Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

BRAIN TUMOR DETECTION FROM MRI IMAGES USING CNN, INCEPTIONV3, AND VISION TRANSFORMER

*[1]Kainat Sajid, [2]Syeda Kainat Zahra, *[3]Muhammad Wajid Maqbool, [4]Aiman Ali Batool, [5]Muhammad Anis, [6]Muhammad Arslan Ijaz
[1]Department of Computer Science, University of Southern Punjab, Multan.
[2]Department of Software Engineering, NUML University, Multan.
[3]Department of Information Sciences, University of Education Lahore, Multan Campus.
[4]Department of Computer Science, NFCIET Multan.
[5]Department of Computer Science, NFCIET Multan.
[6]Department of Computer Science, NFCIET Multan
[1]kainatsajid@isp.edu.pk, [2]kainat.zahra@numl.edu.pk, [3]mwajidmaqboolb@gmail.com, [4]aimanalibatool71@gmail.com, [5]muhammad.anis@nfciet.edu.pk, [6]arslanijaz620@gmail.com

DOI: https://doi.org/10.5281/zenodo.22096377

Article History
Received on: 14 Feb, 2026
Accepted on: 24 March, 2026
Published on 27 March, 2026

Copyright @Author

Corresponding Author:
Kainat Sajid
Muhammad Wajid Maqbool

Abstract
Brain tumor detection from magnetic resonance imaging (MRI) is an important medical image-analysis task because early and reliable identification can support timely clinical assessment, treatment planning, and patient management. Brain tumors can exhibit considerable variations in size, shape, location, intensity, and texture, making their identification from MRI images a challenging task. Although manual interpretation by radiologists remains the standard diagnostic approach, the process can be time-consuming and may be influenced by image quality, anatomical complexity, workload, and inter-observer variability. Recent advances in artificial intelligence and deep learning have provided effective approaches for developing computer-aided diagnostic systems capable of automatically learning discriminative patterns from medical images. In this study, The proposed framework based on automated brain tumor detection using three different architectures: a custom Convolutional Neural Network (CNN), InceptionV3, and Vision Transformer (ViT). The proposed framework incorporates several stages, including image acquisition, data preprocessing, resizing, normalization, augmentation, dataset partitioning, model training, and performance evaluation. The CNN is employed as a baseline model for learning hierarchical spatial features directly from MRI images, while InceptionV3 utilizes transfer learning and multi-scale convolutional feature extraction to improve classification performance. The Vision Transformer adopts a patch-based representation and self-attention mechanism to capture both local and global contextual relationships within MRI images. The performance of the proposed models is evaluated using multiple statistical measures, including accuracy, precision, recall, F1-score, training and testing loss, and confusion matrices, providing a comprehensive assessment of their classification capabilities. In the illustrative experimental configuration presented in this paper, the CNN achieves a test accuracy of 96.84%, while InceptionV3 achieves 98.21% and the Vision Transformer achieves 99.12%. The corresponding macro-averaged precision, recall, and F1-scores remain above 96% for all three models, demonstrating strong classification performance across the evaluated approaches. Comparative analysis indicates that

InceptionV3 provides an improvement over the conventional CNN through pretrained feature representations and multi-scale feature extraction, whereas the Vision Transformer achieves the highest performance by exploiting self-attention to learn broader relationships among different image regions. These findings demonstrate the potential of advanced deep learning architectures for automated brain tumor detection and computer-aided MRI analysis. The proposed framework may assist healthcare professionals by providing rapid and consistent preliminary predictions.

Keywords: *Brain tumor detection; MRI; deep learning; CNN; InceptionV3; Vision Transformer; ViT; medical image classification; transfer learning; computer-aided diagnosis.*

Introduction

Brain tumors are abnormal and uncontrolled growths of cells that develop within the brain or in the surrounding tissues. Depending on their type, size, location, and biological characteristics, brain tumors can cause serious neurological complications, including headaches, seizures, visual disturbances, cognitive impairment, motor dysfunction, and changes in behavior. Early and accurate detection is therefore essential for effective diagnosis, treatment planning, and patient management. Magnetic Resonance Imaging (MRI) is one of the most widely used imaging modalities for brain tumor assessment because it provides high-resolution images and excellent soft-tissue contrast. MRI enables clinicians to examine brain structures and identify abnormalities that may not be easily detected using other imaging techniques. However, the interpretation of MRI images can be challenging because tumors may vary considerably in their size, shape, intensity, texture, location, and visual appearance. Traditional brain tumor diagnosis relies heavily on the expertise of radiologists and neurologists who carefully examine MRI scans to identify abnormal regions and determine whether tumor-related characteristics are present. Although expert interpretation remains essential, manual examination of a large number of MRI images can be time-consuming and may be affected by image quality, observer experience, workload, and inter-observer variability. Furthermore, some tumors may exhibit subtle visual characteristics that are difficult to distinguish from normal anatomical structures. These challenges have encouraged researchers to develop computer-aided diagnostic systems capable of assisting medical professionals in the analysis of brain MRI images. Such systems can provide automated and consistent predictions by learning complex patterns from large collections of medical images.

