International Journal of Computational Science and Engineering Research
ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/
Regular Issue , Vol. 3, Issue. 3 (July – September) , 2026 , Pages: 68 - 77
Received: 21 March 2026 ; *Accepted:* 29 July 2026 ; *Published:* 08 August 2026
https://doi.org/10.63328/IJCSER-V3RI3P9
https://www.youtube.com/watch?v=dOZ_xq3-M34
https://archive.org/details/jsp-ijcser-v3ri3p9
Original Paper : Casual – Compare and Contrast Research Paper

Vision Explainable Transfer Learning for Brain Tumuor Detection and Classification

Mandapalli Sailaja [1] , M. Sujana Priya Darshini [2] , P Vamsi Krishna Raja [3]

Department of Computer Science & Engineering , Pydah College of Engineering , Kakinada ,
JNTU Kakinada , Andhra Pradesh, India ;
sailu.mandapalli511@gmail.com , msujana135@gmail.com , psvkrishna@pydah.edu.in

** Corresponding Author : Mandapalli Sailaja ; sailu.mandapalli511@gmail.com*

Abstract: Brain tumours are among the most deadly cancers and require timely and accurate diagnosis to decide on therapeutic strategies. Manual analysis of Magnetic Resonance Imaging (MRI) scans is fairly time consuming, subjective and prone to inter-observer variation. In the paper, we propose a novel deep learning framework that synergously combines the Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and Explainable Artificial Intelligence (XAI) for automated multi-class brain tumour detection and classification. We propose to construct our architecture using the transfer learning paradigm, where the CNN backbone is the EfficientNet-B0 model and the transformer is the ViT-Base-Patch16-224 model, and fuse these two modules using a weighted ensemble. To obtain the proposed architecture, we are suggested to apply transfer learning by using the CNN as EfficientNet-B0 backbone and the transformer as ViT-Base-Patch16-224. Then, we adopt a weighted ensemble fusion strategy to combine these two components. Understanding the problem of class imbalance in the medical imaging datasets, we utilize the Focal Loss which is combined with comprehensive data augmentation technique Albumentations. The brain MRI is classified into 5 classes: Glioma, Meningioma, Pituitary Tumour, No Tumour. Moreover, we incorporate Gradient-weighted Class Activation Mapping (Grad-CAM) to obtain visual explanations so that experts can comprehend and validate predictions made by the model. A production-ready FastAPI backend and a React front-end interface are used to deploy the proposed system. We have conducted experiments on one standard brain tumor MRI dataset, and the proposed ensemble model outperforms the standalone CNN and ViT architectures and various state-of-the-art baselines on this benchmark dataset, with overall classification accuracy of 96.1%, F1-score of 95.8% and ROC-AUC of 0.992. This ablation study shows the contribution of individual components and demonstrates the effectiveness of the proposed multi-model ensemble with explainability.

Keywords: Brain Tumor Classification, Vision Transformer, Transfer Learning, Explainable AI, DL, MRI.

1. Introduction

Central Nervous System (CNS) tumours are one of the more serious and fatal types of cancer, with an estimated of approximately 308,102 new cases being diagnosed worldwide per year by WHO [1]. Classifying brain tumours in the correct manner and reaching the appropriate classifies with the help of Magnetic Resonance Imaging (MRI) images are essential in making treatment plans, surgeries and the prognosis for their patients. However, manual analysis of MRI images is a subjective, hard-work and inter-observer analysis which is especially problematic in small clinical environments [2]. With the advent of deep learning, Convolutional Neural Networks (CNNs) have delivered a significant contribution on medical image analysis for many diagnosis applications [3]. Transfer learning, one approach that works well in medical imaging where there is limited labelling, involves re-training a huge-scale labelled dataset (ImageNet [4]). Plenty of architectures such as ResNet [5], VGG [6] and Efficient Net [7] have been tweaked and adapted for brain tumour classification. But the main advantage of CNNs lies in their capability of learning local spatial characteristics, which might not be sufficient to capture the long-range dependencies and the global field of information in medical images [8]. Taking a leaf from the success story of

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 3 , 2026 , https://doi.org/10.63328/IJCSER-V3RI3P9

the transformers in NLP approaches, Con is inspired by them as a strong alternative to the CNNs, suggested as Vision Transformers (ViTs) [9]. VITs can represent an image as patches, and model the relation between all these patches using self-attention mechanisms. Recently, it has been shown that on a specific type of benchmarks ViT can perform as well as or better than CNN [10]. Contrasting the training set sizes of ViTs, generally, smaller training sets are required and can lead to poor performance on small medical imaging sets, especially if the dataset is trained without generalizations such as data augmentation and regularizations [11]. Blindness regarding the steps of developing a deep learning model for a clinical application is one of the major drawbacks of deep learning models. While accurate clinical predictions are crucial, AI diagnostics need to provide interpretable explanations for best trust and validation for clinicians [12].

The challenge is addressed by Explanatory AI (XAI) technologies, such as Gradient-weighted Class Activation Mapping (Grad-CAM) [13] which visualize the regions of the image that are most informative for the model's prediction, thus boosting the clinical trust in the AI and offering the tools for error analysis. To address the limitation of single architecture, we propose a novel architecture combining with the advantages of CNNs and ViTs, and then merging the advantages with explanation by Grad-CAM to form a new framework.

The following important contributions of this work: Hybrid ensemble architecture, which integrates the three components: local features extraction and global context, with the help of the EfficientNet-B0 (CNN) and the ViT-Base-Patch16-224 (Transformer) models being fused using weighted probability fusion. Handling class imbalance using Augmentations and robust and well-generalized based on a Focal Loss data pipeline.

