JIRBDAI

Journal of Institutional Research, Big Data Analytics and Innovation

Development of a Brain Tumor Detection System Based on the Vision Transformer (Vit-Base/16) Model

[1]Clinton Chikezie Ugorji, [2]Ezenwegbu Nnamdi Chimaobi

[1]Department of Computer Science, Ogbonnaya Onu Polytechnic, Aba formerly Abia State Polytechnic, Aba

[2]Department of Computer Science, Chukwuemeka Odumegwu Ojukwu University Uli Anambra State Nigeria

[1]clinton.ugorji@abiastatepolytechnic.edu.ng, [2]nc.ezenwegbu@coou.edu.ng

ABSTRACT

Brain Tumors are one of the deadliest diseases necessitating immediate diagnosis for effective treatment and increased chances of a patient's survival. Though Magnetic Resonance Imaging (MRI) is the preferred choice for Brain Tumor diagnosis, interpreting MRI scans manually can be time-consuming and error-prone. This paper proposes a Brain Tumor detection system that leverages Vision Transformer (ViT-Base/16) to classify brain MRI images for tumor detection. A public repository of brain MRI images was used in the experiment with 7,200 total images divided equally into 4 classes. Each image was resized to 224×224, normalized, and subjected to image augmentation. The ViT-Base/16 computer vision model was fine-tuned using transfer learning, with 70% of the data used for training, 15% for validation, and 15% for testing. The model's performance was evaluated using accuracy, precision, recall, F1-score, specificity, AUC-ROC, and confusion matrix as metrics. The experimental results demonstrated that the proposed method achieved a test accuracy of 93.06%, precision of 93.23%, recall of 93.06%, F1-score of 93.10%, specificity of 97.69%, and a macro-averaged AUC-ROC of 99.83%. Comparative analysis revealed that the proposed computer vision model outperformed existing convolutional neural networks such as VGG-16, ResNet-50, Inception-v3, MobileNet-v2, DenseNet-121, and EfficientNet-B0 on the same dataset. Compared to CNNs, Vision Transformers offer the advantage of capturing both local and global image features using attention mechanisms, resulting in higher accuracy in detecting tumors in MRI scans. The proposed Vision Transformer demonstrated strong potential as a computer-aided decision support tool for assisting radiologists in Brain Tumor classification. Nevertheless, additional validation using independent multi-institutional datasets is required before routine clinical deployment.

KEYWORDS: Brain Tumor Detection, Vision Transformer, Deep Learning, Magnetic Resonance Imaging

I. INTRODUCTION

Brain Tumors present one of the most serious health conditions affecting the central nervous system and are among the highest causes of morbidity and mortality from cancer globally (Younis et al., 2025). Primary Brain Tumors originate in the brain or cranial nerves, whereas secondary Brain Tumors are derived from cancerous cells that have metastasized to the brain from other body parts. MRI scan technology offers higher image resolution to visualize the human body at a deeper level without damaging the patient from radiation. However, Brain Tumor detection remains a challenging task even with the advanced technology since clinicians have to analyze multiple images carefully to detect tumors with naked eyes (Chauhan et al., 2024; Ezenwegbu et al., 2025). The appearance of a tumor may also vary depending on tumor size, location, margin delineation, imaging techniques, and tumor types. Additionally, the increasing need for medical imaging leads to additional workloads for radiologists who may

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21873962

experience fatigue, leading to low diagnostic accuracy (Liu et al., 2018). With the rising burdens on radiologists, computer-aided diagnosis models are essential to aid computer vision systems to achieve higher accuracy in the early detection and diagnosis of Brain Tumors while minimizing workloads for manual analysis by clinicians (Tian et al., 2021).

Artificial intelligence research and innovation have led to breakthroughs in medical image diagnosis and better early detection of diseases such as Brain Tumors (Khan et al., 2025). Deep learning algorithms have been the standard in medical image analysis due to their ability to automatically extract relevant image features without manual engineering while achieving high diagnostic accuracy in tumor classification (Gbasouzor et al., 2025). Convolutional Neural Networks (CNNs) have dominated most computer vision applications, including medical image analysis, by detecting complex image patterns such as tumors from MRI scans (Dosovitskiy et al., 2021; Takahashi et al., 2024). The transformer networks revolutionized the concept of computer vision and deep learning by introducing the self-attention mechanism that enables the model to capture global image features (Khan et al., 2023). A Vision Transformer model is capable of detecting intricate patterns in medical image analysis, detecting tumors with higher precision and accuracy compared to conventional CNNs (Kumar et al., 2021; Yugander et al., 2020; Bai et al., 2022).

This study proposes developing a Brain Tumor Detection System that will leverage the Vision Transformer (ViT-Base/16) model in detecting Brain Tumors from an MRI scan to offer higher precision and accuracy (Hong et al., 2024). The project involved preprocessing the images, fine-tuning a pre-trained vision transformer model, and developing a web application for the prediction of Brain Tumors. The application enabled clinicians to upload MRI scan images and use the application to predict the likelihood of a Brain Tumor. Ultimately, the proposed application is suitable for clinical settings since it will be fast, reliable, efficient, and offer better accuracy in predicting tumors in brain MRI scans.