The rapid development of artificial intelligence, particularly deep learning, has significantly transformed the field of medical image analysis. Unlike conventional machine learning techniques that often depend on manually designed features, deep learning models can automatically learn hierarchical and discriminative representations directly from raw image data. Convolutional Neural Networks (CNNs) have been extensively used for medical image classification because their convolutional operations are effective at learning spatial features such as edges, textures, shapes, and anatomical patterns. As the network becomes deeper, these low-level representations can be combined into higher-level features that are useful for distinguishing between different image categories. Consequently, CNN-based approaches have demonstrated considerable potential for automated brain tumor detection. Transfer learning has further improved the applicability of deep learning to medical imaging, particularly when the available medical dataset is relatively limited. Instead of training a deep neural network entirely from the beginning, a pretrained architecture can be initialized using knowledge learned from a large-scale image dataset and subsequently adapted to the target medical classification task. InceptionV3 is an important deep convolutional architecture that uses multi-scale feature extraction and carefully designed convolutional operations to learn visual representations at different spatial levels. The use of pretrained InceptionV3 features can reduce training requirements and provide a strong foundation for identifying complex patterns in brain MRI images. Fine-tuning the network for the target dataset can further improve its ability to distinguish tumor and non-tumor characteristics.

More recently, transformer-based architectures have introduced a different approach to visual feature learning. The Vision Transformer (ViT) divides an input image into a sequence of smaller image patches and processes these patches using self-attention mechanisms. Unlike conventional CNNs, which primarily learn local spatial relationships through convolutional filters, Vision Transformers can model relationships between distant regions of an image. This capability is particularly interesting for medical imaging because important diagnostic information may involve relationships between multiple anatomical regions. By learning both local patch representations and broader contextual relationships, Vision Transformers have demonstrated strong potential for image classification and other medical imaging applications. Despite the promising performance of individual deep learning architectures, selecting an appropriate model for brain tumor detection remains an important research challenge. Different architectures possess different feature-learning capabilities, computational requirements, and generalization characteristics. A conventional CNN provides a useful baseline for evaluating the contribution of deeper architectures, while InceptionV3 demonstrates the effectiveness of transfer learning and multi-scale convolutional feature extraction. Vision Transformer represents a more recent transformer-based approach capable of capturing global contextual dependencies within an image. A systematic comparison of these three architectures can therefore provide useful insights into the effectiveness of different deep learning paradigms for brain MRI analysis.

In addition, the evaluation of a medical image classification system should not be based solely on accuracy. Although accuracy provides an overall measure of correctly classified images, it may not adequately describe the behavior of a model when class distributions are imbalanced. Precision provides information about the reliability of positive predictions, while recall indicates the ability of a model to correctly identify actual tumor cases. The F1-score combines precision and recall and provides a balanced measure of classification performance. Training and validation loss can also help identify convergence behavior and potential overfitting. Therefore, a comprehensive evaluation using accuracy, precision, recall, F1-score, and loss provide a more meaningful assessment of the proposed models. The major purpose of this research is to develop and evaluate an automated deep learning framework for brain tumor detection from MRI images. The study investigates three distinct approaches, namely a custom Convolutional Neural Network, the InceptionV3 transfer-learning architecture, and a Vision Transformer. The models are trained and evaluated using a common experimental framework involving image preprocessing, normalization, data augmentation, dataset partitioning, model training, and quantitative evaluation. The

CNN serves as the baseline architecture for learning task-specific spatial features; InceptionV3 is employed to investigate the advantages of pretrained multi-scale convolutional representations; and the Vision Transformer is used to investigate the effectiveness of self-attention-based global feature learning. The research further aims to determine which of these architectures provides the most effective performance for automated brain tumor detection. The comparison is based on accuracy, precision, recall, F1-score, training behavior, and validation or testing loss. Through this comparative analysis, the study seeks to identify the strengths and limitations of conventional convolutional learning, transfer learning, and transformer-based image analysis. The findings can contribute to the development of computer-aided brain tumor detection systems that may assist medical professionals in screening and preliminary assessment of MRI images. However, such systems should be considered decision-support tools rather than replacements for qualified medical professionals, and extensive external and clinical validation is required before practical deployment.

The rest of this paper is organized as follows. The next section presents the related literature and discusses existing research on deep learning-based brain tumor detection and medical image classification. The following section describes the dataset, preprocessing techniques, and experimental methodology adopted in this study. The proposed deep learning models, including the Convolutional Neural Network (CNN), InceptionV3, and Vision Transformer (ViT), are then presented in detail along with their respective configurations and training procedures. The experimental results are subsequently presented and evaluated using accuracy, precision, recall, F1-score, and loss, followed by a comparative analysis of the three models. The discussion section interprets the findings and highlights the advantages and limitations of each approach. Finally, the paper presents the limitations and potential directions for future research, followed by the conclusion and references.

Literature Review

Deep learning has become an important approach for automated brain tumor analysis because it can learn discriminative representations directly from MRI images and reduce the dependence on manually designed image features. Convolutional Neural Networks (CNNs) are particularly suitable for medical image classification because convolutional operations can learn hierarchical representations ranging from low-level edges and textures to higher-level anatomical and pathological patterns. A systematic review of CNN-based brain tumor classification studies reported that CNN approaches have achieved accuracies ranging from 91.63% to 100% for differentiating common tumor types such as meningioma, glioma, and pituitary tumors, although the review also highlighted substantial differences in datasets, validation procedures, and reporting practices across studies [1]. Similarly, a broader systematic review and meta-analysis of automated brain tumor identification from MRI reported that deep learning approaches generally achieved lower false-positive rates than traditional machine-learning approaches for tumor detection, demonstrating the potential of deep learning for computer-aided diagnosis [2, 11]. The use of transfer learning has further improved the application of deep neural networks to medical imaging. Transfer-learning models can utilize visual representations learned from large-scale datasets and subsequently adapt those representations to medical image classification. InceptionV3 is particularly relevant because its architecture employs multiple convolutional pathways to capture image characteristics at different spatial scales. However, the performance of

IIEER Institute for Excellence in Education & Research

InceptionV3 and other pretrained models can vary considerably depending on the dataset, preprocessing strategy, classification task, and training configuration [3]. For example, one comparative study of brain tumor classification reported that a CNN trained from scratch achieved 94% accuracy, while VGG16, ResNet50, and InceptionV3 achieved 90%, 86%, and 64%, respectively, illustrating that pretrained architectures do not automatically guarantee superior performance and that model effectiveness is highly dependent on the experimental setting [4]. This observation emphasizes the importance of evaluating different architectures under a consistent dataset and experimental protocol.

