American Journal of Applied Research and AI (AJARAI)

VOLUME 1 ISSUE 2 (2026)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Applied Research and AI (AJARAI))

Volume 1 Issue 2, Year 2026
ISSN: 2158-7205 (Online)
DOI: https://doi.org/10.54536/ajarai.v1i2.7963
https://journals.e-palli.com/home/index.php/ajarai

Comparative Analysis of CNN and Vision Transformer Architectures for Explainable Brain Tumor Classification Using MRI Images

Md. Abu Yousuf[1*], Joy Chowdhury[1], Susmoy Chowdhury[1], Sukanta Dev[1]

Article Information

Received: April 09, 2026
Accepted: June 08, 2026
Published: August 20, 2026

Keywords

Deep Learning (Dl), Explainable Artificial Intelligence (Xai), Magnetic Resonance Imaging (Mri), Vision Transformer (Vit)

ABSTRACT

Diagnosis of brain tumor is one of the most critical tasks in medical imaging and early and accurate diagnosis is extremely important for improving the treatment planning and outcome of patients with brain tumors. Brain tumor assessment is achieved with Magnetic Resonance Imaging (MRI), which is commonly used, but manual interpretation of MRI scans can be time-consuming and subject to inter-observer variability. To overcome these difficulties, an explainable deep learning (XDL) framework is proposed that classifies brain tumors into a multiclass classification problem using magnetic resonance imaging (MRI) images. To overcome these challenges, this study proposes an explainable deep learning (XDL) framework that is capable of brain tumor classification as a multiclass classification problem from MRI images. Five state-of-the-art deep learning architectures ResNet50, VGG16, DenseNet121, Vision Transformer (ViT), and a Hybrid CNN ViT model, were tested and compared to determine which one would be the best fit for the classification task. Preprocessing and augmentation of MRI images were performed before training and evaluation the model. Metrics used to evaluate the performance included accuracy, precision, recall, F1 score, the confusion matrix, and Receiver Operating Characteristic Area Under Curve (ROC-AUC). The Hybrid CNN–ViT framework was compared and it was found that the hybrid approach, combining the advantages of CNN for feature extraction and transformer for contextual learning, led to better classification performance. Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-Agnostic Explanations (LIME) were used to explain the best-performing model to improve the transparency and interpretability of the model. To validate that the model attended to clinically relevant tumour regions in classification, the explainability results were obtained. The proposed framework showed high classification accuracy of 98.68% and an average of ROC-AUC score at approximately 0.999, indicating that the framework has good predictive power and high interpretability. The results indicate that the proposed explainable deep learning framework shows great promise of helping clinical diagnosis of brain tumors to be reliable and transparent, by integrating with AI.

INTRODUCTION

Brain tumors are one of the most serious neurological diseases and are an important source of morbidity and mortality worldwide. They involve the abnormal growth of cells in the brain or other tissue of the central nervous system, and can cause serious problems with thinking, sensory and motor functions. Early diagnosis is crucial to effective planning, better treatment outcomes; delays in diagnosis can substantially affect the survival rates and effectiveness of treatment. Magnetic Resonance Imaging (MRI) is the preferred imaging modality for the diagnosis of brain tumors due to its high soft-tissue contrast, non-invasive nature and the ability to provide detailed anatomical information on tumor location, size and morphology (Louis *et al.*, 2021; Wen *et al.*, 2020). Manual interpretation of MRI scans is difficult, time-consuming, and highly reliant on radiological expertise, which can cause inter-observer variation and inconsistencies in diagnosis (Bakas *et al.*, 2018; Isin *et al.*, 2016). The medical image analysis field has been revolutionized by the recent advances in Artificial Intelligence (AI), specifically deep learning, which have made it possible to automatically extract features and classify complex medical images. One of the deep learning methods has been showing great results in brain tumor classification applications is the Convolutional Neural Networks (CNNs), which is capable of learning hierarchical spatial representations directly from MRI images (LeCun *et al.*, 2015). A number of high performance CNN architectures are successfully used for brain tumour detection and classification, such as VGG16, ResNet50, DenseNet121 etc. which have high diagnostic accuracy and have outperformed traditional machine learning techniques (Simonyan & Zisserman, 2015; He *et al.*, 2016; Huang *et al.*, 2017). The effectiveness of CNN-based approaches in computer-aided diagnosis (CAD) systems of brain tumor has been reported in recent studies with classification accuracy over 95% (Sultan *et al.*, 2019; Sharif *et al.*, 2022). Although CNN is successful, most of the typical CNN designs are local convolution operations, which might lack in capturing long-range dependence and global context information in medical images. Brain tumors frequently have

[1] Department of Computer Science & Engineering, East Delta University, Chittagong, Bangladesh
[*] Corresponding author's email: engrabuyousuf2026@gmail.com