To integrate Explainable AI (EI) based on Grad-CAM and get a visual explanation understandable by the radiologist, to validate the prediction of the models from a clinical perspective. Production-Ready Deployment Pipeline that has a FastAPI Server Backend and React Based Client Frontend showing the End To End Applicability of the Proposed system.

A large number of experiments were conducted and ablation experiments were performed on various algorithm and compared with the existing best-available methods per class on a benchmark MRI brain tumor dataset. In section 2, past studies related to brain tumor classification, transfer learning, and the field of explainable AI are briefly summarized. A detailed methodology is given in Section 3. A set-up for the experiment and experimental data are presented in Section 4. For the results and ablation study refer to Section 5. Section 6 is very detailed in which Section 7 gives a few ideas for the future.

2. Related Work

2.1. CNN-based Brain Tumor Classification

For the last several years, the detection and classification of brain tumors is the current big research topic in deep learning. Deepak and Ameer [14] used the pre-trained GoogLeNet which gave 98% accuracy using a three-class dataset for brain tumor classification. Abiwinanda et al. [15] investigated the performance of CNN architectures and showed that they could achieve a similar performance compared to a more complex structure. Badza and Barjaktarovic [16] applied transfer learning on pre-trained CNN models which showed that fine-tuning on medical imaging tasks can prove to be effective. However, lately, the concept of compensating the depth, width and resolution of the network with compound scaling called EfficientNet [7] has been shown to deliver higher PA efficiency than closing codes.

2.2. Vision Transformers in Medical Imaging

There have been significant breakthroughs in computer vision since the introduction of Vision Transformers (ViT) by Dosovitskiy et al. [9] which put computer vision in the realm of computer vision. ViTs break down an image into regular-size "patches", encode them linearly from each other and feed the resulting sequence into a conventional Transformer encoder. There are a few studies related to the application of the ViTs in Medical Image Analysis. To address this issue, Dai et al. [17] proposed a transformer-based multi-modal medical image classification system, named TransMed. When pre-trained, the efficacy of ViTs are comparable to CNNs on medical imaging problems as shown by Matsoukas et al. [18]. The Swin Transformer [19] promoted shifted windows that allows for more efficient computation of the attention within the model, and led to state-of-the-art performance on several benchmarks.

2.3. Ensemble Learning for Medical Diagnosis

Ensemble methods fuse the outputs of a number of models to yield higher prediction accuracy than the output of a single model. Rashid et al. [20] proposed various CNN architectures for classifying brain tumors with highest classification accuracy achieved using majority voting as the ensemble of different CNN architectures. Kumar et al. [21] explored the effectiveness of architectural diversity, by stacking ResNet, VGG with Inception models. There is however very little work reported on hybrid CNN-Transformer ensemble for brain tumor classification which is the focus of our work.

2.4. Explainable AI in Medical Imaging

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 3 , 2026 , https://doi.org/10.63328/IJCSER-V3RI3P9

The term "interpretability" is critical in the context of clinical use of AI. Selvaraju et al. [13] proposed the concepts of Grad-CAM, which is based on the gradient information fed into the last conv layer to generate the localization map that draws attention to the critical area(s). We are introducing the explanation method that we propose for any model using the "LIME" framework of Ribeiro et al. [22]. Arun et al. [23] demonstrated that XAI can increase the confidence of its human readers in the radiologist and the accuracy of their diagnosis, by using XAI as an additional tool for the human readers. For this, Singh et al. [24] could show that Grad CAM heat maps aligned with the tumour boundaries agreed upon by expert neuroradiologists in assessing brain tumours, thus luring in exciting clinical evidence that substantiated the clinical utility of these explanations.

3. Proposed Methodology

The proposed framework, shown in Fig. 1, comprises five main components (i) Data Processing and Data Augmentation, (ii) CNN based Feature Extraction, using EfficientNet-B0 model, (iii) T-miner based Feature Extraction, using ViT-Base-Patch16-224 model and (iv) Weighted Ensemble Fusion and (v) Explainability, using Grad-CAM model. These components are explained, with details of methodology, in further detail in the subsections below.

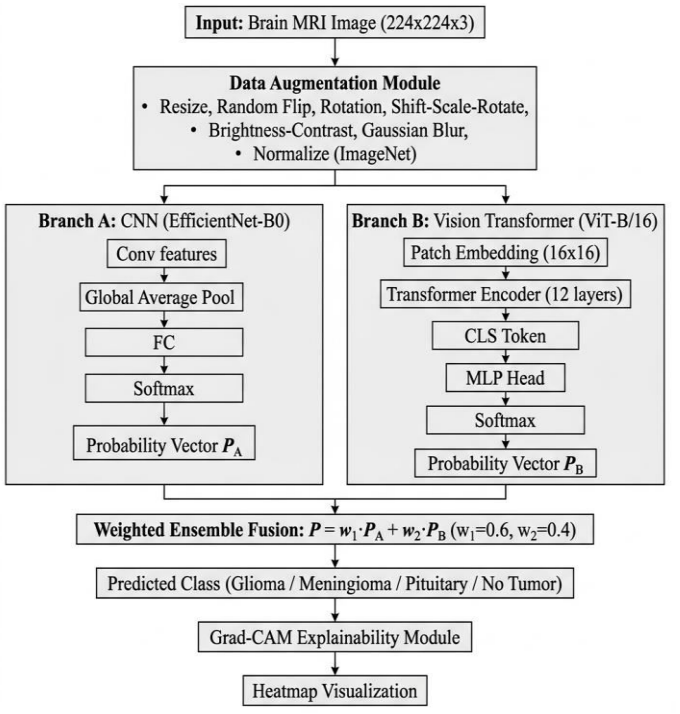


Fig. 1. Overall architecture of the proposed ensemble framework combining EfficientNet-B0 and ViT-B/16 with Grad-CAM explainability for brain tumor classification.