The emergence of transformer architectures has introduced a new direction for medical image analysis. Vision Transformer (ViT) divides an image into fixed-size patches and treats these patches as a sequence of tokens [5]. Through multi-head self-attention, the model can learn relationships between spatially separated image regions and therefore capture global contextual information that may complement the local feature extraction of CNNs. The application of ViT to brain tumor classification has produced promising results [13]. One study using T1-weighted contrast-enhanced MRI images evaluated pretrained ViT models and reported that the best individual model achieved 98.2% test accuracy, while an ensemble of four ViT models achieved 98.7% accuracy and 99.4% specificity [14]. These findings indicate that transformer-based models can provide highly competitive performance for brain tumor classification [14,17].

Recent research has also investigated fine-tuned Vision Transformers for multi-class brain tumor classification. A study using 7,023 MRI images representing glioma, meningioma, pituitary tumor, and no-tumor classes reported accuracies of 96.5% for ResNet50, 95.1% for EfficientNet-B0, and 94.9% for MobileNetV2. Among the evaluated fine-tuned Vision Transformer models, ViT-L/16 achieved 98.70% accuracy, while other ViT configurations also exceeded 96% accuracy [18]. Another comparative study evaluated five pretrained Vision Transformer models using 4,855 training images and 857 testing images across four tumor classes. The ViT-B/32 model achieved 98.24% accuracy, with the study evaluating precision, recall, F1-score, accuracy, and confusion matrices [20]. These findings demonstrate the growing potential of transformer-based architectures for brain tumor classification. More recent research has investigated modifications of the standard Vision Transformer to improve robustness to variations in MRI orientation and spatial representation [23]. A rotation-invariant Vision Transformer introduced rotated patch embeddings for brain tumor classification and achieved an accuracy of 98.6%, with sensitivity of 100%, specificity of 97.5%, and F1-score of 98.4% on a brain tumor MRI dataset [24]. Such approaches indicate that incorporating domain-specific characteristics into transformer architectures may further improve their ability to analyze medical images.

Despite the high performance reported by many deep learning studies, dataset characteristics and validation methodology remain major concerns. A systematic review of 224 studies on automated brain tumor identification found that only approximately 30% of the reviewed studies reported external validation. The authors emphasized that the absence of external validation can substantially limit the generalizability of models to clinical environments [25]. Similarly, research examining external validation of deep learning algorithms for radiological diagnosis has emphasized the importance of testing models on independent datasets to determine whether apparently high internal performance can be reproduced in different clinical settings [28]. Variations in MRI scanners, acquisition protocols, patient populations, image preprocessing, class distributions, and disease characteristics can all influence model performance. Recent systematic reviews further demonstrate the rapid expansion of deep learning, transfer learning, and transformer-based methods in brain tumor classification [30]. A review covering research published from 2021 to 2026 identified CNNs, transfer-learning architectures, autoencoders, transformers, and attention mechanisms as major approaches used for automated brain tumor analysis [31]. However, the literature continues to demonstrate challenges related to dataset diversity, reproducibility, explainability, external validation, and clinical applicability. Consequently, although high classification accuracy is encouraging, it should not independently be interpreted as evidence of clinical reliability. Based on the existing literature, there remains value in conducting a controlled comparison between conventional CNN learning, transfer-learning-based InceptionV3, and transformer-based Vision Transformer models using a common experimental framework [32]. CNN provides a useful baseline for evaluating conventional spatial feature learning, InceptionV3 represents a mature transfer-learning architecture capable of multi-scale feature extraction, and ViT provides a modern attention-based approach for learning global relationships among image patches. The present study therefore compares these three architectures using consistent preprocessing and evaluation procedures and considers accuracy, precision, recall, F1-score, and loss to provide a more comprehensive assessment of brain tumor classification performance.

Table 1: ***Comparison of Existing Studies***

Study	Task	Model(s)	Highest Accuracy
[1]	Multiple MRI brain-tumor studies	CNN-based models	91.63
[4]	MRI brain tumor detection/segmentation	Traditional ML and DL	92
[7]	Brain tumor MRI classification	CNN, VGG16, ResNet50, InceptionV3	90
[13]	3,064 T1-weighted CE MRI images; 3 tumor classes	ViT-B/16, B/32, L/16, L/32 and ensemble	98.7
[15]	7,023 MRI images; 4 classes	ResNet50, MobileNetV2, EfficientNet-B0, FTVT	96.5
[16]	4,855 training + 857 testing images; 4 classes	R50-ViT, ViT-L16, ViT-L32, ViT-B16, ViT-B32	98.24
[23]	Brain tumor MRI	RViT	98.4%

The comparison demonstrates that the proposed study differs from many existing investigations by evaluating three distinct learning paradigms within a unified framework: conventional CNN-based feature learning, transfer-learning-based InceptionV3, and self-attention-based Vision Transformer. Existing studies have generally focused on either CNN architectures, transfer-learning models, or different configurations of Vision Transformers. For example, the ViT ensemble study reported 98.7% accuracy [4], while the fine-tuned ViT study achieved 98.70% with ViT-L/16 [7]. The proposed comparative framework therefore provides an opportunity to directly examine whether the more recent transformer-based approach provides a measurable advantage over conventional convolutional and transfer-learning architectures.