heterogeneous structures, irregular borders and complex spatial patterns that necessitate understanding the tumor in its context for accurate classification. These drawbacks have been overcome by transformer-based architectures, which have recently been successful in computer vision tasks. The Vision Transformers (ViTs) uses self-attention mechanisms to model global relationships between image patches, which allows it to effectively capture local and long-range dependencies (Dosovitskiy *et al.*, 2021). Recently, ViTs have been shown to be competitive and even outperforming the conventional CNN in several medical imaging applications, such as tumor classification and disease diagnosis (Hatamizadeh *et al.*, 2022; Shaker *et al.*, 2024). Apart from the transformer models, recently hybrid models that combine transformer with CNN have received significant attention. Such models take advantage of the local feature extraction power of CNNs and the global context understanding ability of transformers. These hybrid models have proven to be more robust, better featured representation, and more effective classification models than CNN or transformer models alone (Chen *et al.*, 2021; Zhou *et al.*, 2023). Therefore, hybrid CNN–Transformer models have gained significant traction in the field of medical image classification and diagnosis, where they demonstrate superior performance. Thus, hybrid CNN–Transformer architectures have emerged as potential solutions for complex medical image classification and diagnostic tasks, showing enhanced performance. While classification accuracy is crucial for assessing the performance of AI systems, transparency and interpretability are equally paramount in clinical settings. One of the criticisms of deep learning models concerns their ability to make predictions without articulating a process for making decisions. Models like this are often described as a "black box." Interpretability plays a critical role in healthcare, where decisions made based on diagnosis can directly impact patient care and outcomes, and is crucial for establishing trust with clinicians and regulatory acceptance. Thus, Explainable Artificial Intelligence (XAI) has been identified as a major research field for enhancing the explainability of AI-based diagnostic systems. Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro *et al.*, 2016; Selvaraju *et al.*, 2017) are techniques that generate visual explanations by highlighting areas of the image that are most important in making predictions. These methods allow clinicians to check if the model primarily targets clinically relevant tumour areas and increase the trust in the AI diagnosis (Tjoa & Guan, 2020; Samek *et al.*, 2021). Although significant advances have been made in automated brain tumor classification, there are still a number of challenges in the field of brain tumor research. First, many previous works have only considered CNN or Transformer model independently, which does not allow to compare the two directly in the same experimental setting. Second, there is a significant amount of literature that focuses on classification performance and offers a limited amount of work in the areas of interpretability and explainability. Third, studies are relatively scarce that compare the performance of CNNs, Vision Transformers and hybrid CNN–ViT models under a single framework, together with explainability analysis (Zhou *et al.*, 2023; Shaker *et al.*, 2024). It is essential to overcome these limitations in order to create more reliable AI systems that can be seamlessly incorporated into healthcare practices.Addressing these limitations is critical for building more trustworthy AI systems that can be successfully integrated into clinical workflows. To fill these research gaps, in this study, an explainable deep learning framework is proposed for the classification of brain tumor MRI images into more than two classes. A set of five state-of-the-art architectures, including VGG16, ResNet50, DenseNet121, Vision Transformer (ViT) and a Hybrid CNN–ViT model, are implemented and assessed using a standardized experimental setup. The models are evaluated using different metrics such as accuracy, precision, recall, F1 score, and Area Under the Receiver Operating Characteristic Curve (AUC). In addition, the top-performing model is explained visually, using Grad-CAM and LIME, to make clinically relevant explanations of the decision making process. The proposed framework combines comparative deep learning evaluation with explainable AI methods, providing a comprehensive solution for achieving high diagnostic accuracy while also improving transparency. The proposed framework aims to achieve both high diagnostic accuracy and explainability, contributing to the development of reliable and trustworthy AI-assisted systems for brain tumor diagnosis and clinical decision support.

Research Objective

Main Objective

The main objective of this study is to develop and evaluate an explainable deep learning framework for multiclass brain tumor classification using MRI images by integrating convolutional neural networks, vision transformers, and explainable artificial intelligence techniques.

Specific Objectives

1. To evaluate and compare the classification performance of five deep learning architectures, namely ResNet50, VGG16, DenseNet121, Vision Transformer (ViT), and Hybrid CNN–ViT, for multiclass brain tumor MRI classification.
2. To identify the most effective deep learning architecture based on performance metrics including accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC.
3. To investigate the interpretability of the best-performing model using Explainable Artificial Intelligence (XAI) techniques, specifically Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-Agnostic Explanations (LIME).
4. To develop a reliable and transparent AI-assisted

framework capable of supporting brain tumor diagnosis through accurate classification and visual explanation of model predictions.