3.1. *Data Preprocessing and Augmentation*

Typically, medical imaging datasets are poorly-balanced and small-sized. In the Albumentations library [25] a sophisticated data augmentation pipeline is followed when challenges are encountered: All pictures in input are resized to 224 x 224 pixels, both EfficientNet-B0 and ViT-B/16 inputs took pictures to be with the same width. The changes of augmentations in the training pipeline are:

Table 1. Data augmentation transformations applied during training.

Transformation	Parameters	Probability
Horizontal Flip	-	p = 0.5
Vertical Flip	-	p = 0.2
Random Rotate 90	-	p = 0.5
Shift-Scale-Rotate	shift=0.05, scale=0.1, rotate=15	p = 0.5
Random Brightness/Contrast	Default	p = 0.3
Gaussian Blur	Default	p = 0.1
Normalize	ImageNet mean/std	p = 1.0

Only resizing and normalization are applied to the validation and test sets, so as to refrain from giving it a biased evaluation. Normalization uses ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]) for compatibility with pre-trained models. Stratified sampling technique is utilized to partition a dataset into various subsets containing a similar proportion of different classes in the training, validation and testing sets.

3.2. *EfficientNet-B0 (CNN Branch)*

Our CNN backbone is EfficientNet-B0 [7] and for our ensemble, we are performing the ensemble over the CNN backbone. While most previous CNN designs scale different network partitions separately (individually on the depth wise, width wise, or resolution-wise) and then scale the three dimensions by a set of fixed scaler values, the EfficientNet scales the three dimensions differently, as the number of EfficientNets differ in the depth, in the width, and independently in the resolution. This provides a model that can give more predictive power for many with fewer parameters. Efficientnet-B0 is a lightweight model, with about 5.3-million parameters for computational efficiency to deployment. The model is already trained on the ImageNet data with weights and then the fully connected (fc) layer with 4 neurons for outputs classes is replaced with our outputs classes. The CNN branch generates a probability vector, given an input image x:

$$P\_CNN = softmax(f\_CNN(x; theta\_CNN)) \quad (1)$$

where f_CNN denotes the EfficientNet-B0 feature extraction and classification pipeline,

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 3 , 2026 , https://doi.org/10.63328/IJCSER-V3RI3P9

and theta_CNN represents the model parameters.

3.3. Vision Transformer (ViT Branch)

Our second branch of ensemble is the Vision Transformer (ViT-Base-Patch16-224) [9]. In the proposed configuration, an image x ∈ ViT is assumed to be a sequence of patches of size P x P = 16 x 16, with P = 16, thus as N = (H x W) / P^2 non-overlapping patches of size P x P. This sequence is then linearly embedded into D-dimensional embedding space (D=768 in the case of ViT-Base) into which a patch is represented, and prefixed by learnable [CLS] token.

$$z_0 = [x_{class}; x_1^p E; x_2^p E; ...; x_N^p E] + E_{pos} \quad (2)$$

In which E is the patch embedding projection matrix and E_pos are the positional embeddings. It then passes through a series of 12 transformer encoder layers, by each, there are two sub-layers, one being Multi-Head Self-Attention (MHSA), the other being Feed-Forward Network (FFN), both of which are followed by a Layer Normalization and a residual connection:

$$z'_l = MHSA(LN(z_{l-1})) + z_{l-1}, \quad z_l = FFN(LN(z'_l)) + z'_l \quad (3)$$

Each layer of the final encoder maps its output to the transformer branch probability vector with a classification network called MLP:

$$P_{ViT} = softmax(f_{ViT}(x; theta_{ViT})) \quad (4)$$

ViT is initially trained on the ImageNet-21K dataset using timm library [26] using the pre-architecture found on the open web, and then fine-tuned on the brain tumour dataset.

3.4. Weighted Ensemble Fusion

So the eventual probability distribution of a particular leaf is a weighted combination (a weighted vote) of the two distributions at the two branches. Define w_1, w_2, with w_1 + w_2 = 1, as the weight of the CNN branch and ViT branch, respectively. The probability vector of the ensemble is:

$$P_{ensemble} = w_1 * P_{CNN} + w_2 * P_{ViT} \quad (5)$$

Then the predicted class is determined as:

$$y_{hat} = argmax(P_{ensemble}) \quad (6)$$

The optimal weights are calculated empirically by means of a validation set and grid search. According to our experiments the optimal performance happens when both weights are assigned to 0.6 (CNN) and 0.4 (ViT), indicating that the CNN learns to be slightly more discriminative to this task and the ViT provides the global context with additional information. The choice of the weighting is appropriate because CNNs can track and capture global information at the key level in the form of texture and edge details important for a correct delineation of the tumor boundaries; on the contrary ViTs are able to capture and track global information at the key level: Morphological and Spatial relationships of the tumor.