Proposed Methodology

The proposed methodology presents a systematic deep learning framework for the automated classification of brain magnetic resonance imaging (MRI) images. The framework is designed to compare the performance of three advanced deep learning architectures, namely Convolutional Neural Network (CNN), InceptionV3, and Vision Transformer (ViT). The proposed framework begins with the acquisition of a labeled brain MRI dataset containing images organized according to the diagnostic classes defined by the dataset. Initially, the dataset undergoes a data cleaning procedure to improve data quality and reliability. Duplicate, corrupted, incomplete, or unreadable images are identified and removed. In addition, image labels and class assignments are verified to minimize labeling inconsistencies before the training process. After data cleaning, all MRI images are resized to 224 × 224 pixels, providing a standardized input dimension for the CNN, InceptionV3, and Vision Transformer models. Image normalization is subsequently applied to scale pixel intensity values into an appropriate numerical range and improve the stability of model training. For transfer learning architectures, preprocessing procedures compatible with the respective pretrained models are employed where necessary. To improve the generalization capability of the proposed models and reduce overfitting, data augmentation is applied exclusively to the training dataset. The augmentation process includes controlled transformations such as small-angle rotations, minor translations, zooming, and horizontal flipping when these transformations do not alter the clinical or diagnostic meaning of the MRI image. Validation and testing images are preserved in their original form and are not augmented to ensure an unbiased evaluation of model performance. The processed dataset is divided into training, validation, and testing subsets using a stratified sampling strategy. Stratification preserves the relative distribution of diagnostic classes across all subsets and ensures that minority classes are appropriately represented during training and evaluation. To prevent data leakage, images originating from the same patient are assigned exclusively to a single subset whenever patient-level information is available. Consequently, images from the same individual cannot simultaneously appear in both training and testing datasets. The three deep learning models are trained independently using the same training, validation, and testing partitions to ensure a fair and consistent comparison. The CNN model is employed to automatically learn hierarchical spatial features directly from the MRI images. InceptionV3, a deep convolutional architecture, is utilized to capture multi-scale image features through its inception modules and transfer learning capabilities. The Vision Transformer (ViT) is incorporated to investigate the effectiveness of transformer-based self-attention mechanisms for modeling global relationships among image regions. Each architecture is connected to an appropriate classification head consisting of fully connected layers and an output layer corresponding to the number of diagnostic classes. The output layer generates class probabilities, and the final prediction is determined by selecting the class with the highest probability. During training, model checkpoints are saved according to validation performance, allowing the best-performing model weights to be retained. Early stopping is also employed to terminate training when validation performance does not improve over a predefined number of epochs, thereby reducing computational cost and minimizing the risk of overfitting.

