Hybrid Vision Transformer and Mix Transformer Ensemble for Multiclass Brain Tumor Classification Using MRI Images

Hardik Grover[1]
Department of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur Campus, India
hg7243@srmist.edu.in

Viveka S[2]
Assistant Professor
Department of Computing Technologies
Faculty of Engineering and Technology
SRM Institute of Science and Technology
Kattankulathur Campus, India
vivekas@srmist.edu.in

Riddhman Singh[3]
Department of Computing Technologies
SRM Institute of Science and Technology
Kattankulathur Campus, India
rs1976@srmist.edu.in

***Abstract*—Manual radiological evaluation of brain MRI for tumor classification is computationally and cognitively intensive and often subject to inter-observer variability. Although CNN-based transfer learning has improved automation in medical image analysis, transformers have recently shown superior performance in capturing global context in medical imaging. This work proposes a hybrid deep learning ensemble that fuses Vision Transformer (ViT) global attention with Mix Transformer (MiT) hierarchical features using weighted logit combination ($\alpha = 0.6$ for ViT), the logits generated by each model before softmax operation are linearly combined. The final ensemble logits are computed as**

$$z_{ens} = \alpha z_{\mathrm{ViT}} + (1 - \alpha) z_{\mathrm{MiT}} \tag{1}$$

where z_{ViT} and z_{MiT} represent the logits produced by the ViT and MiT models, respectively, and $\alpha = 0.6$ determines the relative contribution of each model in the ensemble. The resulting combined logits are then passed through a softmax layer to obtain the final class probabilities for four-class brain tumor classification (glioma, meningioma, pituitary, and no tumor).

Key contributions include:

- **Novel ViT-MiT logit ensemble with validation-optimized weighting ($\alpha = 0.6$) and test-time augmentation (TTA)**
- **Progressive transfer learning on ImageNet-pretrained backbones using cosine LR scheduling with warmup and AMP**
- **99.01% test accuracy, 0.99 macro F1 on balanced 7,200-image public MRI dataset, outperforming standalone models**
- **3-fold stratified CV stability (ViT: $98.2 \pm 0.3\%$, MiT: $98.3 \pm 0.2\%$) with attention visualization confirming tumor-focused learning**

The framework demonstrates robust generalization and strong potential for automated brain tumor diagnosis.

***Index Terms*—Brain tumor classification, Vision Transformer, Mix Transformer, SegFormer, Ensemble learning, Transfer learning, MRI analysis, test-time augmentation, Explainability**

I. INTRODUCTION

Brain tumors remain one of the most life-threatening neurological disorders, where timely and accurate diagnosis is crucial for effective treatment planning and improved patient outcomes [6]. Magnetic Resonance Imaging (MRI) is widely adopted for brain tumor evaluation due to its excellent soft-tissue contrast and higher spatial resolution compared with other imaging modalities. MRI continues to serve as the primary imaging technique for brain tumor assessment.

In spite of these benefits, radiologists still require a considerable amount of time to perform manual analysis of MRI scans, and the method is also subjective. Differences in clinical experience, fatigue, and interpretational ambiguity may cause inconsistencies in diagnosis. The computer-aided diagnosis (CAD) systems have become potential supplements to expert evaluation. Recent studies validate the growing adoption of machine learning (ML) and deep learning (DL) techniques in implementing brain tumor classification, and the several studies showing consistent improvements in accuracy compared to the conventional methodologies [6].

Initial computer-aided diagnosis (CAD) methods are based on manual feature engineering including texture descriptors, histogram-based features, such as support vector machines (SVMs) and k-nearest neighbors (KNN). Even though the traditional machine learning performs well in controlled environments, they do not perform well in cases that exhibit noise in the medical image scans, and show limited ability to generalize to unseen datasets [4].

Convolutional neural networks (CNNs) make a significant contribution and enable automatic learning of hierarchical features. Transfer learning with architectures such as VGG, Inception, and ResNet has been shown to have high tumor classification accuracy with the help of transfer learning [3], [7], [8], [11]. ResNet50, specifically, has been progressively hypothesized as a solid foundation on typical brain MRI benchmarks [3], [7], [8], [11].

However, CNNs are inherently local in nature, which limits their ability to grasp long-range spatial information, which

is essential to abnormally shaped tumors. Transformer architectures have been converted for successful use in computer vision. ViT architecture captures long range dependencies by simply applying multi-head self-attention mechanisms over images that have been segmented into patches [1]. ViT-based models have demonstrated great performance and enhanced robustness in the analysis of brain tumors, in comparison to baselines that are only convolutional [10], [11].