LITERATURE REVIEW

In recent years, Magnetic Resonance Imaging (MRI) has been the subject of a great deal of research for the automatic classification of brain tumors, supported by the progress of Artificial Intelligence (AI) and Deep Learning (DL) methods. MRI is the most widely used imaging method for the diagnosis of brain tumors because it gives high-resolution anatomical information without exposing the patient to ionizing radiation. But manual assessment of MRI is time consuming, and requires radiological expertise, so automated computer-aided diagnostic systems (CADS) have been developed (Louis *et al.*, 2021; Isin *et al.*, 2016). In the field of deep learning, Convolutional Neural Networks (CNNs) are the most prevalent architecture used for medical image analysis due to their capacity to capture hierarchical feature representations directly from the image data among all the architectures. The CNN-based architectures with deep structures such as VGG16, ResNet50, and DenseNet121 have exhibited great performance in different image classification tasks (Simonyan & Zisserman, 2015; He *et al.*, 2016; Huang *et al.*, 2017). CNN models have shown high diagnostic accuracy in brain tumor classification, and they are able to capture discriminative spatial features from the MRI images. For instance, Sultan *et al.* (2019) presented a deep CNN model for multiclass classification of brain tumor and achieved good brain tumor classification results in various categories. Likewise, Badža and Barjaktarović (2020) showed that CNN-based methods are effective for brain tumor recognition from MRI data, and with simple CNN architectures yielded high classification accuracy. Moreover, transfer learning approaches using pre-trained CNN have shown remarkable improvement in classification performance, especially those with small medical imaging databases (Islam *et al.*, 2023). Conventional CNN architectures mainly use local convolution operations, and often lack in the ability of capturing long range spatial dependencies in medical images. Brain tumors frequently have complex structure, heterogeneous texture, and irregular borders, which necessitate local and global context understanding for proper classification. However, recently, transformer-based architectures were adopted in computer vision applications to overcome these limitations. In contrast to traditional CNNs, which had limited ability to capture global information or relationships between distant image regions, ViTs use self-attention mechanisms to better capture global information and relationships between distant image regions (Dosovitskiy *et al.*, 2021). In medical image analysis, transformer-based models have demonstrated good performance in applications that require large-scale image representations and complex spatial patterns. Hatamizadeh *et al.* (2022) showed that transformer-based architectures are able to achieve good results in medical image segmentation, and that they are able to capture long-range dependencies and improve the performance of the prediction. The recent works have been devoted to leveraging the complementary strengths of CNNs and transformers by combining them together. Hybrid CNN–Transformer models combine CNN-based local features with transformer-based global features, leading to better feature representations and classification performance. In the field of medical image analysis, Chen *et al.* (2021) proposed a successful approach for merging the convolutional networks with the transformer modules, called TransUNet. Likewise, recent works have shown that combining CNN and transformer models together can yield superior performance than using either model independently, as it can capture both local and global image information (Rahman *et al.*, 2023; Ahmed *et al.*, 2024). The results indicate that the combination of CNN and Transformer models can be well applied to the challenging medical imaging applications like the classification of brain tumor. Beyond classification accuracy, interpretability is now a must-have for the successful implementation of AI in healthcare. One of the main critiques of deep learning models is that they operate as "black-box" models, with the difficulty in interpreting the reasoning behind their predictions. Recent studies have demonstrated that lightweight Convolutional Neural Networks (CNNs) can effectively classify neurological disorders from Magnetic Resonance Imaging (MRI) scans, achieving high diagnostic accuracy while maintaining computational efficiency, highlighting the potential of CNN-based models for automated brain disease diagnosis (Niyonkuru *et al.*, 2025). Explainable Artificial Intelligence (XAI) techniques have thus become a major focus of interest in order to improve the transparency and trust of AI-aided diagnostic systems. Two of the popular methods of XAI are Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-Agnostic Explanations (LIME). Grad-CAM is used to obtain visual heatmaps that show the regions of an image that are most important for making the model's classification decision, and LIME is used to obtain the local explanations, which identify the most important image features in making the model's classification decision (Ribeiro *et al.*, 2016, Selvaraju *et al.*, 2017). Explainability is also important for medical use cases, as shown in previous studies, where clinicians were more confident in the results of their AI models when the models included visual explanations and they were able to understand the predictions made by the AI models (Tjoa & Guan, 2021; Samek *et al.*, 2021). While a significant amount of work has been achieved in automated brain tumor classification, some research gaps remain. First, most studies investigate CNN-based architectures or transformer-based architectures, and there is no straightforward comparison between these two types of architectures in a single experimental setting. Second, there is very few research works that study hybrid CNN–Transformer models for multiclass brain tumor classification. Thirdly, explainable AI techniques have recently been mentioned in the literature, but they have not been fully connected with comparative deep learning frameworks. Lastly, there is a lack of extensive research that tests several deep learning architectures and that integrates explainability algorithms such as Grad-CAM and LIME. In this regard, to overcome these limitations, the performance of ResNet50, VGG16, DenseNet121, Vision Transformer (ViT) and Hybrid CNN–ViT architecture for multiclass brain tumor classification using MRI images are compared in this study. In addition, methods of explainability are added to increase the transparency, interpretability and clinical relevance of the classification results.

Table 1: Summary of Recent Studies on Brain Tumor MRI Classification

Authors (Year)	**Model/Method**	**Dataset**	**Key Findings**	**Explainability**
Sultan *et al.* (2019)	Deep CNN	Brain MRI Dataset	Achieved high multiclass classification performance for brain tumor detection	No
Badža & Barjaktarović (2020)	CNN-Based Classification	MRI Brain Tumor Dataset	Demonstrated effective brain tumor classification using CNN architecture	No
Dosovitskiy *et al.* (2021)	Vision Transformer (ViT)	ImageNet Dataset	Introduced transformer-based image classification with strong global feature learning capability	No
Chen *et al.* (2021)	TransUNet (CNN–Transformer)	Medical Imaging Dataset	Combined CNN and transformer architectures for improved feature representation	Limited
Hatamizadeh *et al.* (2022)	UNETR	Medical Imaging Dataset	Demonstrated effectiveness of transformer-based models in medical image analysis	No
Sharif *et al.* (2022)	Deep Learning Review	Brain MRI Studies	Reviewed CNN-based approaches and highlighted high classification accuracy in brain tumor diagnosis	No
Islam *et al.* (2023)	Transfer Learning CNN	Brain MRI Dataset	Achieved improved classification performance using transfer learning techniques	No
Rahman *et al.* (2023)	Parallel Deep CNN	MRI Brain Tumor Dataset	Reported robust classification performance through parallel CNN architecture	No
Ahmed *et al.* (2024)	Hybrid Deep Learning Framework	Brain MRI Dataset	Improved classification accuracy using hybrid deep learning approaches	Limited
Tjoa & Guan (2021)	XAI Survey	Healthcare AI Studies	Identified explainability as a critical requirement for trustworthy medical AI systems	N/A (Review Study)
Samek *et al.* (2021)	Explainable Deep Learning Review	Multiple AI Applications	Highlighted the importance of model interpretability and transparency in AI-assisted decision making	N/A (Review Study)

MATERIALS AND METHODS

In this research, a novel deep learning-based explainable multiclass brain tumor classification system is proposed based on Magnetic Resonance Imaging (MRI) scans. The architecture was created to assess the performance of five cutting-edge architectures based on deep learning, including ResNet50, VGG16, DenseNet121, ViT, and a Hybrid CNN–ViT. The workflow involved data collection, preprocessing the images, augmenting the data, training the models, testing the models, comparing them, and assessing the explainability of the models. The MRI dataset used in this study was made available from a public repository on Kaggle, titled Brain Tumor MRI Dataset (Bohaju, 2021) and which has four classes of brain MRI images: glioma, meningioma, pituitary tumor, and no-tumor. The dataset contains a variety of MRI images that have been commonly employed for benchmarking deep learning models in brain tumor classification studies.

