International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I8P128

Volume 74 Issue 8, 422-443, August 2026

Original Article

Explainable Deep Ensemble of Brain Tumor Classification based on MRI

Anuj Gupta[1], Anita[2], Manish Gupta[3]

[1,2]*Department of CSE & IT, Jaypee University of Information Technology, Waknaghat, Solan, Himachal Pradesh, India.*
[3]*Department of Radiation Oncology, Indira Gandhi Medical College, Shimla, Himachal Pradesh, India.*

[1]*Corresponding Author : nsanujgupta@gmail.com*

Received: 18 February 2026 Revised: 31 July 2026 Accepted: 07 August 2026 Published: 29 August 2026

Abstract - *Heterogeneity of cerebral neoplasms, loss of information labeling, and inter-class similarity make it difficult not only to differentiate the tumor clinically but also to establish separation of subtypes, taking Magnetic Resonance Imaging (MRI) into consideration. It presents a neurotumor multiclass classifier, which has a probabilistic soft voting ensemble as the classification method, based on the transfer learning model using pre-trained Convolutional Neural Networks (CNNs), using VGG16, ResNet50 and fine-tuned EfficientNetB0 as the basic classifiers. The experimental data is divided into four different categories: glioma, meningioma, pituitary and no visible tumor. The evaluation is carried out with the use of conventional performance indicators, which include recall, f1-score, confusion matrices, and a micro-averaged ROC analysis. Grad-CAM predicts clinical confidence and interpretability. The fine-tuned EfficientNetB0 model was found to have the highest test accuracy of 73.35%, compared to the equal-weight soft-voting ensemble with test accuracy 70.81%. The results indicate that the ensemble outperformed the ResNet50 and was marginally better than the VGG16, but still not better than the best EfficientNetB0 model. More frequent misclassification patterns, especially class confusion, are compared, and suggestions are made to enhance the sensitivity of the tumor.*

Keywords - *Transfer learning, VGG, ResNet, EfficientNet, Brain tumor, MRI, Explainable AI, Grad-CAM.*

1. Introduction

Tumors in the brain rank as one of the most serious brain disorders, as they can involve cognition, motor function, life span and quality of life. Primary Malignant Brain Tumors (PMBTs) account for a disproportionately high share of cancer-related deaths, despite being one of the least prevalent cancers, and are the most frequent types of tumors seen in adults, especially gliomas and meningiomas [1]. Magnetic Resonance Imaging (MRI) is optimal for brain tumor evaluation because it contrasts soft tissues and shows intracranial abnormalities. However, manual MRI scan interpretation is time-consuming and subjective owing to the range of tumor appearances, variations in MRI scan quality, tumor location, and the interpreter's clinical expertise.

Deep learning, specifically CNNs, is a revolutionary technology for automated medical image interpretation that can directly extract structured graphical properties from image data [2, 3]. Transfer learning with ImageNet-trained backbones has been found to be effective in brain tumour classification tasks [2, 4-6], and a number of public-dataset studies have shown very high classification results [2, 4-7]. Architectures that are broadly used are EfficientNet, ResNet and DenseNet, due to their representational ability, training stability and image classification capability [4, 8-10]. Transformer-based and hybrid models are also considered recently because they have the ability to capture the long-range spatial relationships using self-attention mechanisms [11, 12]. Even with these advances, there are a number of methodological and clinical limitations. Many reported results are based on a small number of datasets, image-level data splitting or non-patient-exclusive evaluation protocols, which may raise the risk of data leakage and lead to results that are too promising to be replicated in clinical use [5, 13, 14]. Furthermore, the overall accuracy is not enough as evidence of clinical reliability. A good aggregate accuracy can be attained with a poor sensitivity for a clinically relevant class, like glioma. Thus, class-wise brain tumor classification model evaluation should incorporate accuracy, precision, recall, F1-score, confusion-matrix analysis, and ROC-based evaluation.

Another critical need for medical decision support systems is interpretability. Ideally, model predictions should be determined by tumor-relevant anatomical regions and not background artifacts or irrelevant image patterns. Explainable Artificial Intelligence techniques, that include Grad-CAM and Grad-CAM++, can provide a visual understanding of the region that affects the CNN's predictions and aid in the

qualitative validation of the CNN's behaviour [15, 16]. These visual explanations, however, are meant to be supplementary and not a replacement for clinical validation. The use of ensemble learning to increase the robustness has also been investigated, as it has been found that by combining complementary models, the variance of the predictions can be reduced and the generalization can be improved [19, 20]. However, simple ensembling (unweighted) is not always superior. When a constituent model is weak, poorly calibrated or highly correlated with another model, the equal weighting of each model may degrade the ensemble performance [14, 20, 22, 29, 30]. Therefore, in ensemble classification of medical images, just reporting the final accuracy is not enough; an in-depth understanding of base-model behaviour, class-wise errors and the impact of the fusion strategy is crucial.

In this paper, the quadripartite brain MRI classification problem is tested with three transfer-learning CNN architectures: VGG16, ResNet50, and EfficientNetB0. Reduce data leakage using a stringent patient-wise train-validation-test methodology. A gentle voting methodology is used to amalgamate the results of the three models, and the performance of the ensemble is analyzed with respect to the strengths and weaknesses of the individual models. Further, Grad-CAM visualizations are inspected to see if the model is focusing on regions of the image that have anatomical meaning.

The most important contribution of this study is a clear and class-sensitive analysis of brain tumor MRI classification with the transfer-learning and ensemble approach in a stricter patient-wise protocol. The research emphasizes that it is not always the best single model that performs best when ensemble learning, with the inclusion of a weak learner, is used. It also presents extensive evidence of class-wise failure patterns, particularly a low recall for glioma, and considers the implications of these for the practical application to clinical decision support.

This study's main contributions are:

1. To minimize the risk of data leakage, a patient-wise experimental protocol is employed in brain MRI classification for four classes.
2. VGG16, ResNet50, and EfficientNetB0 are evaluated and compared using a unified preprocessing, augmentation, and evaluation pipeline.
3. The implementation and critical analysis of a probabilistic soft voting ensemble of base learners is implemented.
4. The class-wise performance is calculated by precision, recall, F1-score, confusion matrix and ROC-AUC analysis.
5. To explore model predictions, we use Grad-CAM visualizations to highlight the anatomical plausibility.
6. Limitations are highlighted, such as the sensitivity of gliomas, evaluation at the slice level, external validation and the use of such systems as an aid to the radiologist, not a replacement.

The rest of this paper follows this structure. In Section 2, deep learning, ensemble learning, and explainable AI for brain tumor MRI classification are discussed. Part 3 describes the dataset, pre-processing, data augmentation, model architectures, training settings, and ensemble approach. Section 4 presents the experiment and class-wise analysis results. Findings, clinical implications, limitations and future directions are discussed in section 5. This study ends with Section 6.

2. Related Work

The systematic categorization of brain tumors via MRI has been extensively explored since MRI offers abundant contrast from soft tissues and the ability for non-invasive evaluation of abnormalities within the brain. Previous computer-aided diagnosis techniques were based primarily on hand-crafted features, texture descriptors, morphological properties and traditional machine learning classifiers. While these approaches gave a good approximation of performance, they were limited in effectiveness as they required manual design of features and also did not adequately capture the complexity of tumor appearance, shape, intensity changes, and intra-class variability. Medical image classification is particularly popular in Computational Deep Learning (DL) using “Convolutional Neural Networks (CNNs)”, which can learn discriminative features from medical images [3, 24].

Transfer learning approaches were applied to tumor classification of MR images of the brain, particularly when the number of images in the dataset is small. Pre-trained CNN architectures have been widely adopted, including VGG, ResNet, DenseNet and EfficientNet, which enjoy good feature-extraction properties and minimise the requirement of training a network from scratch [4, 8-10, 25]. Despite having a simple stacked convolutional architecture, VGG-based models are still useful baseline models, and ResNet-based models overcome the vanishing gradient problem by utilizing a residual architecture [9, 25].

EfficientNet-based models have been gaining traction due to the desirable compromise of depth, width, and input resolution offered by compound scaling, along with the ability to achieve competitive results with a lower number of parameters than many traditional CNNs [4, 8]. Recent classification studies of brain tumors have shown that fine-tuned CNNs are able to achieve high accuracy on public MRI datasets [2, 4-6, 14, 26]. The reported performance is, however, very different across studies due to the differences in how data is acquired, classes are distributed, the preprocessing, augmentation, train-test splitting and evaluation strategy.