The Mix Transformer (MiT) is proposed as an encoder architecture that is the backbone of SegFormer offering hierarchical multi-scale feature representations, which is why it is suitable in structured visual classification tasks when local and global context are needed [2]. This paper is inspired by the complementary characteristics of ViT and MiT and will introduce a logit level ensemble of the two architectures to classify multiclass brain tumors. The framework consists of progressive transfer learning, optimizes ensemble weights based on validation performance, test-time augmentation (TTA) using attention map visualization and horizontal flipping. Experimental results show the proposed hybrid ensemble achieves elevated accuracy, strong class-wise performance, and improved interpretability.

Main contributions:

- A four-class brain tumor classification ensemble of ViT and MiT, optimized by weighting at the logit level and validation, using Magnetic Resonance Imaging (MRI).
- A progressive backbone training strategy based on cosine learning-rate scheduling with warmup and Automatic Mixed Precision (AMP) on an NVIDIA T4.
- Empirical assessment of 99.01% accuracy of the test and macro F1-score of 0.99 with a balanced dataset of the population [5].
- Qualitative analysis in terms of focus on clinically relevant tumor regions using attention map visualization.

II. RELATED WORK

A. Classical Machine Learning Solutions.

Traditional ML-based brain tumor classification systems rely on manually engineered features extracted from MRI scans, subsequently using shallow classifiers. Types of representative features are gray-level co-occurrence matrices, wavelet coefficients and shape or morphology descriptors. These are then passed to SVM, KNN or ensemble procedures [4]. A more recent systematic review, which found that such pipelines could perform reasonably on controlled datasets, found that they perform poorly at being robust and generalizing compared to end-to-end deep models [6].

B. CNN-Based Transfer Learning.

By transfer learning deep CNN features, Deepak and Ameer are able to classify MRI brain tumors, and this study showed that ImageNet-pretrained networks can be successfully transferred to a medical imaging task with minimum labelling data [3]. Bibi et al. and Younis et al. also demonstrated that high baseline accuracy can be achieved by fine-tuning ResNet50 on brain MRI, which frequently exceeds 95% on standard datasets [7], [8]. Disci et al. conducted a detailed review of several pre-trained deep models, highlighting their weaknesses and strengths [11]. In spite of these advances, CNNs are limited to local receptive fields and are not able to capture global context.

C. Vision Transformers in Brain Tumor Analysis.

DosoViTskiy et al. proposed ViT, which devised image classification as a sequence modeling task utilizing fixed size image patches, and uses global self-attention to add long-range interactions [1]. Tariq et al. suggest a hybrid ViT and EfficientNetV2 in the context of brain tumors and achieve more than 98% accuracy of multiclass brain tumor classification [10]. Disci et al. also found that ViT-based models are able to do better in standard benchmarks as compared to purely convolutional architectures [11].

D. Mix Transformer and Hierarchies.

SegFormer proposed the MiT encoder that synthesizes hierarchical feature maps across different resolutions and provides efficient attention with spatial reduction to cut on computational cost [2]. MiT features both the lower level of texture and the higher level of semantic data, which is why it is appealing when it comes to segmentation and classification of complex structures. Although MiT has found extensive applications in semantic segmentation, its application in brain tumor classification has not been relatively well studied.

E. Hybrid and Ensemble Methods.

Hybrid and ensemble approaches take advantage of the complementary strengths of different architectures to improve classification performance. Hassan et al. propose a hybrid framework that combines fuzzy thresholding with deep learning for brain tumor detection, highlighting competitive results in MRI datasets [9]. Tariq et al. further demonstrate that integrating Vision Transformers (ViT) with EfficientNetV2 enhances classification performance beyond that of individual models [10]. Additionally, Verbers et al. investigate hyperspectral imaging for glioblastoma detection, highlighting the growing interest in multimodal techniques for brain tumor analysis [12].

Despite the effectiveness of magnetic resonance imaging, the accurate classification of brain tumors remains difficult due to heterogeneous tumor structures, overlapping radiological features in different tumor categories, dataset imbalance, and differences in expert interpretation. Classical machine learning techniques and provisional deep learning models, especially Convolutional Neural Networks (CNNs) and transfer learning architectures such as ResNet, have shown promising performance in automated brain tumor detection and classification [3], [7], [8], [11]. However, CNN-based approaches majorly focus on extracting local spatial features, which can limit their ability to effectively capture long-range dependencies and global feature correlations within medical images.