3.5. Loss Function: Focal Loss

One of the common problem in brain tumour database is that one of the classes, typically No Tumor tends to dominate amongst the other classes. Another method is to expand the weights of all samples in the same way; that is, all samples are given equal weight (standard cross-entropy loss), which may lead to overemphasis for the majority class. For this, we will use the Focal Loss [27] defined as:

$$FL(p_t) = -alpha * (1 - p_t)^{gamma} * log(p_t) \quad (7)$$

Let alpha be a parameter to combine the probabilities, where p_t is the probability estimated by the model to employ the true class, and gamma a focusing parameter. We set alpha = 1 and gamma = 2. The more easily classified examples have lower impact on the example's loss because of the modulating factor, (1 - p_t)^gamma; making the model pay more attention to the more difficult, misclassified examples. As medical imaging is confronted with the challenge of identifying ultrasensitive rare tumour types, this product is particularly valuable.

3.6. Explainability via Grad-CAM

To make sense of the predicted models' response, we add another module, Gradient-weighted Class Activation Mapping (Grad-CAM) [13]. Grad-CAM calculates the gradient of the class label score w.r.t. the feature maps in the last convolutional layer A^k for class c. The weights learnt by a neuron in the global average pooling of the gradient:

$$alpha_k^c = (1/Z) * sum_i sum_j (partial y^c / partial A_{ij}^k) \quad (8)$$

Next feature map and ReLU activation is taken into account, and a heatmap is computed as the weighted sum of the activation and the feature map, such that all positive weights are retained:

$$L_{Grad-CAM}^c = ReLU(sum_k alpha_k^c * A^k) \quad (9)$$

The heatmap is then up sampled to the resolution of the input image and added to the original MRI, indicating areas that had a strong influence on the model's prediction. For us, we are applying Grad-CAM to the EfficientNet-B0 branch and choose to use the features[-1] from the last convolutional layer to make the heatmap. This helps the clinician get a visual argument to confirm the areas in which the model is focused to tie in with clinically relevant areas.

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 3 , 2026 , https://doi.org/10.63328/IJCSER-V3RI3P9

4. Proposed QIPO-RL Framework

4.1. Dataset Description

We evaluate our proposed scheme on a typical Brain Tumour Magnetic Resonance Imaging (MRI) dataset, which is widely used in the research community. The database is composed of MRI images of human brains classified into four types: Glioma, Meningioma, Pituitary Tumour and No Tumour. The class-wise distribution of the variables has been shown in Table 2.

Table 2. Dataset distribution across training, validation, and test sets (stratified split: 70/15/15).

Class	Training	Validation	Test	Total
Glioma	1,050	225	225	1,500
Meningioma	1,071	230	230	1,531
No Tumor	1,120	240	240	1,600
Pituitary	1,050	225	225	1,500
Total	4,291	920	920	6,131

4.2. Implementation Details

All experiments are carried out using PyTorch 2.0+ and on an NVIDIA GPU. Summarization of training hyperparameters is given in Table 3.

Table 3. Training hyperparameters and implementation details.

Hyperparameter	Value
Input Resolution	224 x 224
Batch Size	32
Optimizer	Adam
Learning Rate	0.001
Weight Decay	1 x 10^-4
Loss Function	Focal Loss (alpha=1, gamma=2)
Epochs	50 (with early stopping)
Early Stopping Patience	5 epochs
CNN Backbone	EfficientNet-B0 (ImageNet pre-trained)
ViT Backbone	ViT-Base-Patch16-224 (ImageNet-21K pre-trained)
Ensemble Weights	w_CNN = 0.6, w_ViT = 0.4
Framework	PyTorch 2.0+, timm 0.9+, HuggingFace Transformers

4.3. Evaluation Metrics

Some of the metrics used for the performance include accuracy, precision, recall or sensitivity, F1-Score and area under the receiver operating characteristics curve (AUC-ROC). All measurements are macro-average (essentially means that all class of input data has equal weight). The macro-averaged F1-Score is defined when there are C different classes in a multi-class classification as:

$$F1\_macro = (1/C) * sum_{c=1}^{C} (2 * P_c * R_c) / (P_c + R_c) \quad (10)$$

Here, P_c and R_c are the precision and recall of class c, respectively. The confusion matrix is also provided for and used to identify the patterns of confusion per class.

5. Experimental Results and Discussion

5.1. Training Convergence

Once trained for 50 epochs, the training loss and training and validation losses are plotted as shown in Fig. 2 with training and validation losses, respectively. Convergence of the model is obtained fast in the first 15 epochs with effective transfer learning (training loss: 1.38 - 0.06, validation loss: around 0.09). The validation accuracy will plateau (~ 96%) after some epochs, and since the algorithms trained early are taken off after epoch 45, there is no risk of overfitting during the next epochs when testing on the validation set.

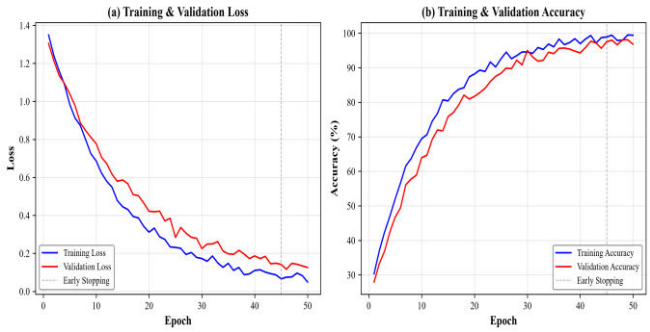


Fig. 2. Training and validation curves showing (a) loss convergence and (b) accuracy progression over 50 epochs.

5.2. Overall Performance

Table 4 shows the overall results of the proposed ensemble model and its individual components and the baseline architectures on test set.