One of the critical points in the literature is that many good studies have adopted image-level random splitting, cross-validation without patient-level separation or augmented data that might not be completely independent of the test set. These practices can lead to overly optimistic results if images from the same patient or images that are very similar in the slices have been involved in both the training and testing sets [5, 13, 14]. This issue is especially relevant in brain MRI classification, as multiple slices from the same patient can have similar pathologies and anatomy. Hence, the reporting of accuracy values above 95% should be used with caution unless the study explicitly provides details about the separation of patients, external validation and class-wise performance. In addition, the overall accuracy may obscure poor sensitivity to clinically relevant tumor types. A model can classify a dominant or visually prominent class well, but suffer from errors in the classification of a small, heterogeneous class like glioma. A model can achieve high accuracy in the classification of a dominant or visually prominent class while being poor at the classification of a small, heterogeneous class like glioma. For a more dependable evaluation, class-wise recall, F1-score, confusion-matrix analysis and ROC-based analysis are required.

Besides CNNs, DL, and hybrid approaches have been proposed for brain tumor classification. CNN-based transfer learning and deep-feature techniques [18, 29] and transformer-based and hybrid CNN-transformer models [11, 12, 27] capture long-range spatial relationships using self-attention mechanisms to improve tumor classification.

Hybrid CNN-classifier pipelines have also been documented, in which handcrafted classifiers and classifiers optimized by the deep features from CNNs are used [18, 29]. These models have potential but can be influenced by the size of the data set, computational resources, optimization approach, and the availability of adequate labeled medical images. The transfer learning approach with well-established CNN backbones remains viable and repeatable for small and moderate size datasets.

Ensemble learning has been applied to boost the robustness of classification and to enhance generalization. The general principle of an ensemble is to use a collection of models, such that the strengths of the individual models can offset the weaknesses of the others [19, 20]. CNN ensembles are found to perform well in the brain tumor MRI classification, as long as the individual models are powerful and sufficiently diverse [30]. There are other methods based on weighted voting, selective ensemble formation, data balancing, and optimized ensemble strategies that have been proposed to improve the fusion of the models [14, 22, 29, 30]. Ensemble learning is not necessarily advantageous, however. If one of the learners has a poor performance or the same calibration (the same level of distribution) or one learner has high correlation error with the other one, it will affect the final ensemble performance [20, 22, 29, 30]. Hence, the design of an ensemble should be complemented with the analysis of base models, class-wise error analysis and justification of the fusion strategy.

Another key field of medical imaging that falls under the umbrella of XAI is the need to explain the images to the clinical user as well. However, another crucial field of medical imaging is the need to explain the images to the clinical user, not just the predicted label. An evidence-based decision-support model should also offer proof of predicting in clinically relevant image regions and not artifacts, borders, scanner-specific patterns or background noise [17].

Two post-hoc visual explanation methods, named Grad-CAM [15] and Grad-CAM++ [16], are popular methods to extract the class-discriminative heatmaps from CNN feature maps. In brain tumor MRI studies, grad-CAM has been used to extract tumor-relevant regions and to aid the qualitative interpretation of model predictions [15, 17, 31, 33]. These techniques can be useful to determine whether network learning is biased toward areas similar to the actual tumor structures or if incorrect predictions are linked to irrelevant activation patterns.

However, though useful, XAI methods have limitations as well. Grad-CAM is not capable of providing a fine localization map and cannot completely reveal the inner mechanism of a deep model. The heatmap can differ based on the layer, model architecture and preprocessing technique. They can even be a good guess if the prediction is incorrect. So, it is essential to consider Grad-CAM as an interpretability tool rather than something clinico-pathological [17, 31, 33]. For the medical application, an explainability should be supplemented with quantitative validation, class-wise analysis, expert review and testing on independent datasets.

The literature studied revealed that deep learning, transfer learning, ensemble learning and XAI have all made a significant contribution to the categorization of cerebral neoplasms with MRI. However, there are still a few things that have not been resolved. First, numerous studies only focus on the overall accuracy with incomplete patient-wise validation.

Second, failure patterns are not studied as much at the class level, particularly for clinically relevant and varied tumour types like glioma. Thirdly, ensemble models are presented as always being better, without any analysis of the poor performers and fusion behaviour. Fourth, the outputs of XAI are sometimes presented as visual means and are seldom discussed critically regarding their limitations. In the present study, we fill in these gaps by comparing VGG16, ResNet50 and EfficientNetB0, using a common patient-wise protocol, assessing a soft-voting ensemble and reporting the class-wise performance, and examining the anatomical plausibility of the model predictions by the Grad-CAM.

3. Dataset and Experimental Protocol

This section describes the features of the dataset, the preprocessing approach, the augmentation strategy used, model architectures, training configuration, the ensemble fusion methodology, evaluation metrics, statistical reliability analysis and explainability procedure used in this work (based on Grad-CAM). The entire schematic of the planned experimental setup is shown in Figure 1. The framework first involves collecting and preprocessing brain MRI images, then dividing and segmenting the images on the patient level, constructing patient-specific transfer-learning models, tests them per model, summarizing the probabilities learned, and qualitatively interpreting them with the Grad-CAM visual explanation. The experimental protocol was designed to be transparent, reproducible and minimize the data leakage in the evaluation of multi-class brain tumor classification using MRI.

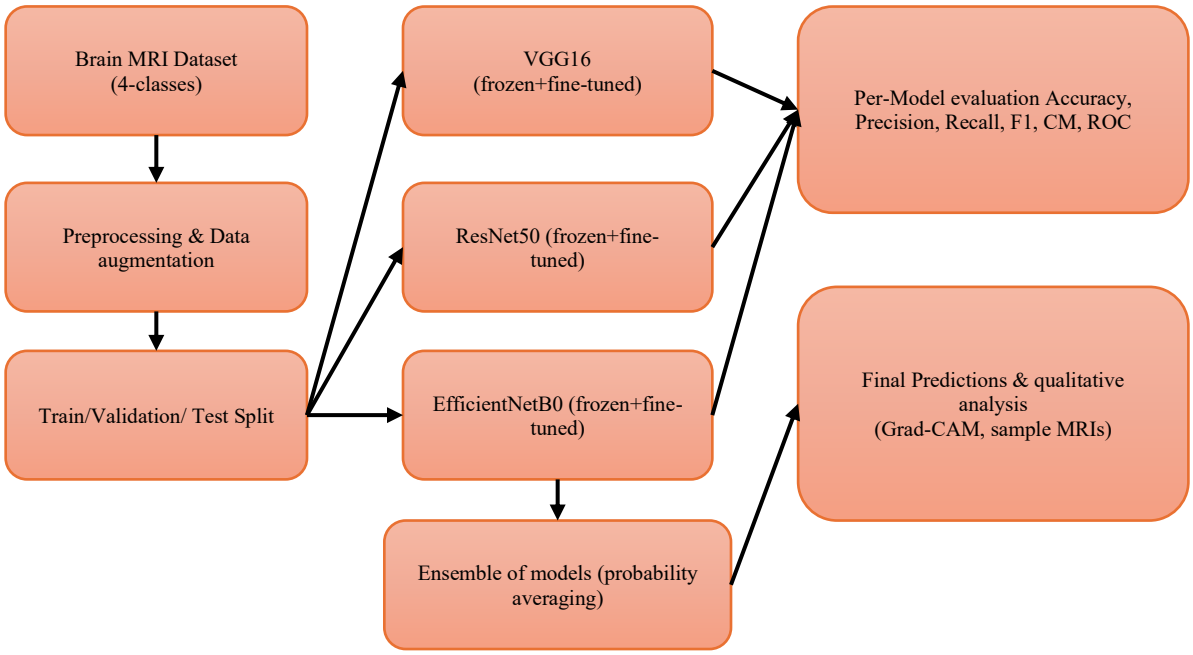

Fig. 1 Generalization of the entire process of the aforementioned ensemble transfer learning architecture for multi-class brain tumor categorization using Magnetic Resonance Imaging

3.1. Dataset

Axial contrast-enhanced T1-weighted brain MRI slices were used in the investigation. The photos were categorized as glioma, meningioma, pituitary, or no tumor.

Sample data for illustrative examples are provided for the four categories in Figure 2. The majority of the tumor images were acquired from the dataset reported by Cheng et al. [25], and normal or no-tumor MRI images were acquired from other public repositories used in previous brain tumor classification studies [2, 4-7]. These four are used because the task is clinically relevant-you have to differentiate the tumor cases from MRI images that look normal, and different tumor types from one another.