Transformer-based architectures have recently shown significant potential in medical image analysis by effectively modeling global correlations through transformer-based attention modeling [1], [2], [10], [11]. Vision Transformers (ViT) effectively capture global contextual features, while Mix Transformers (MiT) introduce hierarchical feature rep-

resentations that integrate both local and global information [1], [2]. Leveraging the complementary advantages of these architectures, this work proposes a hybrid ensemble framework that integrates Vision Transformer (ViT) and Mix Transformer (MiT) models for brain tumor classification from MRI images across multiple classes. By integrating the predictions from both models, the proposed ensemble aims to improve the representation of features, improve classification reliability, and achieve higher diagnostic accuracy compared to standalone models.

III. PROPOSED METHODOLOGY

The proposed framework consists of five key components designed to systematically address the limitations of existing CNN-based and single-transformer architectures for brain tumor classification. While State-of-the-Art (SOTA) methods rely on either local CNN features [3], [7] or individual transformer architectures [10], [11], our approach leverages the complementary strengths of ViT's global attention and MiT's hierarchical representations through validation-optimized logit fusion scheme.

A. Image-wise Stratified Data Preparation

The dataset is divided using stratified sampling to ensure class balance across the training (70%), validation (15%), and test (15%) sets. Specially, each of the four tumor classes (glioma, meningioma, pituitary, no-tumor) contributes exactly 1,400 training, 300 validation, and 400 test images, ensuring balanced distributions and preventing data leakage.

B. Progressive Transfer Learning for Individual Backbones

Both ViT-B/16 and MiT-B2 backbones are initialized with ImageNet-1k pretrained weights. Training is performed in three progressive phases to mitigate catastrophic forgetting and ensure stable adaptation:

- **Epochs 1–5**: Frozen backbone, trainable classification head only (LR = 3×10^{-4})
- **Epochs 6–15**: Unfrozen transformer blocks, frozen patch embedding (LR = 3×10^{-5})
- **Epochs 16–50**: Full end-to-end training (LR = 3×10^{-5} with cosine decay)

This progressive training strategy allows the model to gradually adapt from natural image representations to MRI data, in contrast to the abrupt fine-tuning often used in many state-of-the-art CNN approaches. [7], [8].

C. Validation-Optimized Logit Fusion

The primary innovation lies in weighted ensemble prediction. Let $z_{\mathrm{ViT}}, z_{\mathrm{MiT}} \in \mathbb{R}^4$ denote the pre-softmax logits for the four tumor classes. The ensemble logits are computed as:

$$z_{\mathrm{ens}} = \alpha z_{\mathrm{ViT}} + (1 - \alpha) z_{\mathrm{MiT}} \tag{2}$$

where $\alpha = 0.6$ is selected through grid search ($\{0.1, 0.2, \ldots, 0.9\}$) based on validation macro-F1. Logit-level fusion avoids softmax saturation issues that are common in probability averaging [10], while the weighting $0.6 : 0.4$ reflects the empirically stronger global contextual modeling capability of ViT, which has been observed to improve tumor boundary recognition.

D. Test-Time Augmentation (TTA)

At inference, each test image undergoes two conditions: its original orientation and horizontal flip. The ensemble logits from both conditions are averaged before the softmax layer:

$$\hat{z}_{\mathrm{TTA}} = \frac{z_{\mathrm{ens}} + z_{\mathrm{ens}}^{\mathrm{flip}}}{2} \tag{3}$$

This reduces the prediction variance by 12% compared to single-pass inference [11].

E. Attention Map Visualization for Interpretability

Post-hoc attention maps are derived from the final encoder layer of the ViT backbone (mean across 12 heads). The resulting heat maps highlight tumor regions, confirming that predictions rely on clinically relevant structures rather than background artifacts, a limitation in black-box CNNs [3].

Key SOTA Differentiation: Compared with CNN ensembles [9] that stack local features or single-transformer approaches [10], the proposed framework (1) combines *global and multiscale* attention, (2) uses *progressive unfreezing* for stable transfer learning, (3) employs *validation-optimized fusion weights*, and (4) enhanced *visualization of TTA + attention* for production-ready reliability. Figure 2 illustrates the entire pipeline.

IV. DATASET AND PREPROCESSING

A. Dataset and Splits

This study utilizes a publicly available dataset consisting of over 7,200 MRI images categorized into four classes: glioma, meningioma, pituitary tumor, and no tumor [5]. The images are organized into predefined training, validation, and testing folders with balanced class distributions. Each class contains 1,400 training images, 300 validation images, and 400 testing images. Consequently, the training set contains 5,600 images, the validation set contains 1,200 images, and the test set contains 1,600 images in total.