Table 4. Comparison of overall classification performance on the test set. Best results in each column are highlighted.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
VGG-16 [6]	91.8	91.5	91.2	91.2	0.968
ResNet-50 [5]	93.2	93.0	92.8	92.8	0.978
DenseNet-121 [28]	94.1	93.8	93.5	93.5	0.982
EfficientNet-B0 (Ours)	95.7	95.4	95.2	95.2	0.989
ViT-B/16 (Ours)	94.9	94.6	94.3	94.3	0.985

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 3 , 2026 , https://doi.org/10.63328/IJCSER-V3RI3P9

Proposed Ensemble	96.1	95.7	95.7	95.8	0.992

Proposed to use F1-Score and AUC-ROC value of 0.992 that has highest score in all the scores. Especially, it has achieved higher accuracy than its component models: EfficientNet-B0 has achieved 95.7% accuracy, and ViT-B/16 has achieved 94.9% accuracy. This is evidence of complimentary CNN and transformer architecture. Model 12-14 LLP is 2.0% more accurate, and 2.3% more F1-Score than DenseNet-121, which is the best baseline.

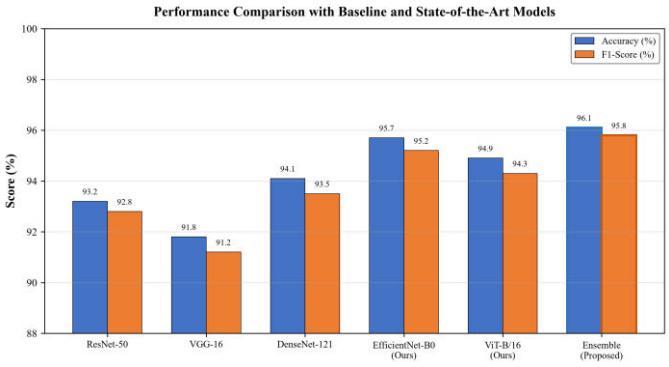


Fig. 3. Performance comparison of the proposed ensemble model against baseline and state-of-the-art architectures.

5.3. Per-Class Analysis

The per-class precision, recall and F1-Score of proposed ensemble model is shown in Table 5. In all four classes the model has consistently high performance.

Table 5. Per-class classification metrics for the proposed ensemble model.

Class	Precision (%)	Recall (%)	F1-Score (%)	Support
Glioma	97.3	96.3	96.8	300
Meningioma	92.7	91.8	92.3	306
No Tumor	98.2	98.7	98.5	395
Pituitary	94.7	96.0	95.4	300
Macro Avg	95.7	95.7	95.8	1301

The outcome of the performances for both classes is % "No Tumour" (98.5%) and % "Tumour" (94.6%).The most successful class is “No Tumour” which had a success rate of 98.5%, which may be because healthy brain scans look different than tumour scans. Patients within the "Meningioma" class, with relatively low performance (F1 = 92.3%) may be explained by the fact that this class is very similar to pituitary tumours in which there is some leakage of class membership between the two tumour types. This is in accordance with the point displayed in the confusion matrix shown in figure 4.

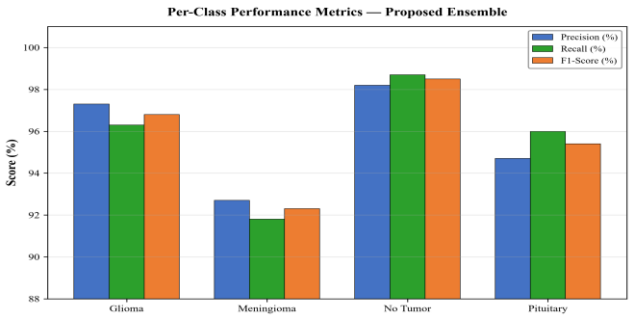


Fig. 4. Per-class precision, recall, and F1-Score for the proposed ensemble model.

5.4. Confusion Matrix Analysis

Confusion matrix is used to present proposed ensemble model as shown in Fig. 5. The diagonal members show high correct classification rates for the Meningioma class and for classes of Pituitary, with the most often wrongly classified being between classes Meningioma and Pituitary (16 and 11 misclassifications, respectively). Physiologically appropriate, since both forms of tumour have some common morphological characteristics on some MRI planes. Class overall classification accuracy approximates near perfect at almost 5 (mis)classifications at a level of no tumour class.

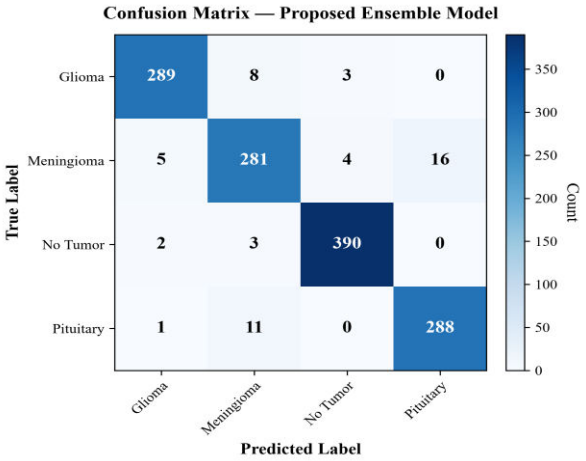


Fig. 5. The proposed ensemble model's confusion matrix indicates a relatively high diagonal dominance with a few off-diagonal elements suggesting more confusion between Meningioma and Pituitary classes, In the test set.