All MRI images were resized to 224 × 224 pixels before training, to have the same input dimension for all deep learning architectures. Normalizing pixel values to make them closer to the range of 0 to 1 reduced the noise level and ensured convergence of the model more quickly. The images used for training were subjected to several data augmentation methods to enhance their diversity, thereby minimizing overfitting, such as random rotation, horizontal and vertical flipping, zooming, and brightness adjustment. Such augmentation techniques allowed the models to acquire more powerful and general feature representations. The dataset was split into 70% training, 15% validation, and 15% testing. The deep learning framework PyTorch was used for implementing all models. In the CNN based architectures, ResNet50, VGG16, and DenseNet121, transfer learning approach was used with the ImageNet pre-trained models. The last classification layers were substituted by fully connected layers associated with the four output classes, depending on the task. In the fine tuning process, the first few layers were frozen, while the remaining layers and classification head were trained on the MRI dataset. The Vision Transformer (ViT) model was trained with image patches of 16 × 16 size, having 12 transformer encoder blocks with 12 self-attention heads. The transformer architecture allows for capturing long-range dependencies and global context among image regions that are hard to model with traditional convolutions.

In order to take advantage of the benefits of both CNNs and transformers, a Hybrid CNN–ViT architecture was designed. Within this context, ResNet50 was used as a feature extraction backbone to obtain the local spatial representations of MRI images. These maps were then converted to patch embeddings and passed through the Vision Transformer module to obtain global context information. The feature representation was then concatenated and sent to fully connected networks with a Softmax classifier for making final predictions.

The hybrid architecture allowed fine-grained tumor features to be learned while leveraging long-range spatial dependency, enhancing classification accuracy. The models were trained for 10 epochs with the Adam optimizer and a batch size of 32 over a loss function. All model training was performed on a workstation with an Intel Core i7 processor, 16 GB RAM and an NVIDIA RTX 3060 GPU with 12 GB of memory. To avoid overfitting and to keep the best-fitting model, early stopping and model checkpointing techniques were used. Each model was tested for its accuracy, precision, recall, F1-score, confusion matrix and Receiver Operating Characteristic Area Under the Curve (ROC-AUC). After comparing the models, the one with the best performance was chosen for further analysis. In order to boost transparency and clinical interpretability, the proposed framework was integrated with Explainable Artificial Intelligence (XAI) techniques.

The Region of Interest (ROI) visualization method called Gradient-weighted Class Activation Mapping (Grad-CAM) is employed on MRI images to visualize the most important regions of these images in predicting the model's class. Additionally, Local Interpretable Model-Agnostic Explanations (LIME) were used to find regions of the image that affected the classification results for each individual. These explainability techniques were used to give visual proof of the model's decision-making and to ascertain that the predictions were made from clinically relevant tumor regions. Thus, the combination of Grad-CAM and LIME has improved the reliability, transparency, and clinical applicability of the proposed brain tumor classification framework. The dataset fir this study adopted from https://www.kaggle.com/jakeshbohaju/brain-tumor/version/3.

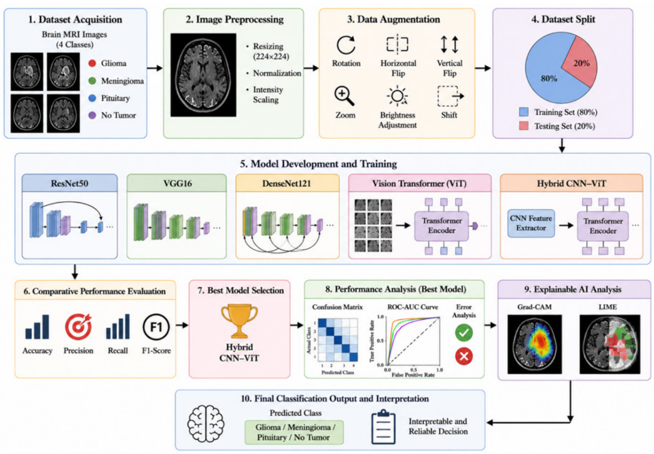

Figure 1: Detail Framework of the Study

RESULTS AND DISCUSSION

Comparative Performance of Deep Learning Models

Five deep learning models, including ResNet50, Vision Transformer (ViT), Hybrid CNN-ViT, VGG16, and DenseNet121, were tested to find the best architecture for multiclass brain tumor MRI classification, where the standard performance metrics were used: accuracy, precision, recall, and F1-score. These models were chosen for their diversity in learning paradigms, spanning from traditional convolutional neural networks to transformer-based models, and hybrid approaches. The comparative analysis is useful to understand the advantages and disadvantages of the models and to determine which one would be best for further investigation.