The images in the data set were first filtered to exclude low-quality, corrupted and duplicate images before training. This step was essential as repeated or near-repeated slices of the MRI would bias the model learning and could cause the model to evaluate the performance in an optimistic way if the same MRI image appears in both the training and the testing sets. 3264 MRI slices were kept for the final experiment after cleaning. Patient-wise partitioning divided the images into training, validation, and testing sets. The training set had 2297 images, the validation set 573, and the independent test set 394.

Patient-wise split and not a random split of Images. This was necessary because the MRI slices in question may be from the same patient and/or have very similar anatomical structures. When slices from the same patient are used in training and testing sets, the model may acquire patient-specific visual patterns instead of tumor discriminative features.

This may overestimate performance and impair clinical interpretation validity. Thus, the train, validation, and test sets were independent. The validation set was utilized to select the model for learning-rate scheduling and early stopping, while the test set was kept only for the last performance reporting.

The data set has two practical issues. Firstly, there are differences in anatomical location, texture, boundary definition and intra-class variation between the four classes.

The appearances of gliomas can be diffuse and heterogeneous, meningiomas tend to be located in the vicinity of the dural or peripheral areas, pituitary tumors are more likely to be found in the vicinity of the sellar area, and no-tumor images show no evidence of a mass lesion. Second, the analysis is performed on two-dimensional MRI slices while clinical diagnosis is typically based on full volumetric MRI scans and multiple imaging sequences. These features render the task more challenging than just binary tumor detection and more realistic.

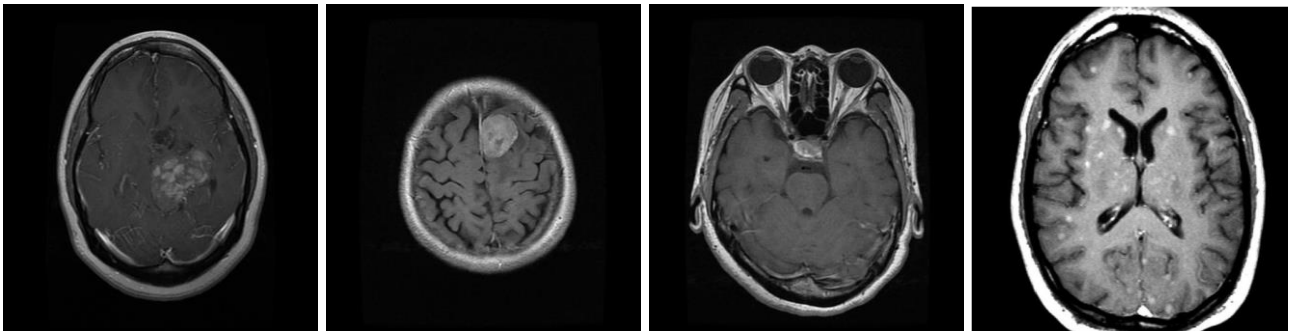

Fig. 2 MRI specimens of glioma, meningioma, pituitary tumors, and no-tumor tissues of the Brain Tumor Dataset

3.2. Preprocessing and Augmentation

Magnetic Resonance (MR) image slices were scaled to 224 × 224 pixels to match the input sizes of pre-trained CNNs (VGG16, ResNet50, and EfficientNetB0). Resizing was done using bilinear interpolation to maintain the general anatomy of the MRI slice when resizing it in space. The grayscale MRI slices were transformed to a three-channel image wherever needed by the selected transfer-learning model, which was pre-trained on images of ImageNet. This conversion was done by copying the grayscale channel to three channels to keep the input dimension the same as the pretrained convolutional filters.

In the entire dataset, the preprocessing procedure was maintained similarly, while a normalization was implemented based on the requirements of the respective architectures. The intensity values of the pixels were linearly scaled to the range 0 to 1 for VGG16 and ResNet50, which were implemented with PyTorch. Following this scaling, the images were standardized with ImageNet channel-wise mean and standard deviation values recommended for such models trained on ImageNet [35], which were equal to (0.485, 0.456, 0.225) and (0.229, 0.224, 0.225), respectively. This was required since the weights of these CNN backbones are pre-trained on a distribution of images that is similar to the distribution in the ImageNet training set. If not, the feature maps extracted from the network might not be stable and/or transferable to the target MRI classification task.

The images had been resized to 224 × 224 and passed through the EfficientNetB0-specific preprocessing routine provided by the Keras/TensorFlow library. This meant that all the input images were processed according to the same pre-processing standard as part of the original EfficientNetB0 pre-training. The model-specific preprocessing was used to preserve compatibility with pre-trained parameters, while minimizing any risk of variations in performance due to differing input data scaling. Data augmentation was only employed on the training set, not the validation or test. This design was followed to guarantee that the validation and test results were not from artificially transformed images, but from images not used in the model construction. The validation and test images were thus only resized and normalized. Augmentation vs. validation and testing was significant to keeping the protocol of evaluation fair and realistic.

To avoid overfitting and make the model more robust, the images were augmented online during model training to give more diversity to the samples. Small rotations (within 15 degrees), translations (in the horizontal and vertical directions up to 10%), zooming and scaling (between 10% and 15%), and horizontal flipping were performed as augmentation operations. The selected transformations are reasonable variations of the MRI acquisition, patient positioning, and alignment of the field of view. However, the changes were kept reasonable to avoid creating images that were unrealistic or misleading for clinical purposes.

No strong intensity manipulation, heavy cropping, aggressive deformation or elastic warping. This limitation was essential as the intensity pattern in MRI and the contrast of tumors could contain clinically relevant information. If the intensity or structure is changed too dramatically, it can change the diagnostic cues and can add artificial patterns that aren't in real MRI images. As the images are grayscale MRI slices, color-based augmentation was also disregarded.

The augmentation strategy is significant due to a small number of images in the dataset and high intra-class variation. The images of gliomas, meningiomas, and pituitary tumors vary in tumor location, shape, boundary characteristics, and texture, and no-tumor images might undergo additional anatomical changes that could impact feature learning. The models were able to learn more generalizable spatial and

textural representations with the help of moderate augmentation rather than memorizing fixed training examples.

Augmentation has also been recognized as a key element for improving the generalization of the brain tumor MRI classification task in previous studies [2, 5, 18]. A lack of augmentation caused early overfitting in the current study, particularly for the VGG16 architecture, where the training accuracy reached to high levels and stabilization, or saturation was observed for the validation accuracy.

Following the introduction of augmentation, the validation accuracy was increased, and the difference between training and validation accuracy was decreased. Initial findings showed that augmentation could improve the VGG16 model's prediction performance by 5%. Therefore, augmentation was incorporated as an essential regularization protocol in the last experimental protocol.

The overall preprocessing and augmentation process aimed to satisfy three criteria: to keep the MRI compatible with the previously trained CNN models, to retain clinically relevant MRI data, and to minimize overfitting. The images were sized to the same input size to ensure the same input pipeline for all models evaluated, normalized using a scheme specific to the architecture, limited to augmentation of the images used in training, and only used realistic transformations.

3.3. Model Architectures and Training Configuration

Three base learners obtained from CNNs, trained on ImageNet, were chosen for the multiglass brain tumor classification task in this paper: VGG16, ResNet50, and EfficientNetB0.

In order to present three different architectural design philosophies, these models were chosen to be used for image classification and medical image analysis, which is a large field. The flexibility of different pretrained CNN backbones also enabled the study to investigate the effect of ensemble fusion on the robustness of classification.

The classical transfer-learning baseline VGG16 was added to the baseline. It consists of a deep convolutional network with primarily stacked 3×3 convolutional layers with max pooling operations. It has a uniform architecture, which makes it easy and effective to extract features. Nevertheless, the model VGG16 has a large number of trainable parameters, particularly when dense layers are employed, which can lead to overfitting if the medical imaging dataset is small [36].

Because brain MRI datasets are tiny, dropout, augmentation, early halting, and careful fine-tuning were needed to minimize overfitting in the VGG16 model. We chose to use ResNet 50, a deeper residual-learning architecture. It adopts identity skip connections that enable the gradients to traverse the network more smoothly and facilitate deeper learning models [9]. The usefulness of these residual connections lies in their ability to make up for the vanishing gradient problem and to learn hierarchical representations of images. ResNet50 is a popular baseline architecture for transfer learning image categorization because it can learn complex visual patterns. ResNet50 was introduced to assess residual feature learning for glioma, meningioma, pituitary tumor, and no tumor discrimination.