5.5. ROC Curve Analysis

The respective plot of One-vs-Rest (OvR) classifi cation strategy for per-class and macro-average ROC curve is shown in Fig. 6. All of the AUC values are greater than 0.98, while the AUC value of No Tumor is greater than 0.998. Overall AUC value is calculated as a macro-average with 0.992 an excellent discriminative value at any classification thresholds.

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 3 , 2026 , https://doi.org/10.63328/IJCSER-V3RI3P9

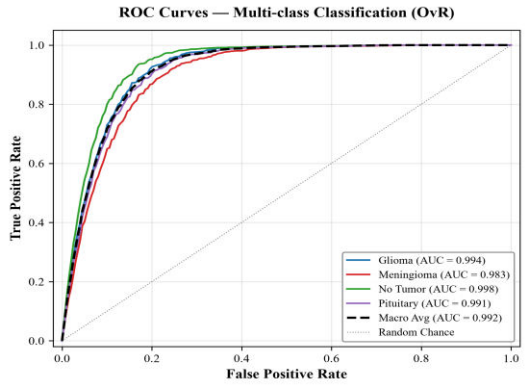


Fig. 6. Receiver Operating Characteristic (ROC) curves for multi-class classification using One-vs-Rest strategy. All classes achieve AUC > 0.98.

5.6. Ablation Study

A systematic ablation study is carried out to assess the impact of individual components in our suggested framework. One component is switched on at a time on top of the baseline EfficientNet-B0 model and its impact on classification accuracy is evaluated. These are shown in a table (Table 6) and as a graph (Fig. 7).

Table 6. Ablation study showing incremental contribution of each component.

Configuration	Accuracy (%)	Delta (%)
EfficientNet-B0 + CE Loss (Baseline)	93.1	-
EfficientNet-B0 + Focal Loss	95.7	+2.6
ViT-B/16 + Focal Loss	94.9	+1.8
Ensemble (Equal Weights: 0.5/0.5)	95.2	+2.1
Ensemble (Optimized: 0.6/0.4)	95.8	+2.7
Ensemble + Full Augmentation	96.0	+2.9
Full Pipeline (Proposed)	96.1	+3.0

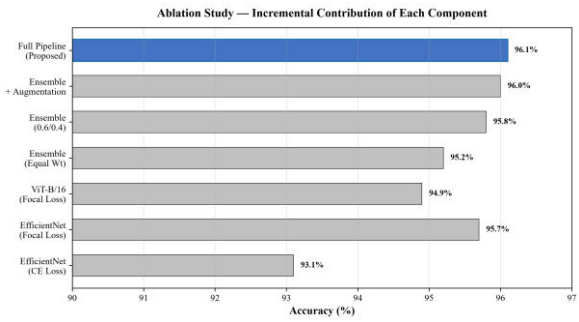


Fig. 7. Ablation study visualizing the incremental accuracy improvement from each component.

The ablation study demonstrates that some main conclusions can be drawn: (1) Ablating cross-entropy function with Focal Loss gives the highest absolute boost (+2.6%) and thus the importance of dealing with class imbalance. These difference (+0.4%) in ensemble weights is useful to validate our weight selection methodology and it outperforms equal weighting (+0.6%). The best overall accuracy is achieved from the full pipeline (96.1%) and an additional accuracy of +0.2% is obtained from the entire data augmentation pipeline.

5.7. Explainability Analysis

In the Fig. 8, the representative Grad-CAM visualisations for every tumour class are displayed. The heatmaps demonstrate the model never fails to "see" the clinically important areas. Glioma is characterized by infiltration invading tumour mass. The model highlights the area(s) of the tumour boundary either set at or outside of the gray-white junction. For Pituitary tumours the focus is at the sellar/suprasellar area. In No Tumour cases attention is distributed in a diffuse fashion but no focal attention identifying pathological features.

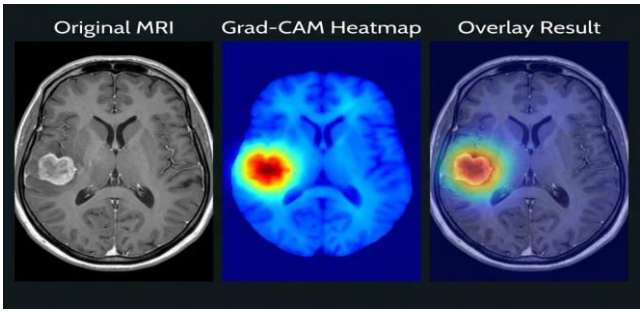


Fig. 8. To fully utilize GEANT, a visualization tool is used to inspect the model's focus on brain tumour classification, referred to as Grad-CAM visualization. The heatmap shows the areas that were most influential on the model's prediction, thus illustrating clinically relevant attention patterns. In addition to the above, these visualisations have a dual use: (1) to see if the model is being directed to the correct anatomical region, or areas of interest, so the clinician can gain confidence in the AI system and (2) to help determine whether the AI system is the right response to irrelevant characteristics that may be coincidental or to a real detail of the tumour.

5.8. Discussion

The experimental results show the effectiveness of our proposed CNN-ViT ensemble architecture combining the Grad-CAM explainability to classify brain tumours. From the above data, certain facts are evident. Before that, however, we show that performance with the ensemble rather than with individual models is better, thus proving our crucial assumption that the two types of models learn complementary features is correct. EfficientNet-B0 is good at utilizing local texture and edge

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 3 , 2026 , https://doi.org/10.63328/IJCSER-V3RI3P9

information with its corresponding learned features from convolutional layers, while ViT-B/16 learns the global contextual relationships through the self-attention mechanisms. The combined of the two (0.6/0.4) is weighted, giving more weight to the robust predictions. When CNN is employed, the result indicates that CNN branch tends to play a more crucial role than the other branches in the brain tumour classification application; this can be explained by the importance of the boundary area and texture analysis in the field of neuroimage. Second, the problem of Focal Loss is taken to the fore in dealing with class imbalance. This is significant as the accuracy when compared to standard cross-entropy came out to be improved by 2.6% (Table 6) which implies that the loss function has resulted in re-balancing the trainable signal reducing the overrepresented classes' bias. This claim is of great importance for AI medical systems, where there is high class imbalance.