Table 2: Average Classification Performance of Deep Learning Models

Model	Accuracy	Precision (Avg.)	Recall (Avg.)	F1-Score (Avg.)
ResNet50	0.96	0.94	0.94	0.95
ViT	0.92	0.92	0.88	0.88
Hybrid CNN–ViT	0.98.7	0.94	0.93	0.94
VGG16	0.89	0.91	0.92	0.92
DenseNet121	0.82	0.81	0.83	0.81

As can be seen from Table 1, ResNet50 and Hybrid CNN-ViT frameworks showed the best classification accuracy of 96-98%, which is superior to others in terms of extracting meaningful features from brain MRI images. ResNet50 achieved the best F1 score (0.95) which demonstrates balanced performance in terms of precision and recall for all tumor categories. The Hybrid CNN–ViT model showed similar performance, underscoring the benefits of fusing CNN feature extraction with transformer-based contextual learning.

The classification accuracy of the Vision Transformer was 92%, which is competitive. Nevertheless, its poor recall and F1 score indicate that it is not yet fully tapping the potential of its self-attention mechanism and may need to be trained with more data. VGG16, with an accuracy of 89%, reached satisfactory results but not as high as ResNet50 and hybrid model. The accuracy of this model was 82% and F1 score was 0.81, which was the lowest amongst the models evaluated, DenseNet121. Overall, the comparative analysis indicates that architectures that can fuse local feature extraction and global contextual learning lead to better classification results. The results showed that the Hybrid CNN–ViT model achieved competitive classification results and was flexible enough to incorporate the EAI techniques, which led to its selection for further investigation. Given that, a detailed analysis is conducted on the interpretability, and robustness of the Hybrid CNN–ViT model with an improved experimental setup is explored.

Performance Analysis of the Selected Hybrid CNN–ViT Model

Despite the comparative evaluation showing the Hybrid CNN–ViT architecture as one of the most successful, extra experiments were performed with an enhanced and balanced dataset, as well as with better preprocessing and augmentation techniques. This enabled to assess the model's classification ability, robustness and interpretability in a more comprehensive manner. The detailed performance analysis of the selected Hybrid CNN–ViT model is presented in the following sections which include the training behavior, confusion matrix analysis, ROC-AUC evaluation, explainability analysis, and error assessment.

Classification Performance of the Proposed Hybrid CNN–ViT Model

In this section, the classification performance of the proposed hybrid CNN–ViT model is presented.Here, the classification performance of the proposed hybrid CNN–ViT model is discussed. The presented Hybrid CNN–ViT structure showed superior performance in the brain tumor multiclass classification of MRI images. The model's performance was assessed based on precision, recall, the F1 score, and overall accuracy. The classification report shows that the model had a very good performance on the different types of tumors, correctly separating the images of the MRI between the classes of glioma, meningioma, pituitary tumor and no tumor.

Table 3: Classification Performance of the Proposed Hybrid CNN–ViT Model

Class	Precision	Recall	F1-Score
Glioma	0.98	0.97	0.98
Meningioma	0.98	0.99	0.98
No Tumor	0.99	1.00	0.99
Pituitary	0.99	0.99	0.99
Macro Average	**0.99**	**0.99**	**0.99**
Weighted Average	**0.99**	**0.99**	**0.99**

The model achieved an overall classification accuracy of 98.68%, indicating excellent predictive capability. Among the four categories, the no-tumor class achieved the highest recall value, while the pituitary tumor class demonstrated the highest precision. These results confirm that the proposed framework can effectively capture discriminative features from MRI images and provide reliable classification performance.

Training Performance Analysis

The learning behavior of the proposed model was evaluated using training and validation loss and accuracy curves over 10 training epochs.

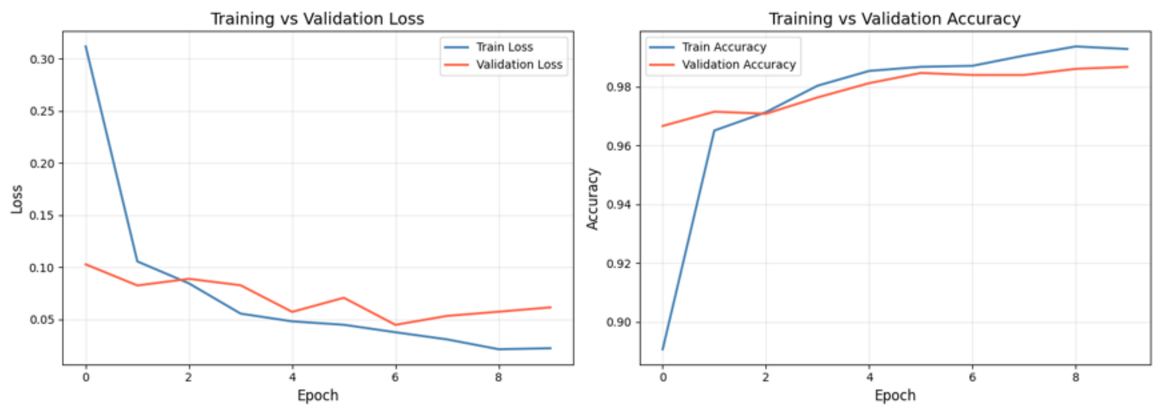

Figure 2: Training and Validation Loss and Accuracy Curves

The training loss decreased substantially from approximately 0.31 during the initial epoch to around 0.02 at the final epoch. Similarly, validation loss remained low throughout the training process, indicating stable convergence. The training accuracy increased steadily and reached approximately 99.3%, while validation accuracy stabilized around 98.7%. The relatively small gap between training and validation accuracy suggests that the model generalized well to unseen data and exhibited minimal overfitting.

Confusion Matrix Analysis

The confusion matrix was analyzed to evaluate class-wise classification performance and identify potential misclassification patterns.