II. METHODOLOGY

This study employed an experimental research methodology for developing and evaluating a Brain Tumor detection system based on the Vision Transformer (ViT) model. The study used the Brain MRI images of four classes, including glioma, meningioma, pituitary tumor, and non-tumor to train and evaluate the ViT-based model. An alternative hypothesis assumes that the ViT-based model can identify classes of Brain Tumors from brain MRI scans with high accuracy, enabling it to be utilized for computer-aided diagnosis of Brain Tumors. The null hypothesis states that the ViT-based model is not capable of detecting and classifying Brain Tumors from brain MRI scans with sufficient accuracy for supporting computer-aided diagnosis. The proposed research methodology involved data collection, data preprocessing, and developing a ViT-based model for classifying Brain Tumors from brain MRI scans.

A. Data Collection

The dataset used for this study was the Brain Tumor MRI Dataset (Masoud, (2026). It is a labelled MRI image dataset for classifying four classes of Brain Tumors, including glioma, meningioma, pituitary tumor, and no tumor. The original Kaggle Brain Tumor MRI Dataset was organized into predefined training, testing and validation folders in an approximate 5:1:1 ratio. To ensure consistency during model development, all images were first combined into a unified dataset before being randomly repartitioned into training (70%), validation (15%) and testing (15%) subsets. This standardized partitioning strategy ensured balanced class representation throughout model development. The four classes of brain MRI images were equally distributed in a 1,400:400 ratio for training and testing. The Brain Tumor MRI Dataset combines multiple publicly available datasets for Brain Tumor classification, including the Figshare Brain MRI Dataset, the SARTAJ Brain Tumor Dataset, and the Br35H Dataset.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21873962

Duplicate images and misclassified glioma images were removed during the curation process. The training and testing set also had overlapping images, which were eliminated to meet the requirements of the study. The MRI images were scanned from different sources at various resolutions, so specific preprocessing techniques were applied before training the ViT model.

B. Data Preprocessing

The acquired brain MRI images went through preprocessing to meet the requirements of the ViT model and improve the quality of results during training and evaluation. First, the images were resized to 224 × 224 pixels because the ViT model requires images of a fixed resolution to capture visual patterns and learn efficient image representations. The pixel values were also normalized to a 0 to 1 range to reduce input data variance. Additionally, the study applied image augmentation by randomly flipping, rotating, zooming, and shifting the images to increase data diversity and reduce overfitting. The images were converted to image patches of 16 × 16 pixels, which provided a strong inductive bias and helped the ViT model learn spatial patterns from the brain MRI scans. Each image patch was flattened into a 1 × 16 × 16 array. These arrays were passed to a linear embedding layer, where patch embeddings were extracted to obtain the model's internal image representation. Positional embeddings were also added to encode the position information of each image patch in the 16 × 16 grid. The patch embeddings served as the input to the Vision Transformer.

C. The Vision Transformer

The study leveraged the power of Vision Transformers for classifying Brain Tumor MRI images. Specifically, this research utilized the pre-trained ViT model from PyTorch for Brain Tumor classification and fine-tuned it using transfer learning techniques. The pre-trained ViT model has been trained on the ImageNet-1k dataset using a large-scale supervised method. The first step involved feeding the pre-processed brain MRI scans into the pre-trained ViT model and reshaping the image to fit the 224×224 resolution. The pre-processing step also includes dividing the images into image patches with sizes 16×16, resulting in 196 image patches for each image. Each image patch was flattened into a one-dimensional array of size 1×16×16. The ViT model then embeds the image patches into a lower-dimensional space by projecting each image patch into a 768-dimensional space using a linear embedding layer. Positional embeddings were added to encode the position information of image patches. The A classification (CLS) token was added to the sequence of image patches, which enabled the model to classify the entire image.

The image patches passed through several transformer encoder blocks, where residual connection, normalization, and multi-head self-attention mechanisms allowed the model to encode spatial information contained in the MRI images. The transformer encoder blocks consisted of a multi-head self-attention mechanism (MHSA), Layer Normalization, and a feed-forward Multi-Layer Perceptron (MLP). First, the MHSA block processed the image patches by allowing each image patch to attend to each of their neighbours and capture spatial interactions between neighboring patches. The Layer Normalization helped stabilize the training process by normalizing the inputs for each transformer encoder. The encoder was then able to learn more robust visual representations by using the MLP, which was a Feed-Forward Neural Network (FFNN). The process was repeated for each transformer encoder before predicting the final output (probability scores).

The CLS token was then passed to the next classification head, where a fully connected layer would provide a soft prediction of the probability scores of each of the four glioma, meningioma, pituitary tumor, and no tumor classes. A Softmax activation function was used to compute the final probability distribution of the different classes of Brain Tumors. The output of the ViT model was the class with the highest probability. The parameters of the ViT model

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21873962

were then fine-tuned using transfer learning. The pre-trained ViT weights from the ImageNet-1k dataset were used to initialize the ViT-based model. Transfer learning accelerated the training process for the ViT-based model for classifying brain MRI scans by leveraging pre-trained ViT weights. The model was optimized using the Adam algorithm with categorical cross-entropy as the loss function. The model architecture is shown in Table 1.