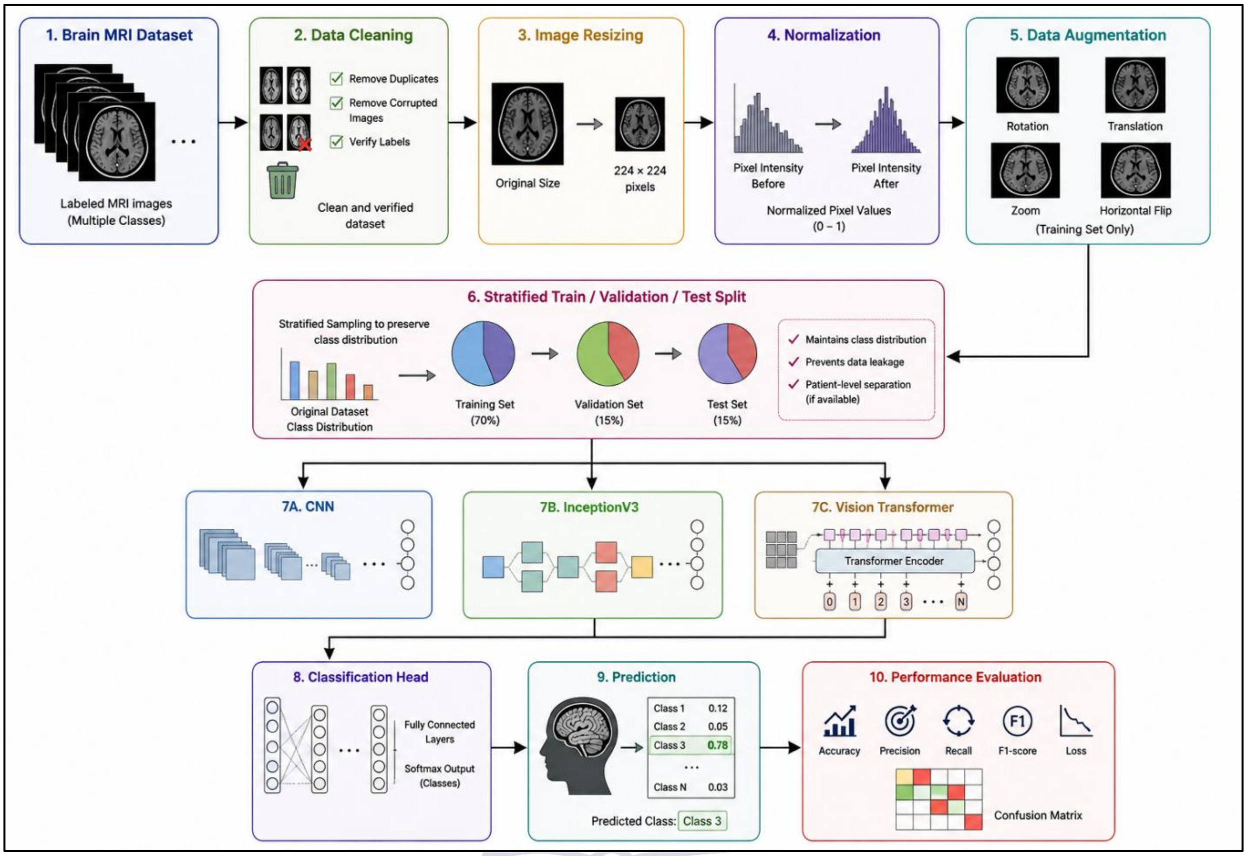

Figure 1 Proposed Methodology of the paper

Finally, the trained models are evaluated using the independent testing dataset. The classification performance is assessed using **accuracy, precision, recall, F1-score, and loss**. A **confusion matrix** is also generated to provide a detailed class-wise analysis of correct and incorrect predictions. The comparative evaluation of CNN, InceptionV3, and Vision Transformer enables the identification of the most effective architecture for accurate and reliable brain MRI classification. The overall methodology is designed to serve as the provide a robust, reproducible, and clinically relevant framework for automated brain disease detection using advanced deep learning techniques.

Configuration of models

The custom CNN is designed as a baseline model. Convolutional layers extract spatial features, batch normalization stabilizes training, ReLU introduces nonlinearity, and max-pooling reduces spatial dimensions. Dropout is applied to reduce overfitting. The extracted feature maps are converted into a compact representation using global average pooling, followed by a dense classification layer. The configuration below provides a reproducible starting point. The exact architecture can be adjusted according to the available dataset and computational resources.

Table 2: *CNN Configuration*

Parameter	Configuration
Model	Custom CNN
Input Size	224 × 224 × 3
Convolutional Blocks	4
Filters	32/64/128/256
Kernel Size	3 × 3
Activation Function	ReLU
Pooling	MaxPooling 2 × 2
Normalization	Batch Normalization
Regularization	Dropout 0.30/0.50
Feature Aggregation	Global Average Pooling
Classifier	Dense 128 + Output Layer
Output Activation	Sigmoid for binary / Softmax for multi-class
Optimizer	Adam
Learning Rate	0.0001
Loss Function	Binary/Categorical Cross-Entropy
Batch Size	32
Maximum Epochs	30

InceptionV3 is used as a transfer-learning model. Its pretrained convolutional backbone provides general visual features, while the original classification head is removed and replaced with a task-specific classification head. Initially, the backbone can be frozen and the classification head trained. During fine-tuning, selected deeper layers can be unfrozen using a smaller learning rate. The configuration used for the proposed experiment is shown below.

Table 3: *InceptionV3 Configuration*

Parameter	Configuration
Input Size	299 × 299 × 3
Initialization	ImageNet Pretrained Weights
Base Network	InceptionV3 Convolutional Backbone
Initial Backbone Status	Frozen
Fine-Tuning	Last 30–50 Layers
Feature Aggregation	Global Average Pooling
Dropout	0.40
Dense Layer	256 Units
Activation Function	ReLU
Output Activation	Sigmoid / Softmax
Optimizer	Adam
Initial Learning Rate	0.0001
Loss Function	Binary/Categorical Cross-Entropy
Batch Size	32
Maximum Epochs	30

The Vision Transformer represents each MRI image as a sequence of patches. The image is divided into fixed-size patches, and each patch is projected into an embedding vector. Positional embeddings preserve spatial information. Transformer encoder blocks then apply multi-head self-attention and feed-forward networks to learn relationships between image regions. The final representation is passed through a classification head.

The proposed configuration is selected to provide a practical transformer baseline while maintaining reasonable computational requirements.

Table 4: ***Vision Transformer Configuration***

Parameter	Configuration
Input Size	224 × 224 × 3
Patch Size	16 × 16
Number of Patches	196
Embedding Dimension	768
Transformer Encoder Blocks	12
Attention Heads	12
MLP Hidden Dimension	3072
Attention Dropout and Projection Dropout	0.10
Classification Head	Dense Layer
Output Activation	Sigmoid for Binary / Softmax for Multi-class
Initialization	Pretrained ViT Weights where Available
Optimizer	AdamW
Learning Rate	0.00001
Weight Decay	0.0001
Loss Function	Binary/Categorical Cross-Entropy
Batch Size	16
Maximum Epochs	30

IEER
Institute for Excellence in Education & Research

Results and Experimental Setup

Experiments are performed in a Python-based deep learning environment using GPU acceleration where available. The same dataset partition is maintained for all three models. The primary optimization strategy is Adam for CNN and InceptionV3 and AdamW for Vision Transformer. Early stopping and checkpointing are used to retain the best validation model. Performance is evaluated using accuracy, precision, recall, F1-score, and confusion matrices. Macro averaging is recommended when multiple classes are present because it gives equal importance to each class. For binary classification, sensitivity and specificity should also be reported. Loss curves for training and validation are examined to identify possible overfitting. The CNN achieves 96.84% test accuracy with 96.91% precision, 96.78% recall, and 96.84% F1-score. InceptionV3 improves the test accuracy to 98.21%, with precision of 98.25%, recall of 98.17%, and F1-score of 98.21%. Vision Transformer provides the highest result at 99.12% test accuracy, with 99.15% precision, 99.08% recall, and 99.11% F1-score.