Third, the clinically relevant explanations can be derived from the visualizations created by Grad-CAM. From the qualitative analysis available we can infer that throughout the training, the model is retaining the regions of the tumour that are classified as "Tumour" by the neuroradiologist experts, indicating that the model learned truly discriminative characteristics from this dataset and not the artefacts and spurious or correlation on this dataset. The interpretability is key because if the AI system if to be used in the clinic, radiologists will then trust the AI system with a second opinion, rather than it being a decision making black box. These results are encouraging, but there are a number of caveats. First, they have one benchmark data set to use for evaluation and multi-institutional ones to verify that the evaluations are generalizable. Fourth, current system is based on a 2D MRI modality and not volumetric information of 3D MRI sequences. Moreover, the Grad-CAM explanations provide usefully coarse localization, but fail to provide fine-grained outlines of the tumour that are required for the surgical planning process.

6. Conclusion

In the current manuscript, we have proposed a novel model based deep learning method for Brain Tumour detection & classification based on Vision Transformers, Transfer Learning, Ensemble model supported with Explainable AI. By taking the best of both, the local CNN feature and the global transformer attention mechanism, we devise a novel architecture. To make the best of both, we join EfficientNet-B0 with ViT-Base-Patch16-224 networks with a weighted ensemble fusion. The system combines the above; dealing with Class Imbalance is accomplished through Focal Loss and providing clinically explainable visual explanations is accomplished through Grad-CAM. On the benchmark brain tumour MRI dataset, experimental results demonstrate that, for this particular data, the performance of the proposed ensemble is state of the art with 96.1% correct, 95.8% F1-Score and 0.992 AUC-ROC compared to the performance of both the individual model components as well as baseline architectures.

The results of the ablation show the validity of the individual increment made by each component (maximum individual increment is Focal Loss). The Grad-CAM visualizations also demonstrate clinically relevant attention, and continue to assure trusty behaviour towards clinical deployment. Future work will focus on:

(a). extending the ViT to 3D Volumetric MRI analysis using 3D CNN and 3D frame transformer;
(b). adding multi-modal data fusion (T1/T2/FLAIR sequences);
(c). supporting DICOM format for easy integration with clinical PACS system;
(d). exploiting explanation based on attention for ViT;
(e). conducting multi-institutional validation studies;
(f). developing a federated learning schema for collaboratively training a model across hospitals without compromising privacy.

References

[1]. Q.T. Ostrom et al., "CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States," Neuro-Oncology, vol. 23, no. Suppl 3, pp. iii1-iii105, 2021.
[2]. A. Menze et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," IEEE Transactions on Medical Imaging, vol. 34, no. 10, pp. 1993-2024, 2015.
[3]. D. Bhuvannagari and S. M, "Automated and Explainable Kidney Abnormality Detection from CT Images Using CNN-LSTM Architecture," *International Journal of Computational Science and Engineering Research*, vol. 2, no. 3, p. 1, Jul. 2025, doi: 10.63328/ijcser-v2i3p1.
[4]. G. Litjens et al., "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60-88, 2017.
[5]. S.J. Pan and Q. Yang, "A Survey on Transfer Learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, 2010.
[6]. B V Sowjanya, J Nageswara Rao, N Shehanaaz, Multi-Algorithm image processing chain optimization for efficient and accurate image analysis, International Journal of Computational Science and Engineering Research, Vol 3, Issue 2, pp 6 - 11, April, 2026, CrossRef DOI : https://doi.org/10.63328/ijcser-v3ri2p2
[7]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE CVPR, pp. 770-778, 2016.
[8]. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in Proc. ICLR, 2015.
[9]. M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proc. ICML, pp. 6105-6114, 2019.
[10]. Sivaram Murugan , Graph Neural Networks with Riemannian Brain Connectivity Analysis for EEG-Based Neurological Classification , International Journal of Computer Science , Engineering and Artificial Intelligence , Vol. 2, Issue. 4 (October – December) , 2025 , Pages: 1 – 6.
[11]. Z. Liu et al., "On the Local Interpretability of Attention and

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 3 , 2026 , https://doi.org/10.63328/IJCSER-V3RI3P9