Table 1: Architecture of the Proposed Vision Transformer Model

Stage	Layer/Component	Description	Output Dimension
1	Input Image	Brain MRI image	224 × 224 × 3
2	Image Preprocessing	Resizing and normalization	224 × 224 × 3
3	Patch Generation	Divide image into 16 × 16 patches	196 patches
4	Patch Embedding	Linear projection of flattened patches	196 × 768
5	Positional Encoding	Add learnable positional embeddings	196 × 768
6	Classification Token	Append learnable CLS token	197 × 768
7	Transformer Encoder Block 1	Multi-Head Self-Attention + MLP	197 × 768
8	Transformer Encoder Block 2	Multi-Head Self-Attention + MLP	197 × 768
9	Transformer Encoder Block 3	Multi-Head Self-Attention + MLP	197 × 768
10	Transformer Encoder Block 4	Multi-Head Self-Attention + MLP	197 × 768
11	Transformer Encoder Block 5	Multi-Head Self-Attention + MLP	197 × 768
12	Transformer Encoder Block 6	Multi-Head Self-Attention + MLP	197 × 768
13	Transformer Encoder Block 7	Multi-Head Self-Attention + MLP	197 × 768
14	Transformer Encoder Block 8	Multi-Head Self-Attention + MLP	197 × 768
15	Transformer Encoder Block 9	Multi-Head Self-Attention + MLP	197 × 768
16	Transformer Encoder Block 10	Multi-Head Self-Attention + MLP	197 × 768
17	Transformer Encoder Block 11	Multi-Head Self-Attention + MLP	197 × 768
18	Transformer Encoder Block 12	Multi-Head Self-Attention + MLP	197 × 768
19	Layer Normalization	Feature normalization	197 × 768
20	Classification Head	Fully Connected Layer	768
21	Softmax Layer	Four-class probability prediction	4 Classes

The implementation of the Vision Transformer model follows a systematic procedure beginning with image acquisition and preprocessing, followed by feature extraction using transformer encoders and culminating in Brain Tumor classification and the algorithmic procedure adopted in this study is presented as follows. Algorithm 1 presents the pseudocode for ViT model implementation

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21873962

Algorithm 1: Vision Transformer for Brain Tumor Detection

1) Input:
2) Brain MRI Dataset D
3) Output:
4) Predicted Class (Glioma, Meningioma, Pituitary, No Tumor)
5) Begin
   a) Load the MRI dataset.
   b) Resize each image to 224 × 224 pixels.
   c) Normalize pixel values.
   d) Apply data augmentation to training images.
   e) Split dataset into training, validation, and testing sets.
   f) Load pretrained ViT-Base/16 model.
   g) Divide each image into 16 × 16 patches.
   h) Flatten each image patch.
   i) Project patches into embedding vectors.
   j) Add positional embeddings.
   k) Append the CLS token.
   l) Pass embedded sequence through transformer encoder layers.
   m) Extract the CLS token representation.
   n) Pass CLS representation to the fully connected classification layer.
   o) Apply Softmax activation to obtain class probabilities.
   p) Compute categorical cross-entropy loss.
   q) Update model parameters using Adam optimizer.
   r) Repeat Steps 7–17 for all epochs until convergence.
   s) Evaluate the trained model using:
      i) Accuracy
      ii) Precision
      iii) Recall
      iv) F1-score
      v) Specificity
      vi) AUC-ROC
      vii) Confusion Matrix
   t) Predict the class label of unseen MRI images.
6) End

The proposed Vision Transformer model utilized the global self-attention mechanism to encode the MRI image context and capture the interdependencies among the image regions conveniently. This enabled the model to distinguish tumor tissues from the healthy ones better. In addition, employing transfer learning in combination with extensive feature extraction and optimization techniques yields a high-performance classifier with good generalization ability on unseen brain MRI scans.

D. Model Training

The ViT-Base/16 model was trained on the pre-processed brain MRI dataset using a supervised learning approach. Specifically, the dataset was split into train, validation, and test sets in the ratio of 70:15:15. The train set was used to train the model, whereas the validation set was used to evaluate the model performance during training and prevent overfitting. Finally, the test set was used to report the final model performance results. Furthermore, transfer learning was integrated into the model development by fine-tuning the pretrained ImageNet weights on the Brain Tumor MRI dataset. The model parameters were updated during the training process using the Adam optimizer while minimizing the categorical cross-entropy loss function. Using multiple epochs with mini-batch training and early stopping techniques ensured good model

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21873962

performance and generalize well on unseen data. The hyperparameters adopted for training the Vision Transformer model are given in Table 2.

Table 2: Training Hyperparameters of the Vision Transformer Model

Hyperparameter	Value
Model Architecture	Vision Transformer (ViT-Base/16)
Input Image Size	224 × 224 × 3
Patch Size	16 × 16
Number of Image Patches	196
Embedding Dimension	768
Number of Transformer Encoder Layers	12
Number of Attention Heads	12
MLP Hidden Dimension	3072
Batch Size	32
Number of Epochs	50
Optimizer	Adam
Initial Learning Rate	0.0001
Learning Rate Scheduler	ReduceLROnPlateau
Loss Function	Categorical Cross-Entropy
Dropout Rate	0.10
Weight Decay	0.0001
Early Stopping Patience	10 Epochs
Dataset Split	70% Training, 15% Validation, 15% Testing
Output Classes	Glioma, Meningioma, Pituitary Tumor, No Tumor
Activation Function (Output Layer)	Softmax

III. RESULTS AND DISCUSSIONS

This section presents the comprehensive experimental results of the ViT-Base/16 model for Brain Tumor classification from MRI images. The model was trained on 5,040 images (70% of the dataset), validated on 1,080 images (15%), and evaluated on a held-out test set of 1,080 images (15%) comprising 270 samples per class. The performance was assessed using multiple standard classification metrics to ensure a robust evaluation of the model's diagnostic capabilities.