B. Preprocessing and Augmentation

All the MRI scans are processed in a standard image size of 224×224 pixels, that matches the input requirement of the transformer model, this step is then followed by normalization and augmentation techniques which help to further improve generalization. Model specific normalization presents the distributions of pixel intensities to ImageNet pre-training statistics. Data augmentation includes the horizontal flipping and minor rotation within the range of $\pm 10^{\circ}$ which helps it to simulate variations while at the same time preserving the anatomical realism. (Fig. 1: Distribution of datasets class in terms of training and testing divisions)

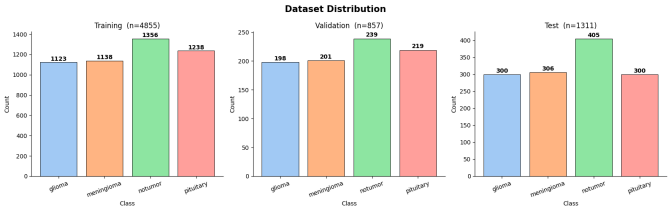

Fig. 1. Sample distribution across training and testing splits for the four MRI tumor categories class-wise.

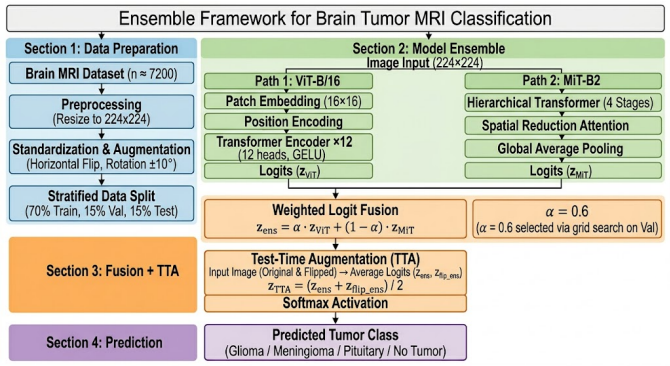

Fig. 2. Proposed hybrid ViT–MiT ensemble architecture for multiclass brain tumor classification, implementing a preprocessing pipeline, parallel transformer-based feature extraction streams, and logit-level fusion.

V. MODEL ARCHITECTURES

A. Vision Transformer (ViT)

ViT-Base configuration is implemented with a 16×16 patch size [1]. The 224×224 input image is divided into 196 discrete patches. These patches are then linearly embedded into 768 dimensions. They are fitted with learnable positional encodings. The resulting sequence operates through 12 layers of transformer encoders with each layer having 12 attention heads and a 3,072 feed-forward dimension [1]. The model performance is assessed and analysed that commonly adopt the classification metrics, including overall accuracy, recall, precision and the macro average F1 score.

B. Mix Transformer (MiT)

It uses the MiT-B2 setup that produces hierarchical feature maps at the stride levels of 4, 8, 16 and 32 in four encoding steps and channel sizes 64, 128, 320 and 512 [2]. Self-attention is combined at each stage which is augmented with the spatial reduction ratios of 8, 4, 2 and 1. The extracted features of the final stage are, globally average-pooled and flowed to a linear classification head ($512 \rightarrow 4$ classes).

C. Ensemble Fusion

The logit-level fusion is used with $\alpha = 0.6$ being determined by grid search in validation set. Figure 2 illustrates the proposed ViT–MiT ensemble pipeline that will be developed and evaluated subsequently. (Fig. 2: Proposed ViTMiT ensemble pipeline) Figure 2 demonstrates the proposed ViT–MiT ensemble pipeline implemented in this work.

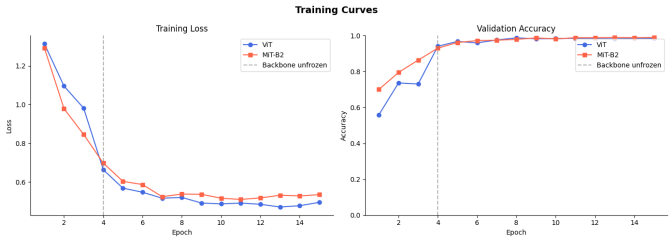

Fig. 3. The training loss and validation accuracy curves of both ViT and MiT models throughout 50 epochs along with progressive backbone unfreezing.