We choose EfficientNetB0 because it balances model size, computational cost, and classification quality. In efficientNet models, depth, width, and input resolution are balanced compound scaled [8]. EfficientNetB0, the family's baseline model, has 5.3 million parameters, less than certain CNN designs but considerable representational capability [4, 8]. This is useful in medical imaging applications where the data is small and a large model could overfit. Thus, EfficientNetB0 was expected to balance accuracy and generalization.

In all three architectures, the original imageNet classification layers were discarded and re-designed a task specific classification head. Global average pooling, dropout, and fully connected softmax output layers with four neurons per target class (glioma, meningioma, pituitary tumor, and no tumor) existed in the new Global average pooling, which reduced overfitting and parameter training, unlike massive flattening-based linked structures.

The softmax layer gave me class probability scores for the four classes. To prevent overfitting, dropout regularization was used in the classification head. In VGG16 and ResNet50 the classification head employed two Dropout layers with drops of 0.4 and 0.3, respectively. Tuned EfficientNetB0 used a dropout rate of 0.4 after batch normalization and dense feature extraction, as it is more parameter efficient and already has built-in more robust regularization behaviour. A similar task-specific head was used for the three models for a fair comparison, but not to prevent architecture-specific regularization.

Due to its flexible learning-rate behavior and deep neural network optimization, the Adam optimization method was used in training. The loss function was categorical cross-entropy because the job had four incompatible categories. The loss function penalizes probability misassignments for all classes, making it fit for multi-class classification.

All the models were trained for 30 frozen-stage epochs and 70 fine-tuning epochs. Training was stabilized using model checkpointing and learning rate reduction, with the objective of keeping the best model that performed best on the validation data. Data was transferred using two-stage transfer learning for all three models. First, only the classification head was trained while the pre-trained convolutional backbone was

frozen. In the second stage, higher-level characteristics for MRI classification were fine-tuned. A learning rate of 1 x 10 -3, was employed in the frozen stage, and 1 x 10 -5 was employed for the fine-tuning stage. Learning rate was reduced using ReduceLROnPlateau when the validation performance stopped improving. A summary of the final architecture and hyperparameter settings for each model is shown in Table 1.

To maintain model comparison consistency, all three models were trained using 32 batches. All CNN models (VGG16, ResNet50, EfficientNetB0) used the identical training, validation, and test partitions.

An NVIDIA A100 GPU was used for training. All models shared the identical patient-wise training, validation and test partitions. Validation loss, early stopping, learning rate reduction, and model checkpoint selection were done with the validation set. The independent test set was not employed for model training or hyperparameter selection. The hyperparameters that were searched were learning rate, dropout rate, batch size, and fine tuning depth with constrained conditions. Final configurations were chosen based on validation behaviour, convergence stability, and control of overfitting. Due to computational constraints, an exhaustive grid search was not performed. The final hyperparameter and architecture settings are listed in Table 1.

Table 1. Hyperparameter and architecture setting

Configuration	VGG16	ResNet50	Tuned EfficientNetB0
Input size	224 x 224 x 3	224 x 224 x 3	224 x 224 x 3
Pre-trained weights	ImageNet	ImageNet	ImageNet
Backbone design	Stacked 3 x 3 convolution blocks	Residual bottleneck block with skip connections	MBConv block with compound scaling
Classification head	Global Average Pooling→ Dense(512,ReLU)→ Dropout(0.4)→ Dense(256, ReLU)→ Dropout(0.3)→Dense(4, Softmax)	Global Average Pooling→ Dense(512,ReLU)→ Dropout(0.4)→ Dense(256, ReLU)→ Dropout(0.3)→Dense(4, Softmax)	Global Average Pooling→ Batch Normalization→ Dense(512, ReLU)→ Batch Normalization→ Dropout(0.4)→Dense(4, Softmax)
Optimizer and loss	Adam; categorical cross-entropy	Adam; categorical cross-entropy	Adam; categorical cross-entropy
Learning rate schedule	1 x 10^{-3} frozen stage; 1 x 10^{-5} fine-tuning; ReduceLROnPlateau	1 x 10^{-3} frozen stage; 1 x 10^{-5} fine-tuning; ReduceLROnPlateau	1 x 10^{-3} frozen stage; 1 x 10^{-5} fine-tuning; ReduceLROnPlateau
Training epochs	30 frozen + 70 fine-tuning = 100	30 frozen + 70 fine-tuning = 100	30 frozen + 70 fine-tuning = 100
Batch size	32	32	32
Data split	2297 train/ 573 validation/ 394 test	2297 train/ 573 validation/ 394 test	2297 train/ 573 validation/ 394 test
Hardware	NVIDIA A100 GPU	NVIDIA A100 GPU	NVIDIA A100 GPU

Overall, the training configuration aimed to ensure a fair comparison among the three CNN backbones, thereby guaranteeing reproducibility and minimizing overfitting.

Furthermore, it was found that the use of pre-trained weights enabled efficient learning, dropout and augmentation were used to regularize the learning, early stopping prevented the over-training of the model, and careful fine-tuning enabled the adaptation to the MRI-specific features.

These trained base models were then evaluated individually and were also used in the next section, where they were utilized to build the soft-voting ensemble.

3.4. Ensemble Method

The individual constituent models were trained and validated, and then an ensemble classifier was formed to integrate the predictive outputs from the individual models. This work examined how VGG16, ResNet50, and EfficientNetB0's complementing feature-learning behavior can improve multi-class brain tumor MRI classification. Three base models record distinct visual qualities according to their designs. The stacked convolutional blocks in VGG16 can enable learning powerful local convolution patterns, while the skip connections in ResNet50 can enable deeper residual representations, and compound scaling in EfficientNetB0 can enable compact and efficient feature representations. The

usefulness of combining these models may then be possible if there are differences in the prediction error they produce, and if the learned representation of the models is complementary.

In the present study, a Probabilistic Soft-Voting (PSV) ensemble strategy was employed. The reason for using soft voting rather than hard voting is that the former makes use of the full class-probability distribution predicted by each model, while the latter uses the class label predicted by each one.

In hard voting, each model gives one class label, and the majority vote determines the forecast. Hard voting ignores model confidence. Soft voting uses the relative confidence level of each classifier based on class probability. If two tumor classes look similar, the confidence distribution may contain more information, which helps with medical image categorization. Each trained model produced probability scores for glioma, meningioma, pituitary tumor and no tumor for each of the test images. So, the ensemble probability was derived as the mean of the class probabilities outputted by all the constituent models, $P_m(c|x)$ represents the probability assigned to class c by the m^{th} model for input MRI image x:

$$P_{ens}(c|x) = \frac{1}{M}\sum_{m=1}^{M} P_m(c|x), \quad (1)$$

where M =3 denotes the number of constituent models: VGG16, ResNet50 and EfficientNetB0. Then the class of the ensemble was predicted as:

$$\hat{y}_{ens} = argmax_c P_{ens}(c|x), \quad (2)$$

where $\hat{y}_{ens}$ represents the final ensemble prediction. This formulation ensured that all models contributed to the final class decision through their probability distributions.

The three base models were all given equal weight in this study. So, each model has been assigned a 1/3 probability to the final ensemble. There were three primary reasons for using equal weighting. First, it offered a straightforward and clear fusion strategy that could be easily reproduced. Secondly, it did not require the training of an extra meta-classifier to combine the classifiers, because the validation set was small.

Thirdly, it enabled the study to directly assess if a simple soft vote strategy could outperform the individual transfer learning models. In the methodological planning, weighted voting was considered with regard to the validation accuracy, but weighting by validation accuracy was not used in the final experiment because of excessive tuning in the validation set due to the use of weighted voting, and to keep the method simple. Weighted voting can contribute to enhancing the performance of the ensemble if stronger base learners are given more weightage and weaker base learners are given less weightage. However, this approach should be carefully validated, since weights derived from a small validation set may not be representative of the test data. For this reason, the unweighted soft voting ensemble was adopted as the main ensemble structure.

The critical analysis of the ensemble method was also performed since ensembling alone does not necessarily lead to improved classification performance. If the constituent models are individually quite reliable but make complementary errors in their predictions, a soft voting ensemble can help decrease the number of errors due to randomness.

However, if one of the models is very poor or gives poor probability estimates, the same approach can become less effective. If so, the averaged class probability in the weak model may be misled away from the right class, and finally, low accuracy of the ensemble may be obtained. The ensemble results were thus evaluated with the performances of the individual base models and not as a "per se" superior result.