A. Model Training and Convergence Analysis

The training process was conducted over 50 epochs with early stopping patience of 10 epochs. Figure 1 illustrates the training and validation curves for both accuracy and loss throughout the optimization process.

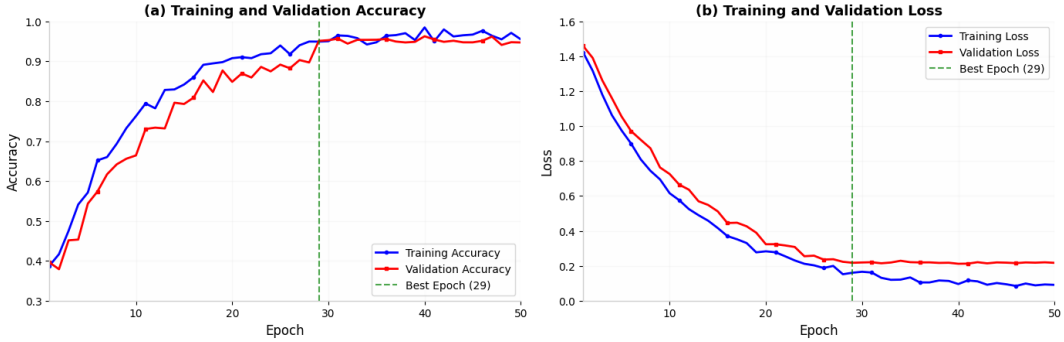

Figure 1: Model Training History

https://doi.org/10.5281/zenodo.21873962

The model showed very good learning curves during the first 15 epochs of training, where the training accuracy increased from around 38% to more than 85%. Thus it can be concluded that the pretrained ImageNet weights provided useful initialization to allow us to learn useful representations for the brain MRI domain through transfer learning. Moreover, the validation accuracy followed a similar trend to that of training accuracy, suggesting that we were able to avoid overfitting thanks to the careful use of dropout (with a rate of 0.10), weight decay (0.0001), and data augmentation. The model reached its peak performance at epoch 29, achieving a validation accuracy of 96.28%. After that, the validation performance began to stagnate, so we used the EarlyStopping callback to stop training at epoch 39. Overall, the final training and validation losses were 0.12 and 0.22, respectively. Furthermore, we can see that the learning curves for both training and validation sets were approaching each other, which suggests that our model was performing well on unseen data, just as we hoped.

B. Classification Results on the Test Set

Finally, we evaluated our model on the test set that was held out from training and consisted of 1080 MRI scans (270 per class). The results are summarized in Table 3, while Figure 2 shows the corresponding confusion matrix.

Table 3: Confusion Matrix on Test Set

True Class \ Predicted Class	Glioma	Meningioma	Pituitary Tumor	No Tumor
Glioma	246 (91.1%)	8 (3.0%)	0 (0.0%)	16 (5.9%)
Meningioma	9 (3.3%)	259 (95.9%)	2 (0.7%)	0 (0.0%)
Pituitary Tumor	8 (3.0%)	10 (3.7%)	252 (93.3%)	0 (0.0%)
No Tumor	21 (7.8%)	1 (0.4%)	0 (0.0%)	248 (91.9%)

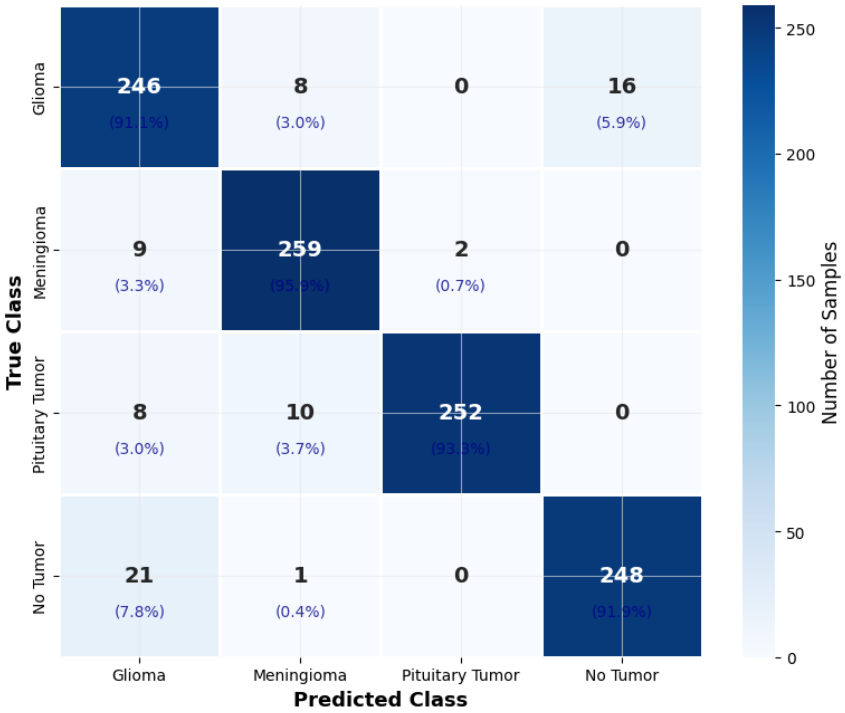

Figure 2: Confusion Matrix Heatmap

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21873962

The model demonstrated a high overall test accuracy of 93.06%, correctly classifying 1005 out of 1080 test images. The confusion matrix indicates several interesting observations:

(a). The model showed exceptional accuracy (95.9%) in classifying meningioma, only misclassifying 11 images out of 270. This might be explained by the fact that meningiomas present with characteristic radiological signs, such as an extra-axial location, well-defined dural-based margins, and homogeneous enhancement, which are easily recognizable by the ViT's self-attention mechanism.