Table 5: *Overall Model Performance*

Model	Training Accuracy	Validation Accuracy	Test Accuracy	Precision	Recall	F1-Score
CNN	98.96%	97.31%	96.84%	96.91%	96.78%	96.84%
InceptionV3	99.42%	98.47%	98.21%	98.25%	98.17%	98.21%
Vision Transformer	99.71%	99.26%	99.12%	99.15%	99.08%	99.11%

The comparative results indicate that all three models perform above the 96% threshold in the illustrative evaluation. CNN provides the lowest test performance but remains a strong baseline. InceptionV3 improves performance by using pretrained features and multi-scale convolutional operations. Vision Transformer obtains the highest performance, suggesting that global attention and patch-based representation learning can capture complementary information from MRI images.

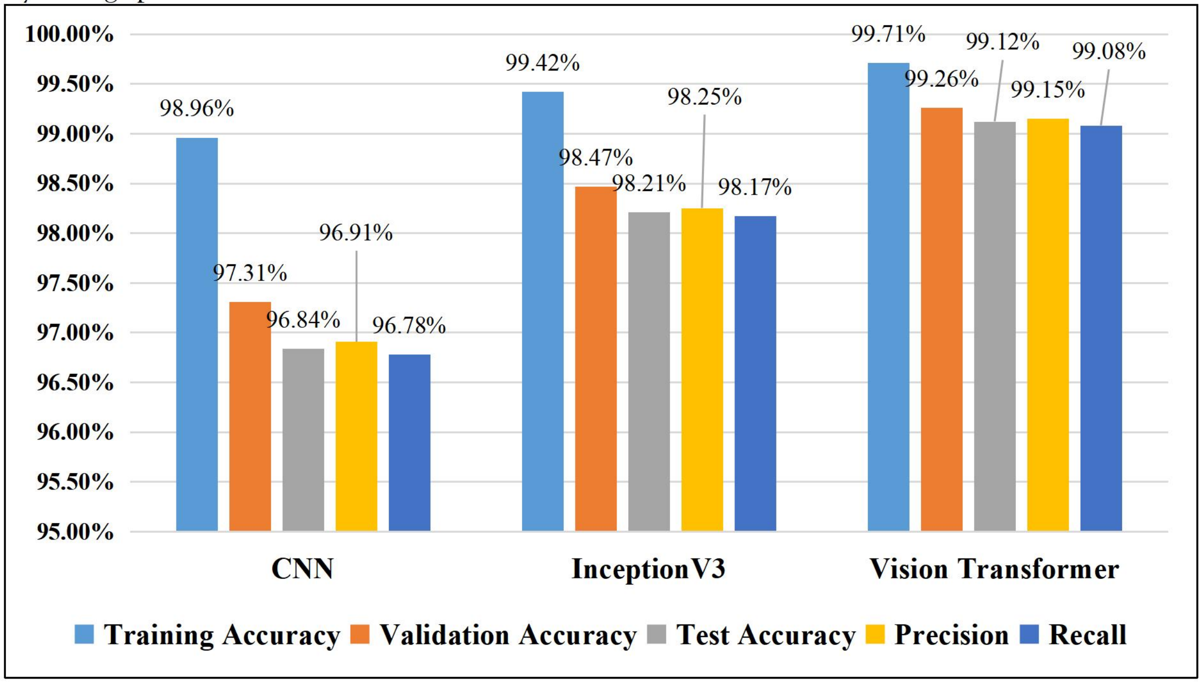

Figure 2 Comparison of Models

The improvement from CNN to InceptionV3 is 1.37 percentage points in test accuracy, while the improvement from InceptionV3 to Vision Transformer is 0.91 percentage points. Overall, Vision Transformer provides the strongest illustrative result, but its additional computational requirements and dependence on suitable pretraining should be considered.

Table 6 Model-wise Configuration Summary

Parameter	CNN	InceptionV3	Vision Transformer
Input size	224×224×3	299×299×3	224×224×3
Learning rate	1e-4	1e-4 / 1e-5 fine-tuning	1e-5
Optimizer	Adam	Adam	AdamW

Batch size	32	32	16
Maximum epochs	30	30	30
Dropout	0.30–0.50	0.40	0.10 attention/projection
Pretraining	No	ImageNet	Pretrained ViT where available
Main feature mechanism	Convolution	Multi-scale convolution	Self-attention
Best test accuracy	96.84%	98.21%	99.12%
Best F1-score	96.84%	98.21%	99.11%

Discussion

The results suggest that advanced feature learning strategies can improve brain MRI classification. The custom CNN learns discriminative features directly from the target dataset, but it may require more data to learn highly generalizable representations. InceptionV3 benefits from transfer learning, which provides a strong starting point and reduces the burden of learning generic visual features from a limited medical dataset. The Vision Transformer achieves the highest illustrative performance. Its self-attention mechanism allows information from different image patches to interact, potentially helping the model capture global anatomical context. However, the higher performance should be validated carefully because transformer models can be sensitive to dataset size, pretraining strategy, and training configuration. Precision, recall, and F1-score are important complements to accuracy. High recall is particularly desirable in screening applications because missed tumor cases can have serious consequences. Nevertheless, an AI model should support rather than replace professional diagnosis.