This is very significant in the current task of four-classes classification of MRI. Two-dimensional MRI slices may have similar visual features of tumor classes like glioma and meningioma. When consistently misclassifying a base model as a different class, the probability contribution of a base model can decrease the ensemble sensitivity for glioma.

The same applies if a model is poorly calibrated or assigns high probability to the wrong class, as probability averaging can reduce the impact of more accurate models. The ensemble was assessed for accuracy, class-wise precision, recall, F1-score, confusion matrices, and ROC-AUC.

The soft voting ensemble was thus used for two purposes in the study. First, it gave a combined classifier using the prediction output of VGG16, ResNet50 and EfficientNetB0.

Second, it enabled an empirical investigation in a clinically relevant MRI classification setting about whether a simple ensemble fusion is enough. The results of this analysis are discussed in subsequent sections with regard to the impact of the weaker base learner and the necessity to use weighted or selective ensemble strategies in the future.

Table 2. Proposed soft-voting ensemble framework pseudocode

"**Step**	**Operation**
1.	Input: training set D_{train}, validation set D_{val}, test set D_{test}, and class set $C = \{$glioma, meningioma, no tumor, pituitary$\}$.
2.	Preprocess all MRI slices to 224 × 224 × 3 and apply augmentation only to D_{train}.

3.	Train VGG16, ResNet50, and EfficientNetB0 independently using the same train-validation split.
4.	For each test image x, obtain probability vectors. $P_{VGG}(x)$, $P_{ResNet}(x)$, and $P_{EffNet}(x)$.
5.	For each class c , compute the ensemble probability as $P_{ens}(c\|x) = \{P_{VGG16}(c\|x) + P_{ResNet50}(c\|x) + P_{EfficientNetB0}(c\|x)\}/3$
6.	Assign the final class as $\hat{y}_{ens} = argmax_c P_{ens}(c\|x)$.
7.	Evaluate the ensemble using accuracy, precision, recall, F1-score, confusion matrix, ROC-AUC, confidence intervals, and paired statistical tests."

The most important drawback of the ensemble strategy used is that no discrimination was made among the models based on their validation behaviour and/or the test performance. It helped keep things transparent, but also allowed it to fall prey to weak models in its individual parts.

Other methods of fusion, like those based on validation accuracy, class-specific weighting, selective model exclusion, stacking or calibration-based fusion, can be used to get better results. However, these methods require further proving and could add to the computational complexity. Hence, in the present study, the unweighted soft voting approach was taken as a reproducible baseline ensemble method, and its advantages and disadvantages were investigated in detail based on experimental results.

4. Results

In this section, the results of the single transfer-learning models and the soft voting ensemble are reported on the independent test set. Training and validation behaviour are evaluated; test accuracy is calculated; interpretation of the confusion-matrix is provided; ROC-AUC analysis is carried out; precision and recall for each class are computed; F1 score is computed; reliability assessment is performed, and the results are displayed; and a visual explanation is provided using Grad-CAM analysis. The results are presented not only as overall accuracy, but also on a class-wise basis, as a medical image classification model could perform well even if it is not able to detect a clinically significant tumor type.

The final examination made use of an independent test set containing 394 MRI pictures. The test set was not used for model training, hyperparameter selection, early halting, or checkpoint models. This separation ensured that reported figures were based on model performance on new data. This section tests VGG16, ResNet50, EfficientNetB0, and an ensemble model created by soft voting the three models' probability outputs.

4.1. Training Dynamics

VGG16, ResNet50, and EfficientNetB0 training and validation accuracy and loss curves are shown in Figure 3. Each architecture's convergence behavior, optimization stability, and generative capability are shown by the curves. During training, the VGG16 model converged quickly. It has been provided with a rapid rate of learning and has attained a high level of accuracy, around 95-100%. However, the validation accuracy stagnated around 70% and became unstable during the later portion of training. Overfitting suggests that the training performance is worse than the validation performance, as there is a widening gap.

This is normal because VGG16 has many parameters and can memorize training examples if medical imaging data is scarce. The augmentation, dropout, and early stopping approaches reduced overfitting, but the training curves show that VGG16 learned something more useful in the training set than in the validation data.

The training patterns of the ResNet50 model were the least promising and fluctuating out of the three architectures. The accuracy of validation was relatively low and ranged from about 40-50%, and the validation loss was high and unstable. This means that the chosen ResNet50 model was not successfully transferred to the current MRI classification task. The instability might be due to the small data size, learning-rate sensitivity or fine-tuning the depth or difficulty of fitting the deep residual features to the two-dimensional slices in a contrast-enhanced MRI. Later, the lack of validation behaviour of ResNet50 was demonstrated by its poor test-set performance.

The EfficientNetB0 had the most consistent learning behaviour, however. The training and validation curves of its were closer than those of VGG16, and the pattern of its validation loss was relatively stable. The validation accuracy was about 73%, which is the highest of all the architectures tested. This finding suggests that EfficientNetB0 provided a better balance between model complexity and generalization. Compound scaling efficiently scales depth, width, and input resolution in EfficientNetB0 [8]. Previous work using EfficientNet topologies on small or moderate medical imaging datasets yielded similar results [4, 14].

The sum of all the training dynamics shows 3 behaviours. VGG16 had a nice performance in training; however, its performance was quite overfitting. Under the currently used configuration, ResNet50 did not converge and was not transferable. After training, each model was evaluated on 394 images. Tables 3 and 4 show individual and ensemble model accuracy on test data, precision, recall, and F1 scores for each class.

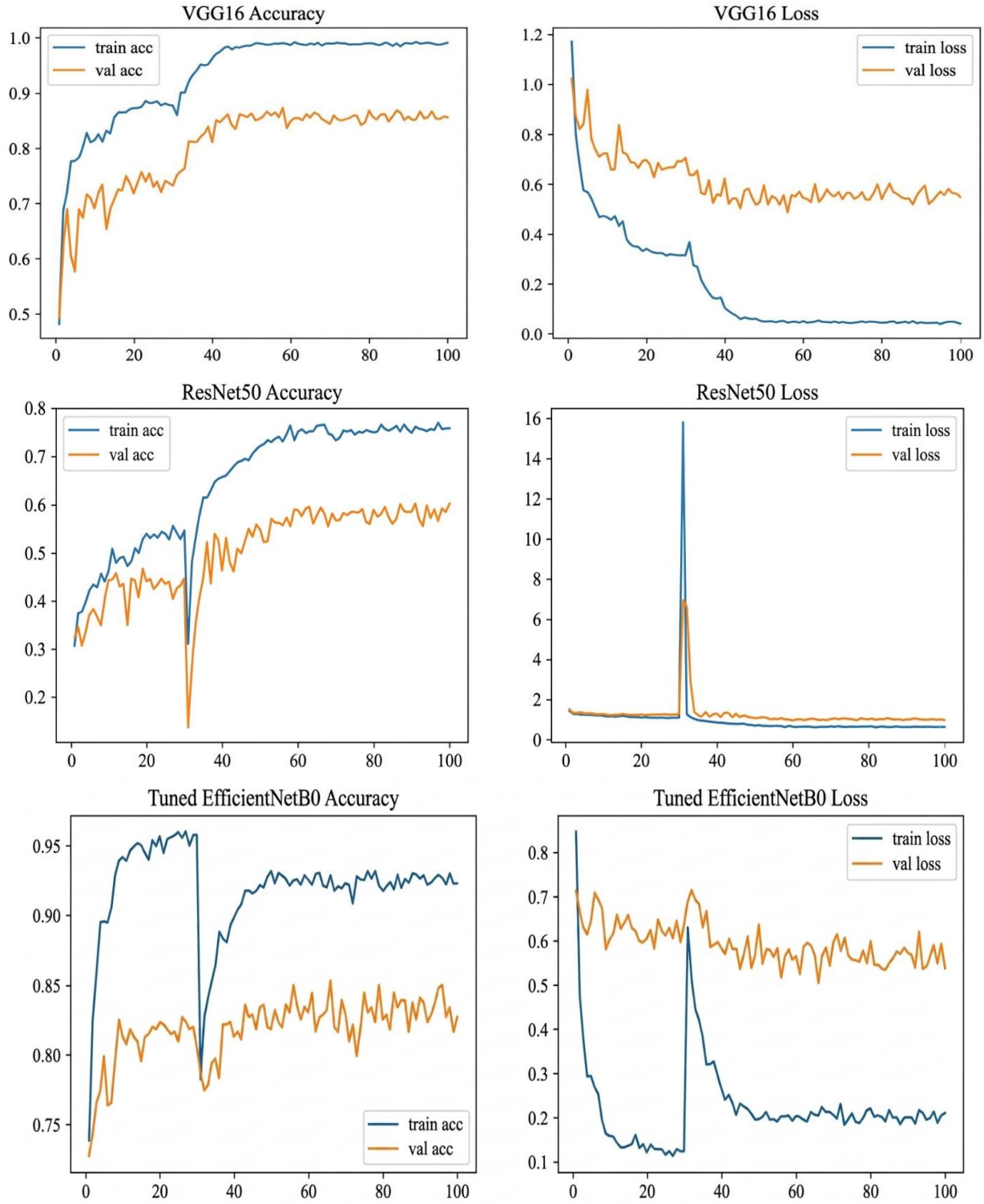

Fig. 3 Training & validation curves for VGG16, ResNet50, and EfficientNetB0, highlighting VGG16 overfitting, ResNet50 instability, and EfficientNetB0's stable best performance (~73%)

4.2. Test Set Performance

Each model had been assessed on a 394-image dataset after training.