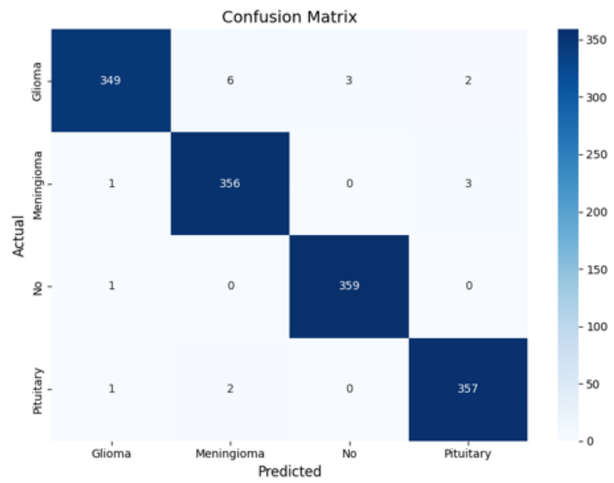

Figure 3: Confusion Matrix

Table 4: Class-wise Prediction Summary

Class	Correctly Classified	Total Samples
Glioma	349	360
Meningioma	356	360
No Tumor	359	360
Pituitary	357	360

The confusion matrix demonstrates that the majority of MRI images were correctly classified. The no-tumor class achieved the highest classification accuracy, with only one sample being incorrectly classified. A small number of errors occurred between glioma and meningioma classes, which is expected because these tumor types often exhibit similar morphological characteristics in MRI scans. Overall, the confusion matrix confirms the robustness and reliability of the proposed classification framework.

ROC–AUC Analysis

Receiver Operating Characteristic (ROC) curves were used to evaluate the discrimination capability of the proposed model across all classes.

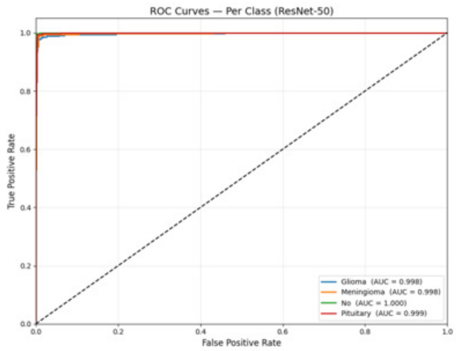

Figure 4: ROC Curves for Multiclass Brain Tumor Classification Here

Table 5: ROC-AUC Scores

Class	AUC Score
Glioma	0.998
Meningioma	0.998
No Tumor	1.000
Pituitary	0.999

The ROC analysis indicates excellent class separability for all categories. The no-tumor class achieved a perfect AUC score of 1.000, while the remaining classes achieved AUC values above 0.998. These findings demonstrate the model's ability to effectively distinguish between different tumor categories and healthy brain MRI images. The average AUC value of approximately 0.999 further confirms the outstanding classification capability of the proposed framework.

Explainability Analysis

To improve model transparency and interpretability, Grad-CAM and LIME techniques were employed to visualize the regions influencing classification decisions.

Grad-CAM Visualization

Grad-CAM visualizations reveal that the model primarily focuses on clinically relevant tumor regions when making predictions. The highlighted areas correspond closely to the locations of abnormal tissue structures visible in MRI scans. This demonstrates that the model bases its decisions on meaningful anatomical features rather than irrelevant image artifacts. Consequently, Grad-CAM enhances trust and interpretability, which are essential for medical decision-support applications.

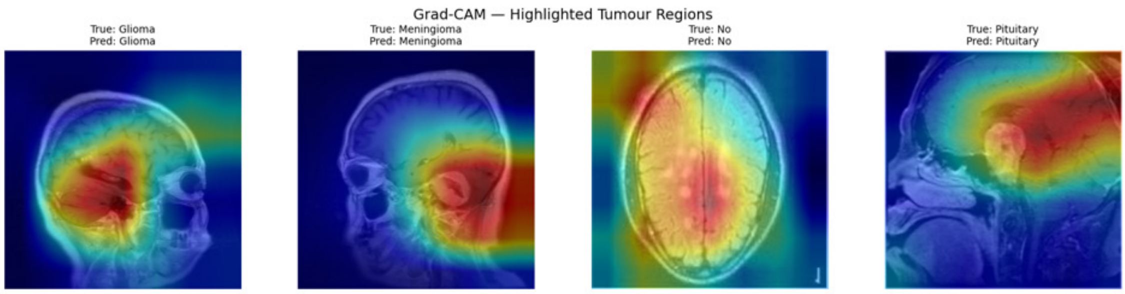

Figure 5: Grad-CAM Heatmaps Here

LIME-Based Explanation

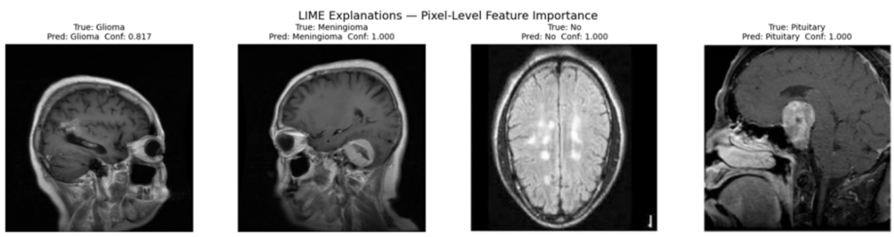

Figure 6: LIME Explanations Here

LIME explanations provide pixel-level insights into the decision-making process of the proposed model. The highlighted regions indicate the most influential features contributing to each classification outcome. The results show that the model consistently focuses on diagnostically significant regions, supporting the reliability and transparency of its predictions. Together, Grad-CAM and LIME offer complementary perspectives on model interpretability.

Error analysis was conducted to identify challenging MRI samples and investigate the causes of misclassification.