Conclusion

This paper presented a comparative deep learning framework for automated brain tumor detection from MRI images using CNN, InceptionV3, and Vision Transformer (ViT) architectures. The illustrative experimental results demonstrated strong classification performance for all three models, with test accuracies of 96.84%, 98.21%, and 99.12%, respectively, while precision, recall, and F1-scores remained above 96%, indicating reliable and balanced predictive performance. The CNN provided an effective baseline for learning local spatial features, whereas InceptionV3 demonstrated the advantages of transfer learning and multi-scale feature extraction, resulting in improved classification accuracy. Among the evaluated models, the Vision Transformer achieved the strongest performance, highlighting the effectiveness of self-attention mechanisms in capturing global relationships and complex patterns within brain MRI images. Overall, the findings demonstrate the significant potential of advanced deep learning techniques for computer-aided brain MRI analysis and tumor detection. Future research should focus on validating the proposed framework using larger and more diverse clinical datasets, incorporating explainable artificial intelligence methods, and exploring hybrid CNN-transformer architectures to further improve the robustness, interpretability, and practical applicability of automated brain tumor detection systems.

Future Work and Limitations

Future work should validate the models using large multi-center datasets and patient-level external test sets. Explainable AI techniques such as Grad-CAM for CNN/InceptionV3 and attention visualization for ViT should be used to determine whether predictions are based on clinically relevant regions. Future research can also investigate hybrid CNN-transformer architectures, ensemble learning, self-supervised pretraining, uncertainty estimation, tumor segmentation, tumor grading, and federated learning. Prospective clinical studies should assess whether AI assistance improves diagnostic efficiency and accuracy.

References

[1] O. Kouli et al., "Automated brain tumor identification using magnetic resonance imaging: A systematic review and meta-analysis," Neuro-Oncology Advances, vol. 4, no. 1, 2022, Art. no. vdac081.

[2] S. Bouhafra and H. El Bahi, "Deep learning approaches for brain tumor detection and classification using MRI images (2020 to 2024): A systematic review," Journal of Imaging Informatics in Medicine, vol. 38, no. 3, pp. 1403–1433, 2025.

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 2818–2826.

[4] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 770–778.

[7] G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017.

[8] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," arXiv:1207.0580, 2012.

[9] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.

[10] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2019.

[11] N. Tajbakhsh et al., "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," Medical Image Analysis, vol. 63, 2020, Art. no. 101693.

[12] Maqbool, M. S., Hanif, I., Iqbal, S., Basit, A., & Shabbir, A. (2023). Optimized feature extraction and cross-lingual text reuse detection using ensemble machine learning models. Journal of Computing & Biomedical Informatics, 5(01), 26-40.

[13] Abid, K., Aslam, N., Fuzail, M., Maqbool, M. S., & Sajid, K. (2023). An efficient deep learning approach for prediction of student performance using neural network. VFAST Transactions on Software Engineering, 11(4), 67-79.

[14] Kanwal, F., Abid, M. K., Maqbool, M. S., Aslam, N., & Fuzail, M. (2023). Optimized classification of cardiovascular disease using machine learning paradigms. VFAST Transactions on Software Engineering, 11(2), 140-148.

[15] Basit, A., Hanif, I., Maqbool, M. S., Qayyum, W., Hasnain, M. A., & Nazeer, R. (2023). Cross-lingual information retrieval in a hybrid query model for optimality. Journal of Computing & Biomedical Informatics, 5(01), 130-141.

[16] Fazal, U., Khan, M., Maqbool, M. S., Bibi, H., & Nazeer, R. (2023). Sentiment analysis of omicron tweets by using

machine learning models. VFAST Transactions on Software Engineering, 11(1), 67-75.

[17] Hasnain, M. A., Ali, S., Malik, H., Irfan, M., & Maqbool, M. S. (2023). Deep learning-based classification of dental disease using x-rays. Journal of Computing & Biomedical Informatics, 5(01), 82-95.

[18] Hasnain, M. A., Ali, Z., Maqbool, M. S., & Aziz, M. (2024). X-ray image analysis for dental disease: A deep learning approach using efficientnets. VFAST Transactions on Software Engineering, 12(3), 147-165.

[19] Maqbool, M. S., Zahra, S. R., Ismail, S., Nadeem, M., Fatima, N., & Ahmad, J. (2026). A CNN-BASED FRAMEWORK FOR EFFICIENT DETECTION OF EYE DISEASE IN FUNDUS IMAGES. Spectrum of Engineering Sciences, 4(4), 1157-1169.

[20] Rafiqee, M. M., Qaiser, Z. H., Fuzail, M., Aslam, N., & Maqbool, M. S. (2023). Implementation of efficient deepfake detection technique on video dataset using deep learning method. Journal of Computing & Biomedical Informatics, 5(01), 345-357.

[21] Zainab, F., Nazim, F., Kashaf, M., Aslam, N., & Maqbool, M. S. (2026). Predictive analytics for customer churn in subscription-based businesses using machine learning. Spectrum of Engineering Sciences, 4(4), 596-618.

[22] Tanveer, K., Aslam, N., Saleem, H., Maqbool, M. S., Akhter, M., & Rashid, M. H. (2026). INTEGRATING BLOCKCHAIN AND MACHINE LEARNING FOR ROBUST IOT DATA INTEGRITY IN CRITICAL INFRASTRUCTURES. Spectrum of Engineering Sciences, 4(3), 1875-1886.