Table 3 demonstrates individual and ensemble model accuracy on test data, whereas Table 4 shows class precision, recall, and F1 score.

Table 3. Overall test accuracy for all evaluated models (on 4-class classification)

Model	Accuracy (%)
VGG16	70.56
ResNet50	40.10
EfficientNetB0 (fine-tuned)	73.35
Ensemble (VGG+ResNet+EffNet, soft voting)	70.81

The best test accuracy overall was 73.35% obtained by the fine-tuned EfficientNetB0 model. VGG16 was able to achieve a competitive accuracy of 70.56% while ResNet50 was able to achieve a very low accuracy of 40.10%.

The soft-voting ensemble model obtained 70.81% accuracy, which is slightly better than the VGG16 model but poorer than the EfficientNetB0 model. This means that the equal-weight ensemble model did not perform better than the best individual model.

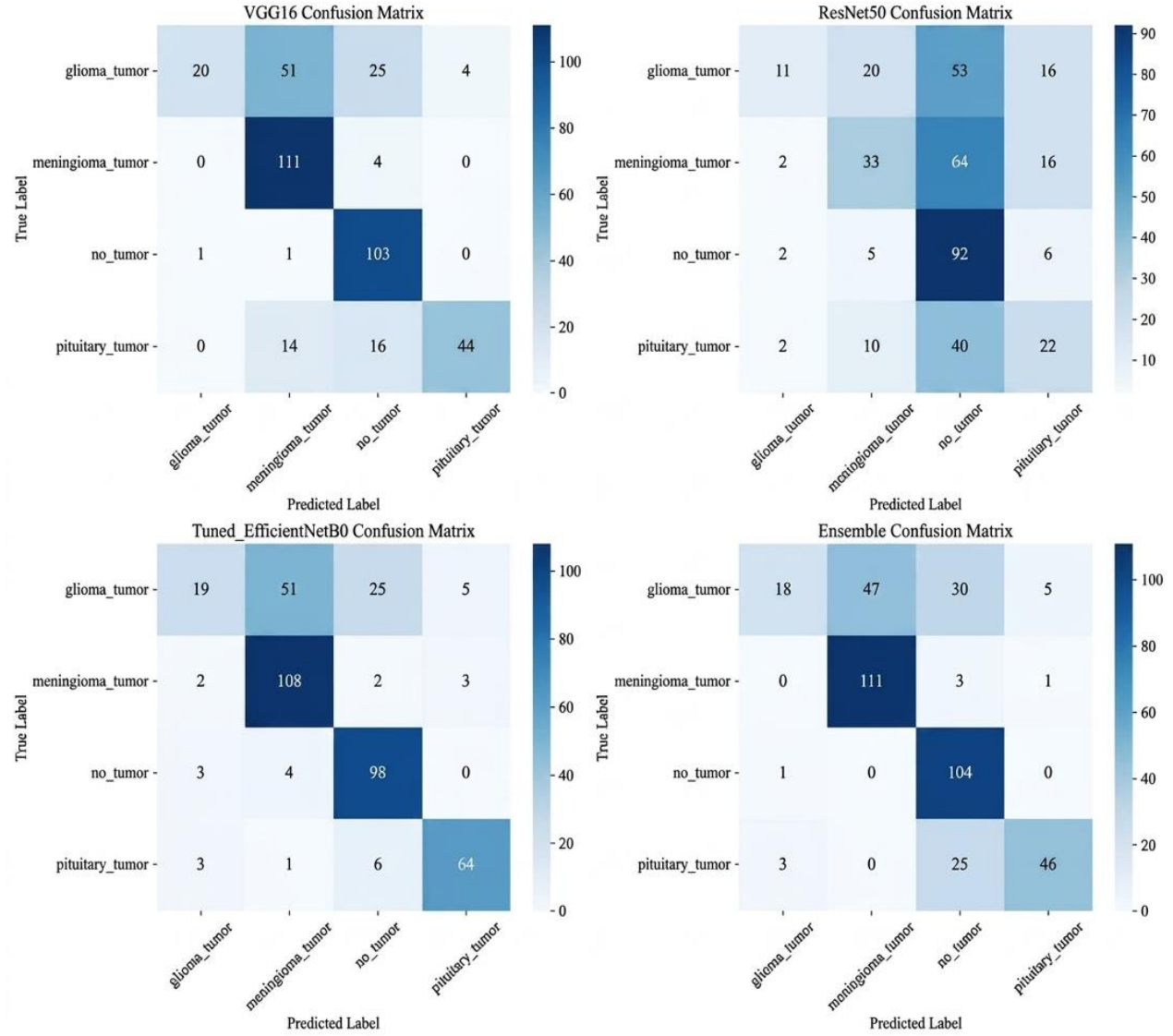

Fig. 4 Test-data confusion matrices for VGG16, ResNet50, EfficientNetB0 and the ensemble, EfficientNetB0 best class balance, poor glioma recall in VGG16 and the ensemble and poor overall performance for ResNet50

The result of ResNet50 was comparatively poor with an accuracy of 40.10%, which was significantly below the VGG16, EfficientNetB0 and Ensemble. This performance validates the fact that the chosen configuration of ResNet50 was not well transferred to the current MRI classification task. This rate is statistically above the random classification rate of

25% for a 4-class problem, but is not high enough to be reliable for medical classification. It is worth noting that ResNet50 performs worse than the other networks on test data, which reflects the instability of training and validation behaviour exhibited by ResNet50, as seen in Figure 3. This suggests that ResNet50 was not able to learn very transferable representations with the chosen experimental setup.

The soft voting ensemble gave an accuracy of 70.81%. This was slightly better than VGG16 but still worse than the best model, EfficientNetB0. This indicates that there is no obvious performance improvement using equal-weight probability averaging over the best base learner.

This is a significant result as it demonstrates that ensemble learning alone is not necessarily the best individual model. The low contribution of ResNet50 had a negative impact on the ensemble. Given the use of the equal probability averaging, ResNet50's individual performance was poor, yet it accounted for one-third of the final ensemble probability. This decreased the strength of the ensemble and moved some predictions off the correct class.

Therefore, there are two important observations to be made from the results of the test. Firstly, the best single architecture for the current dataset was EfficientNetB0. Second, an improperly performing constituent model can make the soft voting fusion based on unweighted voting problematic. This finding suggests the need for weighted voting, selective ensemble formation or model-exclusion strategies in the construction of future ensembles [20, 22, 29, 30].

4.3. Confusion-Matrix Analysis

The confusion matrices of VGG16, ResNet50, EfficientNetB0 and the ensemble model are displayed in Figure 4. The matrices give a class-wise overview of correct predictions and misclassifications among the classes of glioma, meningioma, no tumor or pituitary tumor.

For the meningioma and no-tumor classes, the VGG16 model is good, but it fails to be sensitive for glioma. There were many cases of glioma that were incorrectly identified as meningioma or not a tumour. This indicates that the VGG16 network was able to extract strong features for less ambiguous and more uniform examples of the classes that were visually more clear, but failed to learn strong features of glioma, which may exhibit more intra-class variation and less definite boundaries. VGG16's high precision, low recall for glioma, suggests that the model often correctly identifies glioma samples, but fails to correctly identify many glioma samples. A lot of misclassifications were observed for all four classes in the ResNet50 confusion matrix. It often misclassified tumor classes and exhibited poor discrimination between glioma and other classes. This is in line with the conclusion that ResNet50 failed to generalize well to the present dataset. This instability seen while training was thus echoed in the final test predictions. The most balanced class-wise behaviour was observed in the case of EfficientNetB0. It performed better than the other models in detecting gliomas and maintained good performance in other tumor types (meningioma, no tumor and pituitary tumor). Glioma recall was still quite low, but EfficientNetB0 mitigated some of the extreme misclassification patterns of the VGG16 and ResNet50. This is because the EfficientNetB0's overall test accuracy is the highest.