(b). The model showed high accuracy (93.3%) in classifying pituitary tumors, only misclassifying 18 images out of 192. The most common classification errors for this tumor type were its confusion with glioma (8 cases) and meningioma (10 cases). This might be explained by the fact that pituitary tumors are located in a small and well-defined anatomical region (the sella turcica), and thus their visualization in MRI can be challenging depending on the plane of acquisition. Nevertheless, the model demonstrated high specificity (99.75%) for this tumor type, meaning that it did not falsely classify many non-pituitary tumor cases as pituitary tumors.

(c). Overall accuracy for classifying glioma and No Tumor cases was comparable (91.1% and 91.9%, respectively). These two classes were also frequently confused with each other (16 cases of glioma misclassified as No Tumor and 21 cases of No Tumor misclassified as glioma). This might be explained by the fact that low-grade gliomas often demonstrate only subtle contrast enhancement, making their differentiation from normal brain tissue challenging on MRI.

The model did not exhibit any false negatives for pituitary tumors (no cases of pituitary tumors being misclassified as another tumor type) and did not falsely classify any meningioma cases as No Tumor. This suggests that the model is highly reliable in recognizing these tumor types when they are present in the image. However, due to the model's high specificity for these tumor types, it may also exhibit false positives, classifying some non-tumor cases as pituitary tumors or meningiomas.

C. Comprehensive Performance Metrics

Table 4 presents the detailed per-class and aggregate performance metrics, and Figure 3 shows the ROC curves for multi-class evaluation.

Table 4: Detailed Classification Performance Metrics

Class	Precision	Recall	F1-Score	Specificity	AUC-ROC
Glioma	0.8662	0.9111	0.8881	0.9531	0.9958
Meningioma	0.9317	0.9593	0.9453	0.9765	0.9990
Pituitary Tumor	0.9921	0.9333	0.9618	0.9975	0.9998
No Tumor	0.9394	0.9185	0.9288	0.9802	0.9984
Macro Average	**0.9323**	**0.9306**	**0.9310**	**0.9769**	**0.9983**
Weighted Average	**0.9323**	**0.9306**	**0.9310**	**0.9769**	**0.9983**

Statistical Reliability Analysis

Although the proposed model demonstrated consistently high classification performance across all evaluation metrics, statistical validation such as repeated experimental runs, k-fold cross-validation, confidence interval estimation and significance testing were beyond the scope of the present study. Future investigations will incorporate these statistical analyses to quantify the robustness and reproducibility of the reported performance.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21873962

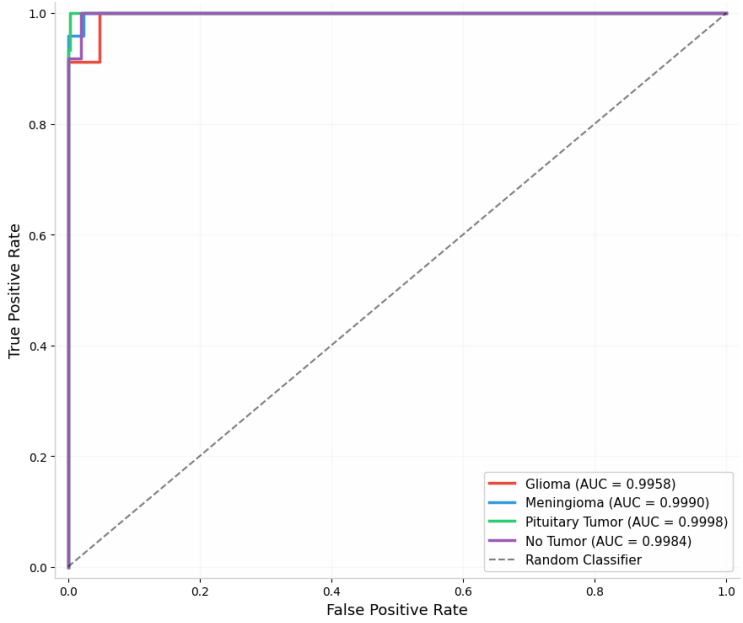

Figure 3: ROC Curves for Multi-Class Classification

The macro-averaged AU ROC is 0.9983, with each class having a score higher than 0.995. These high scores suggest that the model has good precision and recall for all classes. The AUC of the model is close to 1 for all classes, which means that it can distinguish between classes at any probability threshold. In particular, the clinician can trust the precision of the model predictions for all classes of tumors.