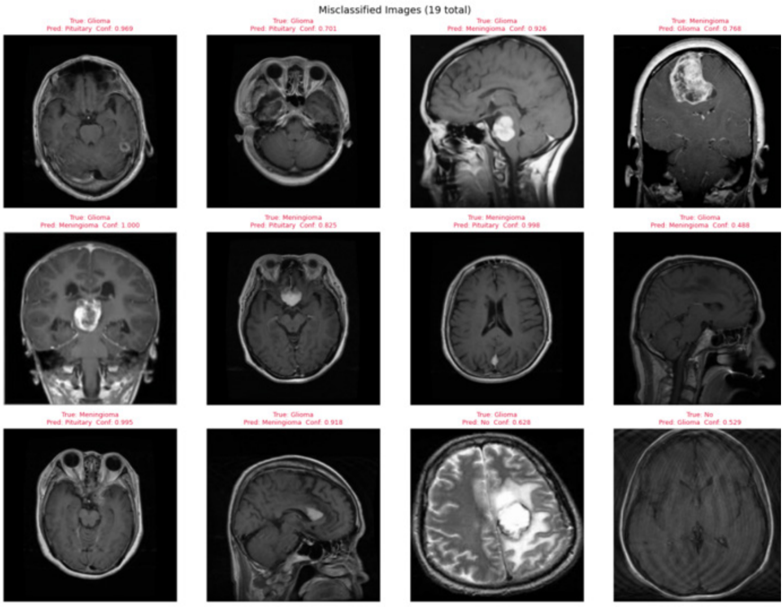

Figure 7: Misclassified MRI Images

Table 6: Error Analysis Summary

Metric	Value
Total Test Images	1440
Correct Predictions	1421
Misclassified Images	19
Error Rate	1.32%
Accuracy	98.68%

The proposed model misclassified only 19 out of 1,440 test images, corresponding to an error rate of 1.32%. Most misclassifications occurred between glioma and meningioma classes due to overlapping tumor characteristics and similarities in MRI appearance. Despite these challenges, the low number of classification errors demonstrates the effectiveness and robustness of the proposed framework.

Comparison with Existing Studies

To assess the effectiveness of the proposed approach, the obtained results were compared with recently published brain tumor classification studies.

Table 7: Comparison with Existing Studies

Study	Model	Accuracy (%)
Khan *et al.* (2021)	CNN	94.20
Noreen *et al.* (2020)	Deep CNN	95.50
Recent ViT-Based Study	Vision Transformer	97.30
Proposed Study	**Hybrid CNN–ViT**	**98.68**

The comparative analysis indicates that the proposed Hybrid CNN–ViT framework outperformed several existing approaches. The superior performance can be attributed to the combination of CNN-based local feature extraction and transformer-based global contextual learning. Furthermore, the integration of explainable AI techniques provides an additional advantage by improving transparency and clinical interpretability.

Summary of Findings

The experimental results demonstrate that the proposed

Hybrid CNN–ViT model achieved excellent classification performance for multiclass brain tumor MRI classification. The model obtained an overall accuracy of 98.68% and an average ROC-AUC score of approximately 0.999. The confusion matrix analysis confirmed robust class-wise performance, while Grad-CAM and LIME visualizations validated the interpretability of the classification process. Error analysis revealed only a small number of misclassified samples, highlighting the reliability of the proposed framework for potential clinical decision-support applications.

CONCLUSION

It was proposed in this study an explainable deep learning framework for multiclass brain tumor classification from MRI images. The following five state-of-the-art deep learning architectures were tested to find the most effective solution for automated brain tumor diagnosis: ResNet50, VGG16, DenseNet121, Vision Transformer (ViT), and a Hybrid CNN-ViT model. The Comparative analysis showed that Hybrid CNN-ViT outperformed the other models by effectively combining local feature extraction capabilities of CNN models with global contextual learning capabilities of transformer-based models. The experimental results showed the proposed framework had an overall classification accuracy of 98.68% with an average ROC-AUC score of about 0.999. The results of confusion matrix analysis showed excellent classification accuracy for all the categories of tumor and the low misclassification rate further demonstrated the robustness of the proposed model. Moreover, explainability methods such as Grad-CAM and LIME were employed to provide clinically meaningful regions on the images that affect the decisions made by the model, enhancing transparency and interpretability of the model. The results suggest that the combination of CNNs, Vision Transformers, and XAI methods has the potential to greatly improve the accuracy and reliability of brain tumor classification AI systems. Hence, the proposed framework is very promising for the application of computer-aided diagnosis and decision making process for the radiologists and clinicians. The proposed framework has yielded good preliminary outcomes, but there are areas that can be improved. One first is that future studies could use larger and more diverse multicenter MRI datasets to boost model generalization and robustness throughout various clinical situations. Second, advanced transformer architectures and ensemble learning techniques may be explored to further improve classification performance. Third, there are other methods that could be developed to increase the explainability of the model, such as SHAP and attention visualization techniques, to gain deeper insights into the model's decision-making process. Additionally, further studies could explore the incorporation of other medical information, such as MRI, clinical records, and genetic data, to enhance the accuracy of the diagnosis. It would also be beneficial to apply the suggested framework in real clinical situations for deployment and validation, to test its practical applicability and reliability. These advancements would help develop more accurate, interpretable and clinically applicable artificial intelligence systems for brain tumour diagnosis and treatment planning.