The ensemble confusion matrix demonstrated improvement or maintenance of performance for some classes, like meningioma, but not for glioma detection for soft voting. However, the introduction of ResNet50 led to an overall decrease in the ensemble's ability to classify glioma and pituitary tumor cases accurately. This validates that the final probability distribution was affected by the weak base learner, and the benefit of ensemble fusion was decreased. The confusion-matrix analysis will then give better evidence than just overall accuracy, since it shows where the ensemble fell short and why the best single model did better than the ensemble model.

The most important observation from a clinical point of view is the low recall value for glioma in all the models. Glioma is an important tumour type, and a misdiagnosis of glioma could have significant consequences in a decision support context. So, the accuracy of the model should not be judged only by its overall accuracy. When evaluating the clinical usefulness, class-wise recall should be taken into consideration, particularly for glioma.

4.4. ROC-AUC Analysis

The micro-averaged ROC curves of the models tested are shown in Figure 5. A one-vs-rest approach was adopted to calculate the ROC analysis to measure the capability of class ranking ability of each model. Although VGG16 had low recall for glioma, it gave a micro-averaged AUC of 0.862, which means that the class separability is reasonable. ResNet50 achieved the lowest AUC of 0.619, similar to its poor test accuracy and erratic training performance.

The highest micro-averaged AUC was obtained by EfficientNetB0, with an AUC of 0.886. This confirms the fact that EfficientNetB0 gave the best probability ranking among the models tested. The micro-averaged AUC value of the ensemble was 0.836. The ensemble of these models demonstrated improvement over ResNet50, but failed to outperform EfficientNetB0. This validates the poor probability estimates of ResNet50 that lower the advantage of equal weight probability averaging. Thus, the updated ROC analysis with the ensemble model shows that EfficientNetB0 is still the best model in terms of accuracy and probability-ranking behaviour.

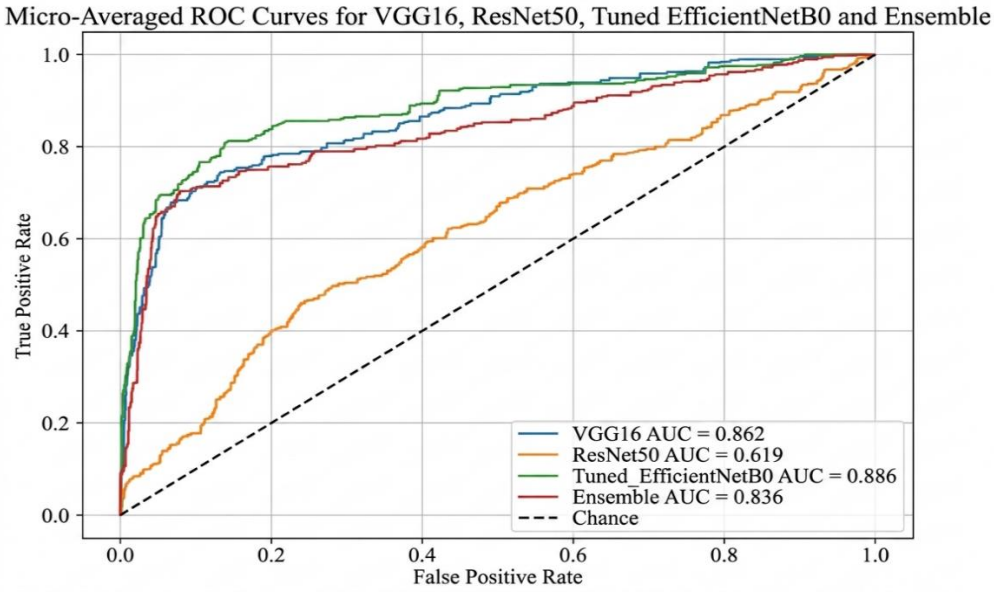

Fig. 5 Micro-averaged ROC curve for VGG16, ResNet50, EfficientNetB0, and the soft-voting ensemble on the independent test set

4.5. Precision, Recall, and F1-score by Class

The precision, recall and F1 scores for each class are presented in Table 4. Glioma was the hardest class for all models. For high gliomas, VGG16 had a high precision of 0.952 and a low recall of 0.200, meaning that many glioma cases were missed.

The recall of 0.110 and the F1-score of 0.188 were the lowest values achieved by ResNet50 for glioma. The efficientNetB0 model was able to recall gliomas with 0.190 and an F1-score of 0.299, whereas the ensemble model was able to recall gliomas with 0.180 and an F1-score of 0.295. As these values illustrate, the current system is still restricted clinically by sensitivity to glioma.

VGG16 and ensemble had a high recall value of 0.965 in the case of meningioma. The ensemble was able to achieve the highest meningioma F1 score of 0.813, which means that soft voting benefited this class. VGG16 had a recall of 0.981, EfficientNetB0 had a recall of 0.933, and the ensemble had the highest recall of 0.990 in the no-tumor class. In the case of pituitary tumor, EfficientNetB0 achieved the highest precision (0.889), recall (0.865), and F1-score (0.877).

The ensemble had the highest no-tumor recall and F1-score for meningioma, while EfficientNetB0 had the highest F1-score for pituitary tumor overall. All models, however, exhibited low glioma recall, which underscores the need for improved sensitivity of the current framework.

The class-wise analysis indicates that EfficientNetB0 has performed best in most of the classes. The ensemble achieved better performance for meningioma patients, but worse performance for glioma, no tumor, and pituitary tumor patients compared to EfficientNetB0. This validates the hypothesis that for some classes, ensemble fusion might help, but for others, it might hurt if one of the learners in the ensemble is weak. Therefore, future ensemble designs should include class-specific weighting, validation-based weighting or discard the models that do not perform well.

Table 4. Class-wise precision, recall, and F1-score for VGG16, ResNet50, EfficientNetB0, and the soft-voting ensemble on the independent test set

Class	VGG16			ResNet50			EfficientNetB0		
	Prec.	**Recall**	**F1**	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**
Glioma	0.952	0.200	0.331	0.647	0.110	0.188	0.704	0.190	0.299
Meningioma	0.627	0.965	0.760	0.485	0.287	0.361	0.659	0.939	0.774
No Tumor	0.696	0.981	0.814	0.369	0.876	0.520	0.748	0.933	0.831
Pituitary	0.917	0.595	0.721	0.367	0.297	0.328	0.889	0.865	0.877
Class	**Ensemble (soft voting)**			------			------		
	Prec.	**Recall**	**F1**	-	-	-	-	-	-
Glioma	0.818	0.180	0.295	-	-	-	-	-	-
Meningioma	0.703	0.965	0.813	-	-	-	-	-	-
No Tumor	0.642	0.990	0.779	-	-	-	-	-	-
Pituitary	0.885	0.622	0.730	-	-	-	-	-	-

4.6. Statistical Reliability of Model Comparison

Bootstrap-based 95 % confidence intervals were calculated, and paired statistical comparison of model predictions was performed to determine the statistical reliability. As all models had been tested on the same test set of 394 images, it was more appropriate to use paired tests rather than comparing absolute accuracy figures. Bootstrap resampling was used to estimate a 95% accuracy confidence interval and McNemar's exact test was used to compare paired correct/incorrect prediction outcomes.

Furthermore, the Wilcoxon signed-rank test was used to compare the confidence scores given by paired models pertaining to the correct class for the true-class probability scores. VGG16 achieved an accuracy of 70.56% with a 95% confidence interval of 65.99%-75.13%. ResNet50 achieved 40.10% accuracy with a 95% confidence interval of 35.28%-44.92%.

The highest accuracy of 73.35% was obtained by EfficientNetB0 with the 95% confidence interval of 69.04%-77.66%. The soft-voting ensemble scored 70.81% with 95% confidence interval 66.24%-75.38%. The results obtained from the paired McNemar test suggested ResNet50 performed significantly poorer than the ensemble and EfficientNetB0 and VGG16. However, the difference in accuracy between EfficientNetB0 and the ensemble was not significant by McNemar's test (p=0.121). McNemar's test for EfficientNetB0 vs. VGG16 had a non-significant p-value of 0.152. Compared to VGG16 and the ensemble, EfficientNetB0 had the highest accuracy, the benefit should be interpreted cautiously on the current test set.