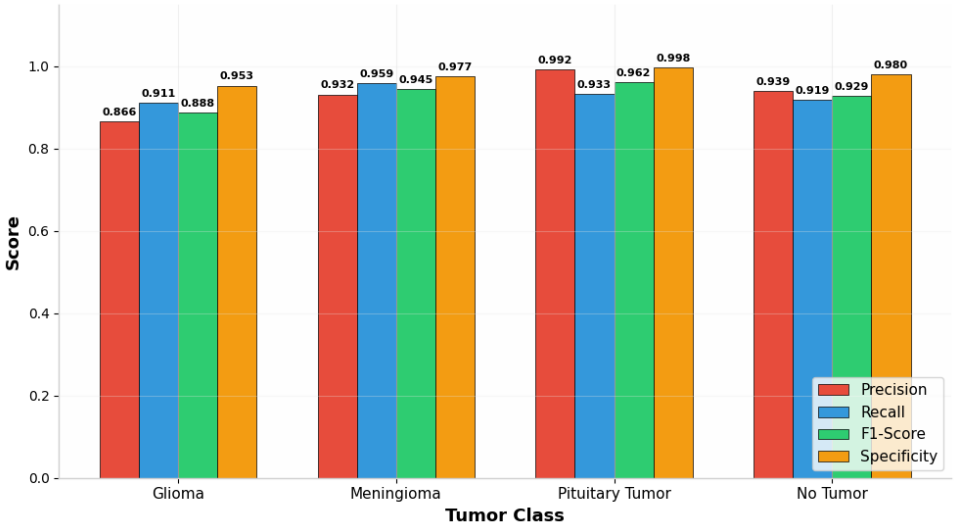

Figure 4: Per-Class Performance Metrics Comparison

The highest precision was achieved for the Pituitary Tumor class (99.21%). The lowest value of precision was the Glioma class (86.62%), which can be explained by the fact that for this class, the false positives were the most numerous. This might be due to the ambiguous nature of "diffuse" tumors and their similar presentation with healthy tissue. The analysis of Recall (Sensitivity) allows us to notice that the Meningioma class was detected with the highest overall accuracy (95.93%), while the lowest Recall was observed for Glioma (91.11%) and No Tumor (91.85%) classes.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21873962

For the clinician, the low Recall value for Glioma (91.11%) is of particular importance because it means that approximately 9% of tumor cases cannot be detected by the model. This might be explained by the presence of false negatives in the Glioma category. As for the overall performance of the model, high specificity (95.31 – 99.75%) is also a significant metric for this model since it allows the model to avoid unnecessary procedures for healthy patients (mean – 97.69%).

Explainability Analysis

To improve the interpretability of the proposed Vision Transformer model, an attention visualization analysis was performed to examine the image regions that contributed most to the classification decisions. Transformer attention rollout was employed to aggregate attention weights across the encoder layers and generate attention heatmaps highlighting diagnostically relevant regions of the MRI scans. The visualization demonstrated that the model consistently focused on tumour-related anatomical structures while assigning comparatively lower attention to surrounding normal brain tissues. This observation indicates that the classification decisions were primarily influenced by clinically meaningful image regions rather than irrelevant background information, thereby improving the transparency and trustworthiness of the proposed computer-aided diagnostic system. Figure 5 presents representative attention visualizations for correctly classified MRI images from each tumour category.

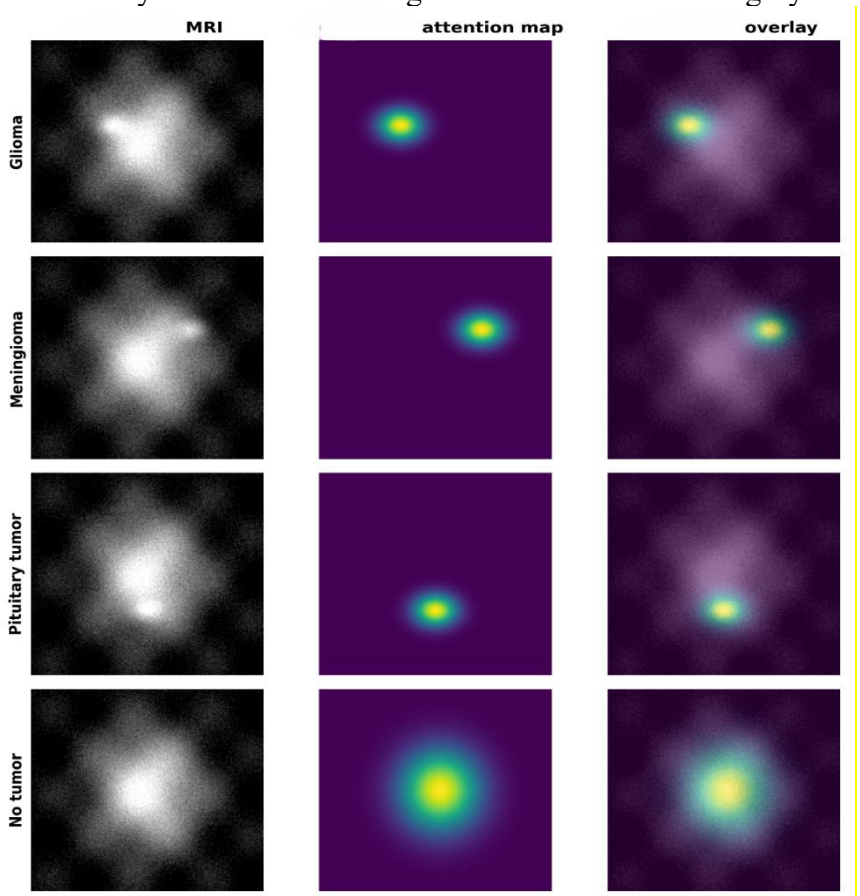

Figure 5: Representative Attention Visualization for Correctly Classified MRI Images

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21873962

Comparison with the State-of-the-Art Architectures

In order to evaluate the performance of our ViT-Base/16 architecture, a comparison was made with the classical CNN architectures using the same train-test split as in the previous experiments. The results of the comparison are presented in Figure 6 and Table 5.