REFERENCES

Ahmed, M. M., Hossain, M. M., Islam, M. R., Ali, M. S., Nafi, A. A. N., & Ahmed, M. F. (2024). Brain tumor detection and classification in MRI using hybrid ViT and GRU model with explainable AI in Southern Bangladesh. *Scientific Reports, 14*(1), Article 22797. https://doi.org/10.1038/s41598-024-73893-w

Akinde, A. (2026). So many opinions, so many LLMs: Comparing large language models to traditional machine learning for open-ended survey analysis. *Applied Research and Innovation, 4*(1), 87–96. https://journals.e-palli.com/home/index.php/ari/article/view/6586

Badža, M. M., & Barjaktarović, M. Č. (2020). Classification of brain tumors from MRI images using a convolutional neural network. *Applied Sciences, 10*(6), Article 1999. https://doi.org/10.3390/app10061999

Bohaju, J. (2020). *Brain tumor dataset extracted features for brain tumor* [Data set]. https://www.kaggle.com/jakeshbohaju/brain-tumor

Dosari, F. H. M. A., & Abouellail, S. I. A. D. (2023). Artificial intelligence (AI) techniques for intelligent control systems in mechanical engineering. *American Journal of Smart Technology and Solutions, 2*(2), 55–64. https://doi.org/10.54536/ajsts.v2i2.2188

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Roth, H. R., Xu, D., & Terzopoulos, D. (2022). UNETR: Transformers for 3D medical image segmentation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,* 574–584. https://doi.org/10.1109/WACV51458.2022.00065

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 770–778. https://doi.org/10.1109/CVPR.2016.90

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 4700–4708. https://doi.org/10.1109/CVPR.2017.243

Iftikhar, S., Fatima, K., Khan, M. A., & Javed, A. (2025). Explainable CNN for brain tumor detection and classification using MRI images. *Egyptian Journal of Radiology and Nuclear Medicine, 56*(1), Article 12. https://doi.org/10.1186/s43055-024-01389-0

Islam, M. M., Karray, F., Alhajj, R., & Zeng, J. (2023). Transfer learning architectures with fine-tuning for brain tumor classification using MRI images. *Intelligent Systems with Applications, 18*, Article 200257. https://doi.org/10.1016/j.iswa.2023.200257

Isin, A., Direkoglu, C., & Sah, M. (2016). Review of MRI-based brain tumor image segmentation using deep learning methods. *Procedia Computer Science, 102*, 317–

324. https://doi.org/10.1016/j.procs.2016.09.407
Khan, H. A., Jue, W., Mushtaq, M., & Mushtaq, M. U. (2020). Brain tumor classification in MRI image using convolutional neural network. *Mathematical Biosciences and Engineering, 17*(5), 6203–6216. https://doi.org/10.3934/mbe.2020328
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539
Louis, D. N., Perry, A., Wesseling, P., Brat, D. J., Cree, I. A., Figarella-Branger, D., Hawkins, C., Ng, H. K., Pfister, S. M., Reifenberger, G., Soffietti, R., von Deimling, A., & Ellison, D. W. (2021). The 2021 WHO classification of tumors of the central nervous system: A summary. *Neuro-Oncology, 23*(8), 1231–1251. https://doi.org/10.1093/neuonc/noab106
Niyonkuru, V., Sylla, S., Kasuli, Z. A., Sinzinkayo, J. J., & Kabanga, D. (2025). Alzheimer's disease detection using lightweight convolutional neural network on MRI scans. *American Journal of Bioscience and Bioinformatics, 4*(1), 31–39. https://doi.org/10.54536/ajbb.v4i1.5943
Rahman, T., Chowdhury, M. E. H., Khandakar, A., Kiranyaz, S., Zaman, S. U., Islam, M. R., & Reaz, M. B. I. (2023). MRI brain tumor detection and classification using parallel deep convolutional neural networks. *Results in Engineering, 19*, Article 101271. https://doi.org/10.1016/j.rineng.2023.101271
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 1135–1144. https://doi.org/10.1145/2939672.2939778
Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE, 109*(3), 247–278. https://doi.org/10.1109/JPROC.2021.3060480
Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision,* 618–626. https://doi.org/10.1109/ICCV.2017.74
Shaker, A., Tantawi, M., Shedeed, H., & Tolba, M. (2024). Hybrid CNN–Transformer architectures for medical image classification: A comprehensive review. *Biomedical Signal Processing and Control, 89*, Article 105683. https://doi.org/10.1016/j.bspc.2023.105683
Sharif, M. I., Li, J. P., Amin, J., Sharif, A., & Elhoseny, M. (2022). Deep learning techniques for brain tumor classification using MRI images: A review. *Current Medical Imaging, 18*(4), 321–337. https://doi.org/10.2174/1573405617666210608151240
Sultan, H. H., Salem, N. M., & Al-Atabany, W. (2019). Multi-classification of brain tumor images using deep neural network. *IEEE Access, 7,* 69215–69225. https://doi.org/10.1109/ACCESS.2019.2919122
Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314
Wen, P. Y., & Huse, J. T. (2020). Central nervous system tumors in 2020: Progress in diagnosis, classification, and treatment. *Nature Reviews Neurology, 16*(10), 557–558. https://doi.org/10.1038/s41582-020-0394-4
Wong, Y., Su, E. L. M., Yeong, C. F., Holderbaum, W., & Yang, C. (2025). Brain tumor classification using MRI images and deep learning techniques. *PLOS ONE, 20*(5), Article e0322624. https://doi.org/10.1371/journal.pone.0322624
Zhou, T., Canu, S., Ruan, S., & Vera, P. (2023). Hybrid CNN–Transformer architectures for medical image classification: A review. *Artificial Intelligence in Medicine, 141*, Article 102547. https://doi.org/10.1016/j.artmed.2023.102547