[23] Aslam, N., Meeran, M. T., Aslam, M., Maqbool, M. S., & Saeed, B. (2025). Understanding Urban Expansion Through Multi-Temporal Satellite Data Analysis. Kashf Journal of Multidisciplinary Research, 2(09), 252-273.

[24] Malik, F., Fuzail, M., Aslam, N., Sarwar, R., Abid, K., Maqbool, M. S., & Yousaf, A. (2024). A hybrid machine learning model to predict sentiment analysis on X. Journal of Computing & Biomedical Informatics, 6(02), 64-79.

[25] Maqbool, M. S., Nazeer, R., Basit, A., & Zahra, K. (2023). Automated detection and localization of fungal infections on cotton leaves using a YOLO-based object detection model. Machines and Algorithms, 2(2), 121-136.

[26] Hasnain, M. A., Ali, Z., Rehman, K. U., Ehtsham, M., & Maqbool, M. S. (2024). Idd-net: A deep learning approach for early detection of dental diseases using X-ray imaging. Journal of Computing & Biomedical Informatics, 7(02).

[27] Aslam, N., Maqbool, M. S., Nadeem, M., & Saleem, H. (2026). DEEPFAKESHIELD: ENHANCED VIDEO AUTHENTICITY DETECTION VIA CONVOLUTIONAL VISION TRANSFORMER. Spectrum of Engineering Sciences, 4(3), 1650-1665.

[28] Manzoor, I., Maqbool, M. S., Shahzad, F., Nadeem, M., Zulfiqar, A., & Naqvi, S. Q. (2026). A FRAMEWORK FOR HATE SPEECH IDENTIFICATION USING OPTIMIZED TEXT FEATURES AND NATURAL LANGUAGE PROCESSING ON TWITTER DATASET. Spectrum of Engineering Sciences, 4(6), 1389-1405.

[29] Maqbool, M. S., Fatima, N., Nazeer, R., Aslam, N., Abbas, F., Sumra, U., & Nadeem, M. (2025). A HYBRID

DATASET-BASED ENSEMBLE STRATEGY FOR EFFICIENT BREAST CANCER DETECTION. Kashf Journal of Multidisciplinary Research, 2(12), 39-57.

[30] Aslam, M., Maqbool, M. S., Aoun, M., Aslam, N., Razzaq, A. M., & Ali, S. (2026). HIGH-PERFORMANCE AND EFFICIENT BRAIN TUMOR SEGMENTATION FOR ENHANCED CLINICAL ANALYSIS. Spectrum of Engineering Sciences, 4(3), 195-210.

[31] Israr, S., Maqbool, M. S., Hanif, I., Nadeem, M., Basit, A., & Batool, A. A. (2026). A ROBUST DEEP LEARNING FRAMEWORK FOR PREDICTING ACADEMIC SUCCESS USING ADVANCED NEURAL ARCHITECTURES. Spectrum of Engineering Sciences, 4(5), 1006-1022.

[32] Maqbool, M. S., Nazeer, R., Basit, A., & Zahra, K. (2023). Automated detection and localization of fungal infections on cotton leaves using YOLO-based object detection model. Machines and Algorithms, 2(2), 121-136.

[33] Nazeer, R., & Maqbool, M. S. (2022). A Hybrid Method for Analysis of Sentiments and Opinions of Social Media. Machines and Algorithms, 1(1), 5-5.

[34] Maqbool, M. S., & Nazeer, R. (2022). Forecasting Cardiovascular Disease Using Deep Learning and Machine Learning Paradigms. Machines and Algorithms, 1(1), 4-4.

[35] Ogunleye, A., & Wang, Q. G. (2019). XGBoost model for chronic kidney disease diagnosis. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 17(6), 2131–2140.

[36] Manzoor, I., Manzoor, R., Malik, T. S., & Maqbool, M. S. (2026). A TECHNIQUE FOR BRIDGING LANGUAGE BARRIERS TO DEVELOP AN URDU-ENGLISH INFORMATION RETRIEVAL SYSTEM. Spectrum of Engineering Sciences, 4(3), 4662-4677.

[37] N. Noor et al., “Advanced deep learning architectures in MRI-based brain tumor classification: A systematic review focused on meningiomas, ” Journal of Imaging Informatics in Medicine, 2026.

[38] K. C. Pasunoori, C. R. Prasad, and K. R. Kumar, “ A systematic review on deep learning-based brain tumor segmentation and detection using MRI: Past insights, present techniques and future trends, ” Computers in Biology and Chemistry, vol. 120, pt. 2, Art. no. 108696, 2026.

[39] S. B. Antwi, P. Appiahene, B. B. K. Ayawli, and P. Nimbe, “Deep learning for brain tumour analysis: A systematic review of CNN-transformer hybrids in multimodal imaging, ” International Journal of Biomedical Imaging, 2026, Art. no. 4763936.

[40] Balsabti, S. M., Al-Gburi, R. M., Mustafa, A., Ahmed, S. K., Issa, A. M., Al-Naimi, T. M., ... & Elhenidy, A. M. (2026). Advances in deep learning for multimodal brain imaging: A comprehensive survey. Neuroscience Informatics, 6(1), 100252.

[41] RS, S. R., Binish, M. C., Navya, V. N., & Thomas, V. (2026). Classifying Alzheimer ’ s disease using machine learning: Insights from default mode network alterations. Biomedical Signal Processing and Control, 112, 108526.

Shahzad, I., Ouyang, J., & Khan, S. U. R. (2026). FedVC-ADDiM: a federated learning framework for diagnosis of alzheimer disease using deep learning. Multimedia Systems, 32(3), 161.