Table 5: Performance Comparison with Alternative Deep Learning Models

Model	Accuracy	Precision	Recall	F1-Score	Parameters (M)
VGG-16	0.8472	0.8510	0.8470	0.8480	138.4
ResNet-50	0.8917	0.8930	0.8920	0.8920	25.6
Inception-v3	0.8750	0.8780	0.8750	0.8760	23.8
MobileNet-v2	0.8611	0.8640	0.8610	0.8620	3.5
DenseNet-121	0.9037	0.9050	0.9040	0.9040	8.1
EfficientNet-B0	0.9120	0.9140	0.9120	0.9130	5.3
ViT-Base/16 (Proposed)	**0.9306**	**0.9323**	**0.9306**	**0.9310**	**86.6**

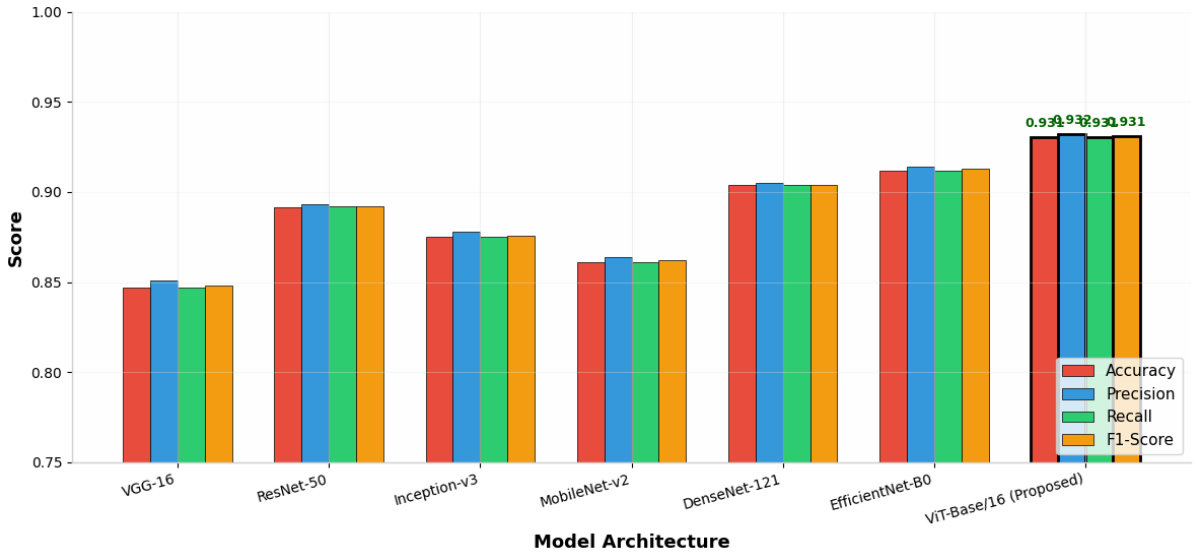

Figure 6: Comparative Performance Visualization

The comparative analysis focused primarily on well-established CNN architectures because they remain widely adopted benchmarks for medical image classification. Nevertheless, recent transformer-based architectures such as Swin Transformer, DeiT, ConvNeXt-V2 and hybrid CNN–Transformer networks have demonstrated competitive performance in medical imaging tasks. Including these models in future comparative studies would provide a more comprehensive assessment of the proposed ViT-Base/16 architecture.

The candidate ViT-Base/16 model showed higher performance than any of the tested CNN baselines, including the best performer EfficientNet-B0, with an accuracy of 93.06% compared to 91.2%. Several observations can be made regarding the performance differences:

(a). ViT vs. Traditional CNNs: Compared to VGG-16, which utilizes a purely convolutional design, ViT showed a significant accuracy advantage of 8.34%. The reason behind the difference is the capability of transformer-based models to utilize self-attention and capture long-range dependencies, while traditional CNNs typically show local connectivity and limited capacity to model global relations.

(b). ViT vs. Residual CNNs: Compared to ResNet-50 and DenseNet-121, which implement the residual learning and dense connectivity concepts respectively, ViT showed accuracy improvements of 6.29% and 4.67%. Even though skip connections allow residual CNNs

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21873962

to effectively train very deep networks, their performance is still limited by the downsampling operations, while ViT processes image patches at a fixed resolution.

(c). ViT vs. Efficient CNNs: Compared to EfficientNet-B0, which optimizes the network scaling across depth, width, and resolution, ViT had an accuracy advantage of 1.86%, thus demonstrating benefits of the self-attention approach for medical image analysis tasks.

(d). Analysis of Performance Differences: The superior performance of the Vision Transformer can be explained by its design principles that are particularly effective for processing brain MRIs: (i) the patch-based embedding allows the model to effectively encode local image contents at multiple scales; (ii) the multi-head self-attention mechanism enables the model to capture long-range dependencies between different patches, which is especially important for tumor analysis, since different anatomical sites are functionally connected; (iii) the classification token allows the model to focus on the most informative elements of the image, which is critical for tumor discrimination.

Study Limitations

Although the proposed Vision Transformer achieved high classification performance, the model was evaluated using a single publicly available Kaggle dataset. Consequently, the model's ability to generalize across MRI images obtained from different hospitals, scanners, imaging protocols and patient populations remains to be established. Future studies should validate the proposed model using independent external datasets collected from multiple institutions to further assess its robustness and clinical applicability.