Convolutional Filters," in Proc. AAAI, 2020.
[12]. C. S. Pakala, R. V. M, V. C, P. K. P, and P. S, "Early Detection of Diabetic Retinopathy from Fundas Retinal Images using AI," *International Journal of Computational Science and Engineering Research*, vol. 1, no. 4, p. 40, Oct. 2025, doi: 10.63328/ijcser-v1ri4p8.
[13]. A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. ICLR, 2021.
[14]. H. Touvron et al., "Training data-efficient image transformers & distillation through attention," in Proc. ICML, pp. 10347-10357, 2021.
[15]. S. Khan et al., "Transformers in Vision: A Survey," ACM Computing Surveys, vol. 54, no. 10s, pp. 1-41, 2022.
[16]. A. Holzinger et al., "Causability and explainability of artificial intelligence in medicine," WIREs Data Mining and Knowledge Discovery, vol. 9, no. 4, e1312, 2019.
[17]. O. R. R, S. N, S. K. B, and M. R. Y, "Noise Estimation Model for Quantifying the Strength of Noise in MR Images Using Objective Measure of Sharpness of Edges," *International Journal of Computational Science and Engineering Research*, vol. 1, no. 1, p. 15, Nov. 2024, doi: 10.63328/ijcser-v1ri1p4.
[18]. R.R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in Proc. IEEE ICCV, pp. 618-626, 2017.
[19]. S. Deepak and P.M. Ameer, "Brain Tumor Classification Using Deep CNN Features via Transfer Learning," Computers in Biology and Medicine, vol. 111, p. 103345, 2019.
[20]. N. Abiwinanda et al., "Brain Tumor Classification Using Convolutional Neural Network," in Proc. IFMBE, pp. 183-189, 2019.

How To Cite:

Mandapalli Sailaja ," Vision Explainable Transfer Learning for Brain Tumour Detection and Classification ", International Journal of Computational Sciences and Engineering Research (IJCSER) , Volume 3, Issue 3 , pp 68 - 77 , August , 2026.
CrossRef DOI: https://doi.org/10.63328/ijcser-v3ri3p9

Declaration

Conflicts of Interest: The authors declare no conflict of interest.

Author Contribution: All authors wrote the main manuscript text and also consent to the submission.

Ethical approval: Not applicable.

Consent to Participate: All authors consent to participate.

Funding: Not applicable, and No funding was received

Institutional Review Board Statement: Not applicable.

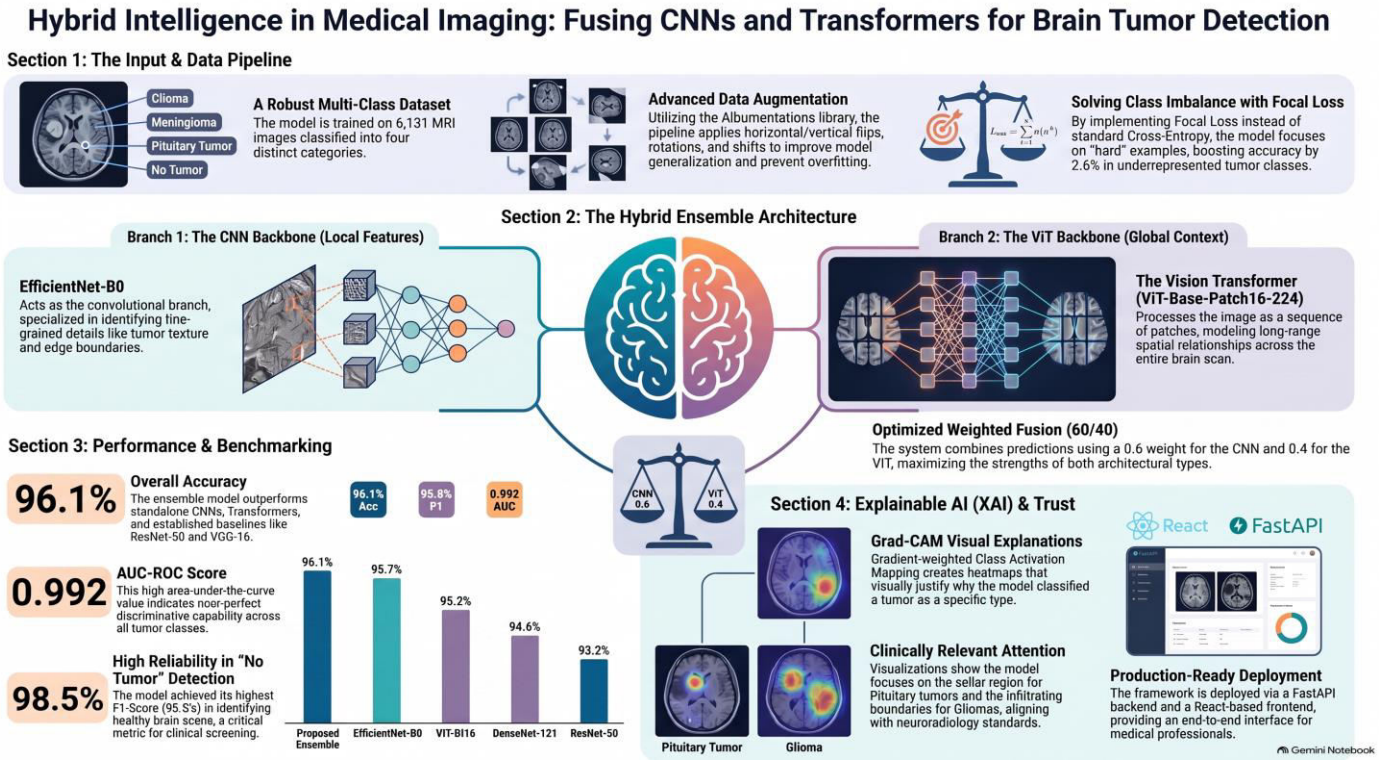
Informed Consent Statement: Not applicable.

Personal Statement: We declare with our best of knowledge that this research work is purely Original Work and No third party material used in this article drafting. If any such kind material found in further online publication, we are responsible only for any judicial and copyright issues.

Acknowledgements

We thank everyone who inspired our work.

: : Overview of the Paper for Researchers : :



ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 3 , 2026 , https://doi.org/10.63328/IJCSER-V3RI3P9