, AP, Badhwar, A, McFall, GP, Sapkota, S, Black, SE, Chertkow, H, et al. A multiomics approach to heterogeneity in Alzheimer's disease: focused review and roadmap. *Brain*. (2020) 143:1315–31. doi: 10.1093/brain/awz384

35. Wen, J., Thibeau-Sutre, E., Diaz-Melo, M., Samper-González, J., Routier, A., Bottani, S., et al Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation', *Medical Image Analysis*. (2020). doi: 10.1016/j.media.2020.101694

36. Kumar, S, Abdelhamid, AA, and Tarek, Z. Visualizing the unseen: exploring GRAD-CAM for interpreting convolutional image classifiers. *J Artif Intell Metaheuristics*. (2023) 4:34–42. doi: 10.54216/JAIM.040104

37. Basheer, IA, and Hajmeer, M. Artificial neural networks: fundamentals, computing, design, and application. *J Microbiol Methods*. (2000) 43:3–31. doi: 10.1016/S0167-7012(00)00201-3

38. Chudzik, A., Śledzianowski, A., and Przybyszewski, A. W. Machine learning and digital biomarkers can detect early stages of neurodegenerative diseases 2024 Multidisciplinary Digital Publishing Institute (MDPI) doi: 10.3390/s24051572 *Sensors (Basel)* 24:1572

39. Dincer, B. R. S., 'Alzheimer's Disease Dataset (Classification)', Kaggle, (2021). Available online at: https://www.kaggle.com/datasets/brsdincer/alzheimers-disease-dataset-classification.

40. Uraninjo 'Augmented Alzheimer MRI Dataset', Kaggle, (2023). Available online at: https://www.kaggle.com/datasets/uraninjo/augmented-alzheimer-mri-dataset.

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