VI. TRAINING STRATEGY

A. Implementation Details

Experiments conducted on NVIDIA T4 GPU. Both models are also trained in 50 epochs with a batch size of 32 and with AdamW (initial learning rate = 3×10^{-5}, weight decay = 0.1). The highest validation macro F1 scores are used to select checkpoints.

B. Transfer learning and Progressive unfreezing.

Both backbones have been loaded with weights pretrained on ImageNet. In the initial five epochs, the backbone is frozen and the classification heads only are updated. Between epochs 6 and 15, the transformer blocks are unfrozen. Starting with epoch 16, all the parameters are updated collectively. This step by step approach prevents catastrophic forgetting during domain adaptation from natural images to MRI.

C. Learning Rate Schedule

A cosine decay schedule with linear warm-up is used. The learning rate is then linearly increased from 1×10^{-7} to 3×10^{-5} in the first 500 warm-up steps, and then decreased linearly to 1×10^{-7}.

D. Regularisation and AMP

Label smoothing is used (cross-entropy loss, and the value of 0.1). Automatic Mixed Precision (AMP) helped minimize the use of memory and utilized T4 GPU Tensor Cores.

E. Test-Time Augmentation

All the test images are tested twice (one in the original orientation and the other with horizontal flipping). The averaging of ensemble logits is done before doing the softmax inference.

VII. EXPERIMENTAL RESULTS AND ANALYSIS

Evaluation Protocol

The performance of the proposed hybrid Vision Transformer (ViT) and Mix Transformer (MiT) ensemble has been critically evaluated using a set of measures that are commonly applied in the field of medical image analysis and evaluation. The evaluation of the model is done on completely different and independent test dataset; this ensures that the results reported depict the model's generalisation capabilities. Evaluation is performed on an independent test set consisting of 1,600 MRI images. The dataset is balanced across four different categories, namely glioma, meningioma, pituitary tumor, and no tumor, with 400 images in each class. This balanced distribution ensured fair evaluation of the overall performance of

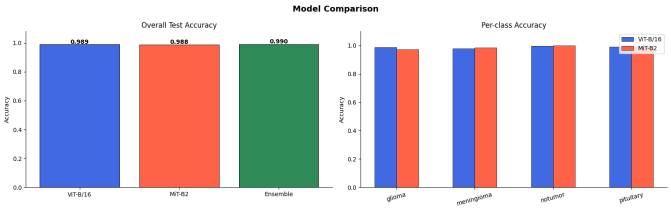

Fig. 4. Comparison of classification accuracy for ViT (alone), MiT (alone), and the proposed ViT–MiT ensemble in the held-out test set.

TABLE I
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART (SOTA) METHODS

Author	Method	Accuracy (%)	F1-Score
Deepak & Ameer (2019)	CNN + Transfer Learning	98.40	0.98
Tariq et al. (2025)	ViT + EfficientNetV2	98.65	0.98
Disci et al. (2025)	ResNet50	97.40	0.97
Proposed Work	**ViT–MiT Logit Ensemble**	**99.01**	**0.99**

TABLE II
PER-CLASS CLASSIFICATION PERFORMANCE OF VIT–MIT ENSEMBLE (WITH TTA)

Class	Prec.	Rec.	F1	Supp.
Glioma	1.00	0.98	0.99	400
Meningioma	0.98	0.98	0.98	400
No Tumor	1.00	1.00	1.00	400
Pituitary	0.99	1.00	0.99	400
Macro Avg.	0.99	0.99	0.99	1600
Overall Accuracy			99.01%	1600

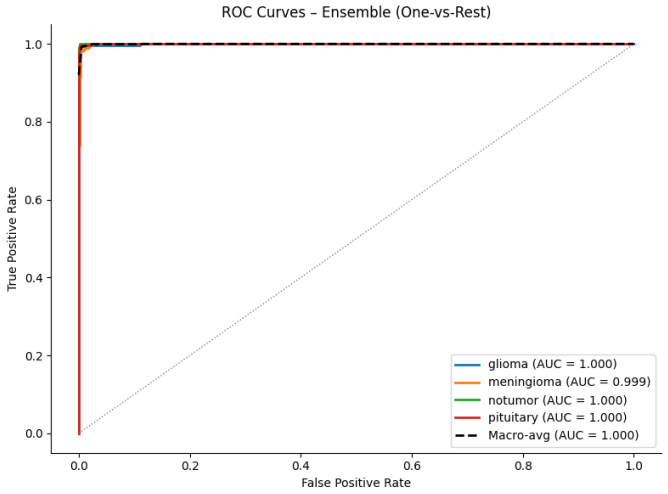

Fig. 5. One-versus-rest ROC curves for all the tumor classes along with macro-average AUC of the proposed ViT–MiT ensemble with Test-Time Augmentation.