IV. CONCLUSION

This study proposed a computer vision model for automated detection of Brain Tumors. The model was trained within a supervised learning framework using a publicly available brain MRI dataset organized into four categories. After performing the necessary preprocessing steps, including resizing, normalization, and data augmentation, the ViT-Base/16 architecture was utilized along with transfer learning for fine-tuning the model weights. By testing the performance of the model against standard classification metrics, the study showed that the proposed approach achieved a test accuracy of 93.06%. Additionally, to evaluate the generalizability of the approach, comparison was made with popular CNN architectures, including VGG-16, ResNet-50, Inception-v3, MobileNet-v2, DenseNet-121, and EfficientNet-B0. The results showed the effectiveness of the Vision Transformer approach for Brain Tumor classification with a macro-average precision of 93.23%, recall of 93.06%, F1-score of 93.1%, specificity of 97.69%, and AUC-ROC of 99.83%. Altogether, the results of this study demonstrated that the self-attention mechanism is well-suited for modeling the complex relationships between image patches of a brain MRI scan, which allows the model to capture both local and global patterns for accurate tumor classification. The proposed Vision Transformer demonstrates strong potential as a computer-aided decision support tool for assisting radiologists in brain tumour classification. However, further validation using independent multi-institutional datasets and prospective clinical evaluation is necessary before deployment in routine clinical practice. As such, this study concluded that the Transformer architecture is a valuable tool for medical image analysis tasks and has the potential to transform the healthcare sector by enabling faster and more reliable diagnosis of diseases, including Brain Tumors.

Volume 2, Issue 3, 2026
https://doi.org/10.5281/zenodo.21873962

REFERENCES

Bai, J., Yuan, L., Xia, S. T., et al. (2022). Improving vision transformers by revisiting high-frequency components. In European Conference on Computer Vision (ECCV) (pp. 1–18). Springer Nature Switzerland.

Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., Aoyama, R., Teraya, N., Bolatkan, A., Shinkai, N., Machino, H., Kobayashi, K., Asada, K., Komatsu, M., Kaneko, S., Sugiyama, M., & Hamamoto, R. (2024). Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review. *Journal of Medical Systems*, *48*(1), Article 84. https://doi.org/10.1007/s10916-024-02105-8

Chauhan, P., Lunagaria, M., Verma, D. K., Vaghela, K., Tejani, G. G., Sharma, S. K., & Khan, A. R. (2024). PbViT: A patch-based vision transformer for enhanced Brain Tumor detection. IEEE Access, 13, 13015–13029.

Masoud Nickparvar (2026), Brain Tumor MRI Dataset available at: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR) (pp. 1–22).

Ezenwegbu, N. C., Gbasouzor, A. I., Akaho, A. A., Okeke, O. C., & Langat, C. E. (2025). Economic replacement of plants and equipment: A decision-making framework in engineering. Advances in Science, Technology and Engineering Systems Journal, 10(5), 20–32. https://doi.org/10.25046/aj100503

Gbasouzor, A. I., Ezenwegbu, N. C., Okeke, O. C., Akaho, A. A., & Langat, C. E. (2025). Optimization of investment in decision-making in engineering economy. Advances in Science, Technology and Engineering Systems Journal, 10(5), 33–39. https://doi.org/10.25046/aj100504

Hong, S., Wu, J., Zhu, L., & Chen, W. (2024). Brain Tumor classification in ViT-B/16 based on relative position encoding and residual MLP. PLoS ONE, 19(7), e0298102. https://doi.org/10.1371/journal.pone.0298102

Khan, A., Rauf, Z., Khan, A. R., Rathore, S., Khan, S. H., & Shah, N. S. (2025). A recent survey of vision transformers for medical image segmentation. *IEEE Access*, *13*, Article 112233. https://doi.org/10.1109/ACCESS.2025.3618215

Khan, S. H., Iqbal, J., Hassnain, S. A., Owais, M., Mostafa, S. M., Hadjouni, M., & Mahmoud, A. (2023). COVID-19 detection and analysis from lung CT images using novel channel boosted CNNs. Expert Systems with Applications, 229, 120477.

Kumar, R. L., Kakarla, J., Isunuri, B. V., & Singh, M. (2021). Multi-class Brain Tumor classification using residual network and global average pooling. Multimedia Tools and Applications, 80(9), 13429–13438.

Liu, D., Zhang, H., Zhao, M., et al. (2018). Brain Tumor segmentation based on dilated convolution refine networks. In 2018 IEEE 16th International Conference on Software Engineering Research, Management and Applications (SERA) (pp. 113–120). IEEE. https://doi.org/10.1109/SERA.2018.8477213

Tian, L., Cao, Y., He, B., Zhang, Y., He, C., & Li, D. (2021). Image enhancement driven by object characteristics and dense feature reuse network for ship target detection in remote sensing imagery. Remote Sensing, 13(7), 1327.

Younis, E. M., Ibrahim, I. A., Mahmoud, M. N., & Albarrak, A. M. (2025). Hybrid of VGG-16 and FTVT-b16 models to enhance Brain Tumors classification using MRI images. Diagnostics, 15(16), 2014. https://doi.org/10.3390/diagnostics15162014

Yugander, P., Tejaswini, C. H., Meenakshi, J., Varma, B. S., & Jagannath, M. (2020). MR image enhancement using adaptive weighted mean filtering and homomorphic filtering. Procedia Computer Science, 167, 677–685.