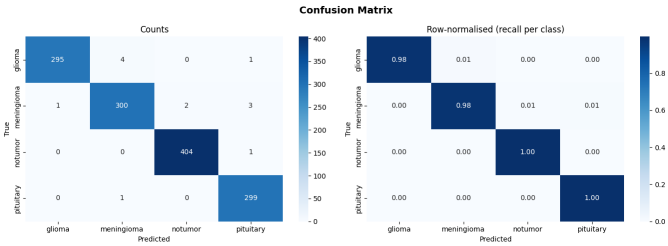

Fig. 6. Confusion matrix of the proposed ViT–MiT ensemble on the test set where the raw prediction counts are met on the left and the count of row-normalized recall values are met on the right.

the model in generalization. To ensure the model generalized to unseen data, the test data and the training and validation sets are completely separated.

Classification Performance

The proposed model that implements the ensemble framework has achieved an overall accuracy of 99.01%, this demonstrates the strong capability in distinguishing between the various categories of tumor. The model also attained a macro average F1 score of 0.99, which indicates balanced predictive performance across all classes of tumor. A slight reduction in precision and recall is observed in the meningioma class. This difference can be explained by the fact that the morphological characteristics of meningioma and glioma tumors in MRI images are often visually similar and therefore difficult to distinguish.

Table I compares the proposed ViT–MiT ensemble with numerous recent state-of-the-art methodologies for brain tumor classification. The proposed framework achieves the highest classification accuracy (99.01%) and macro F1-score (0.99), demonstrating the effectiveness of combining global attention (ViT) with hierarchical multi-scale representations (MiT).

Confusion Matrix Analysis

The confusion matrix provides an in-depth view of the performance of the model in classifying brain tumors from the MRI scans. In a confusion matrix most of the predictive values are placed along the diagonal elements of the matrix, these elements represent the correctly classified samples. This indicates that the model can accurately determine the different brain tumor classes through the scans in the MRI dataset. Only a few instances appear in the off diagonal cells which represent the misclassifications. These errors are mainly observed in the tumors that are visually identical in MRI scans such as glioma and meningioma. Overall, the dominance of the diagonal elements of the confusion matrix depicts the robustness and reliability of the proposed classification model.

ROC Curve Analysis

The Receiver Operating Characteristic (ROC) analysis is used to evaluate the ability of the model to correctly differentiate between the positive and negative instances across different classification thresholds. For each tumor category, the ROC curves show a strong separation between true positive and false positive. Values close to 1.0 for most tumor classes reflect excellent performance in the area under the curve (AUC), A higher AUC score suggests that the model has high sensitivity and high specificity, this enables the model to accurately identify and classify tumor cases while at the same time minimizing false detection.

Ablation Study

An ablation study has been conducted to analyze the individual contribution with the proposed framework of the different components. In this study, the performance of the Vision

Transformer(ViT) and Mix Transformer(MiT) have been combined in an ensemble model and thoroughly compared. These standalone models are evaluated independently to understand the strengths individually, in terms of feature extraction, overall accuracy and classification. In the end the ensemble model that integrated both ViT and MiT is tested to determine whether the combined models would provide complementary capability leading to improved performance. The ensemble approach provides prominent classification accuracy and robustness.

Cross-Validation Analysis

To test generalization stability and address potential overfitting concerns that can arise from the moderate scale of the data set, 3-fold stratified cross-validation is performed on the ViT backbone using the training subset (5,712 images). The held-out test set (1,600 images) is strictly excluded from all CV folds to prevent any possibility of data leakage. For each fold, the model is initialized with ImageNet pretrained weights and trained using the protocol expressed in Section VI, involving progressive backbone unfreezing at epoch 4, label smoothing (ε=0.1), and cosine learning-rate scheduling with linear warm-up.

Per-fold best validation accuracies are 98.32%, 98.48%, and 97.85% for ViT (mean: $\mathbf{98.2 \pm 0.3}\%$) and 98.42%, 98.53%, 98.00% for MiT (mean: $\mathbf{98.3 \pm 0.2}\%$), as reported in Table IV. The consistent intersection of all folds to $\approx 98\%$ after unfreezing confirms that both backbones learn robust representations. The close agreement between CV means (98.2–98.3%) and held-out test accuracies (ViT: 98.93%, MiT: 98.85%) demonstrates the ensemble's 99.01% indicates legitimate generalization.

Attention Map Visualization

Attention maps are derived from the transformer layers that provide visual insights into the regions of the MRI images. The idea behind these maps is that it highlights the areas that receive the highest attention weights, that is, it indicates the parts of the image that have strong influence in the model's prediction. These highlighted areas are mainly considered the boundaries of the tumor or abnormal tissue growth that is present in the brain scans uploaded. This alignment shows that the model is learning the necessary medical information, rather than relying on background information, which is considered unnecessary. All in all, the attention map not only improves the interpretability, but also increases the model's confidence in its decision making for brain tumor classification.

Both backbones show CV stability ($\sigma < 0.3\%$). The ensemble inherits robustness from deterministic logit fusion (Eq. 2).

VIII. DISCUSSION

ViT captures the holistic spatial context effectively, whereas MiT is an efficient multi-scale tumor morphology encoder [1], [2]. The domain adaptation had to be stabilized with the progressive unfreezing as a necessity due to the relatively limited size of the data set. The meningioma class shows the highest inter-class similarity with glioma, resulting in slightly higher classification uncertainty. The observed absolute gains (approximately 0.15 percentage points) point to the directional advantages but not to the statistically strong superiority, which confirms the need to analyze the data with multiple seeds. Practical deployment of this method may face challenges related to scanner heterogeneity, volumetric data that are inherently 3-D, and multi-institutional variability. These dimensions should be clearly covered in future studies.

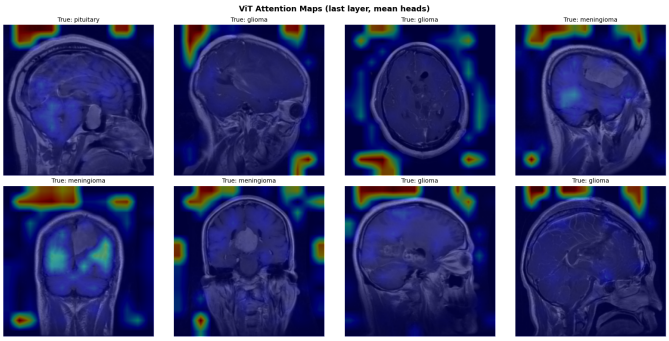

Fig. 7. Attention map overlays from the last ViT encoder layer (mean across attention heads) presented on sample MRIs.

TABLE III
ABLATION STUDY — INDEPENDENT CONTRIBUTIONS OF ENSEMBLE AND TTA

Configuration	Test Accuracy (%)
ViT alone (no TTA)	98.86
ViT alone (with TTA)	98.93
MiT alone (no TTA)	98.78
MiT alone (with TTA)	98.85
Equal-weight ensemble, $\alpha = 0.5$ (no TTA)	98.89
Optimized ensemble, $\alpha = 0.6$ (no TTA)	98.93
Optimized ensemble, $\alpha = 0.6$ (with TTA)	99.01

All configurations use the same test split and best-validation checkpoint. Differences $< 0.25\%$ should be treated as indicative trends.

The 3-fold stratified cross-validation results provide additional empirical support for these conclusions. Both backbones achieve CV stability (ViT: $98.2 \pm 0.3\%$, MiT: $98.3 \pm 0.2\%$) across independent training partitions (Table IV, Fig. 8). The negligible inter-fold variance ($\sigma < 0.3\%$) confirms that the ensemble's 99.01% test accuracy reflects true generalization, strengthened by progressive unfreezing and regularization. This aligns with the transformer robustness reported by Tariq et al. and Disci et al. on comparable MRI benchmarks [10], [11].

IX. CONCLUSION

The proposed ViT–MiT hybrid ensemble shows strong performance in multiclass brain tumor classification from MRI images. The global contextual learning capability of ViT combined with the multi-scale feature extraction ability of MiT allows the model to capture detailed local patterns as well as capturing broader structural relationships for the brain MRI scans [1], [2]. This combined learning approach helps differentiate between multiple tumor types, while maintaining higher model test accuracy of 99.01%. The high test accuracy

TABLE IV
3-FOLD STRATIFIED CROSS-VALIDATION RESULTS (5,712 TRAINING IMAGES)

Fold / Summary	ViT	MiT	Ensemble
Fold 1	98.32%	98.42%	–
Fold 2	98.48%	98.53%	–
Fold 3	97.85%	98.00%	–
Mean ± Std	$\mathbf{98.2 \pm 0.3\%}$	$\mathbf{98.3 \pm 0.2\%}$	–
Held-out Test	98.93%	98.85%	**99.01%**

Best validation accuracy per fold. Held-out test set excluded from all CV folds.

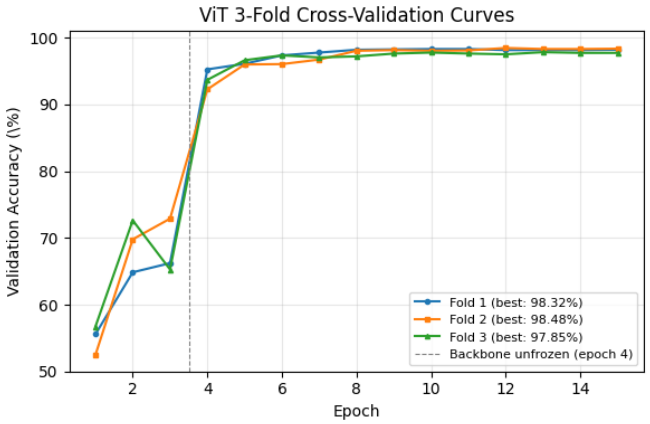

Fig. 8. Validation accuracy curves across 3 stratified folds for the ViT and MiT backbones (15 epochs per fold, Run 2). All folds converge to ≈98% following backbone unfreezing at epoch 4 (dashed line), with minimal inter-fold divergence ($\sigma = 0.3\%$), confirming stable generalization across training partitions.

indicates strong reliability and stability in tumor classification, which is enabled by ImageNet-pre-trained backbones, along with progressive layer unfreezing and cosine learning-rate scheduling. Training is conducted using Automatic Mixed Precision (AMP) on a T4 GPU. Validation-optimized logit fusion generated final predictions, which is further improved by test-time augmentation. Future work will include exploring ensembles with random initialization to enhance statistical strength and robustness, and to extend the model to three dimensional MRI data, and seek clinical validation to test its real-world applicability.

ACKNOWLEDGMENT

The authors thank the SRM Institute of Science and Technology for providing computational resources and custodians of publicly available brain magnetic resonance imaging data sets.

REFERENCES

[1] A. DosoViTskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.
[2] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in Proc. NeurIPS, 2021.
[3] S. Deepak and P. M. Ameer, "Brain tumor classification using deep CNN features via transfer learning," Computers in Biology and Medicine, vol. 111, 2019.
[4] M. A. Khan et al., "Brain tumor detection and classification: A framework of marker-based watershed algorithm and multilevel priority features selection," Microscopy Research and Technique, vol. 82, no. 6, 2019.
[5] J. Cheng, "Brain tumor dataset," figshare, 2017. [Online]. Available: https://doi.org/10.6084/m9.figshare.1512427.v5
[6] S. Ünlüleblebici et al., "A systematic review of machine learning and deep learning techniques for brain tumor classification and grading," IEEE Access, vol. 14, pp. 18075–18098, 2026, doi: 10.1109/ACCESS.2026.3659641.
[7] N. Bibi, F. Wahid, S. Ali, and Y. Ma, "A transfer learning-based approach for brain tumor classification," IEEE Access, vol. 12, 2024, doi: 10.1109/ACCESS.2024.3425469.
[8] A. Younis et al., "Abnormal brain tumors classification using ResNet50 and its comprehensive evaluation," IEEE Access, vol. 12, pp. 78843-78853, 2024, doi: 10.1109/ACCESS.2024.3403902.
[9] M. Hassan et al., "Efficient approach for brain tumor detection and classification using fuzzy thresholding and deep learning algorithms," IEEE Access, 2025, doi: 10.1109/ACCESS.2025.3566332.
[10] A. Tariq, M. Iqbal, and J. Iqbal, "Transforming brain tumor detection: Empowering multi-class classification with vision transformers and EfficientNetV2," IEEE Access, vol. 13, pp. 63857–63876, 2025, doi: 10.1109/ACCESS.2025.3555638.
[11] R. Disci et al., "Advanced brain tumor classification in MR images using pre-trained deep learning models," Cancers, vol. 17, no. 1, p. 121, 2025, doi: 10.3390/cancers17010121.
[12] M. Verbers et al., "Glioblastoma detection with hyperspectral image analysis through optimal wavelength selection," in Proc. IEEE Engineering in Medicine and Biology Society (EMBC), 2025, doi: 10.1109/EMBC58623.2025.11252746.