Several model pairs had significant differences in probability levels as indicated by the Wilcoxon signed-rank test performed on true-class probability scores. However, since Wilcoxon estimates paired confidence distributions, not only the final decisions of correct/incorrect, but also these were considered as complementary evidence. In summary, statistical analysis shows that ResNet50 is statistically much weaker than the other two, and the difference between EfficientNetB0 and the soft-voting ensemble is not statistically significant with respect to the paired accuracy outcome.

Table 5. Paired statistical comparison of VGG16, ResNet50, EfficientNetB0, and the soft-voting ensemble

Comparison	McNemar p-value	Wilcoxon p-value	Interpretation
Ensemble vs VGG16	1.000	6.25e-27	The accuracy difference not significant by McNemar
Ensemble vs ResNet50	4.01e-33	1.24e-54	Ensemble significantly better than ResNet50
Ensemble vs EfficientNetB0	0.121	1.73e-39	The accuracy difference not significant by McNemar
EfficientNetB0 vs ResNet50	3.54e-30	4.50e-51	EfficientNetB0 is significantly better than ResNet50.
EfficientNetB0 vs VGG16	0.152	2.47e-02	The accuracy difference not significant by McNemar
VGG16 vs ResNet50	2.52e-30	3.94e-49	VGG16 is significantly better than ResNet50.

4.7. Grad-CAM Visual Explanations

The interpretability of the model predictions was assessed qualitatively using Grad-CAM visualizations. For this analysis, we chose EfficientNetB0, which obtained the best per-test accuracy, 73.35%, against VGG16's 70.56% and ResNet50's 40.10%. Figure 6 shows some representative correctly classified images of each class with respective Grad-CAM heat maps. The Grad-CAM heatmaps for glioma images were activated over the area of abnormal tissues in the brain region, but the activation was comparatively less and not as concentrated as in other tumor classes. This finding is consistent with the quantitative measures, as glioma had the lowest F1 score and recall. The less pronounced localization could be due to various factors, such as the heterogeneous nature of gliomas, indistinct borders or similar visual characteristics to other tumor groups. The images of meningioma demonstrated greater activation within peripheral or dural-based areas for the Grad-CAM heatmaps. This is anatomically possible, as meningiomas tend to grow in the meninges and are frequently close to the skull or on the surface of the dural membrane. The high recall and F1-score of meningioma were thus reinforced by activation patterns for this lesion type that were visually meaningful.

The heatmaps of Grad-CAM for pituitary tumor images revealed that the activated areas were concentrated around the sellar region. This aids the model's ability to learn pituitary tumor anatomical clues. These localised heat maps were reflected in a strong pituitary performance as shown by EfficientNetB0. No tumor images were associated with a clear focus of activation in Grad-CAM. Rather, the activations were low and widespread. This is an expected behavior since there is no tumor region visible in these kinds of images. Keep in mind, however, that these heat maps can be misleading in that the absence of focal activation is not enough to rule out disease. Based on the Grad-CAM visualization, one can conclude that the most successful models tended to be the ones that paid attention to the anatomically meaningful regions.

Grad-CAM should not be used for quantitative interpretability. Heatmaps can be used to determine whether the prediction is visually convincing, but do not fully capture the model's decision process, and can sometimes be convincing even when the prediction is wrong [17]. Hence, the results of Grad-CAM should be considered in context with quantitative metrics, class-wise results and clinical experts' review [15, 17, 31, 33].

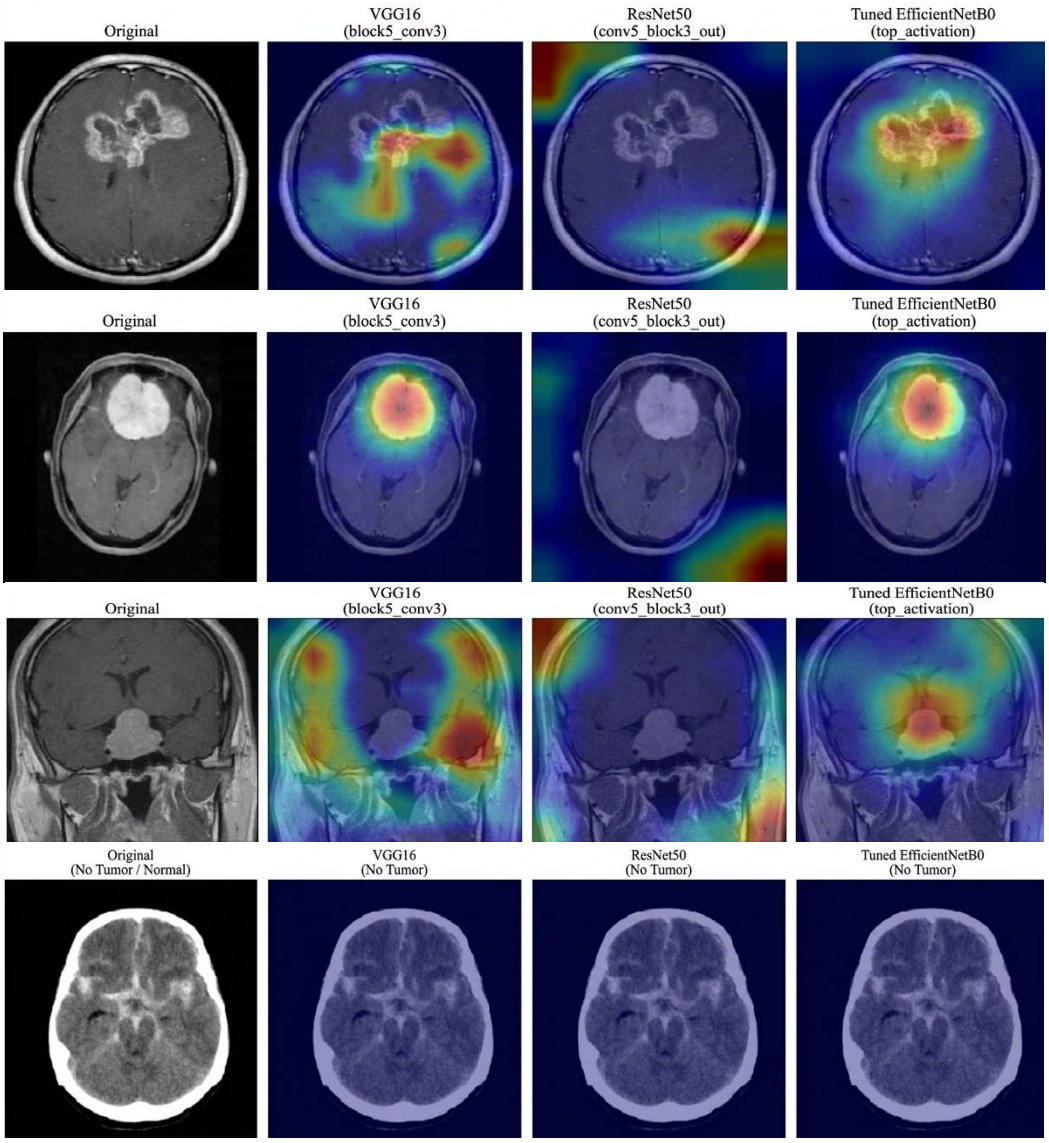

Fig. 6 Grad-CAM visualizations for EfficientNetB0 showing attention on a class specific-per 4 cases (glioma, meningioma, pituitary and no tumor)

4.8. Summary of Key Result Findings

The following are the four main findings of the experimental results. Firstly, the EfficientNetB0 model was the most promising one, with the best test accuracy, highest ROC-AUC and class-wise performance. Second, the VGG16 model had competitive accuracy and exhibited overfitting and a low rate of glioma recall. Third, the training behaviour of ResNet50 was unstable, and its test performance was weak with the selected configuration. First, the unweighted soft voting ensemble did not outperform EfficientNetB0, as the weak ResNet50 model had a negative impact on the averaged probability distribution. The biggest clinical drawback was the low glioma recall for all models. This means that there is still room for improvement before the model can be deemed to be suitable for use in clinical decision-making.

There are potential improvements, including class-specific losses, ensemble-based fusion and weighting, exclusion of weak learners, glioma-specific augmentation, higher resolution input, multi-sequence MRI, three-dimensional modelling and external validation on independent clinical datasets.

5. Discussion

Discussion of experimental outcomes includes model behavior, ensemble effectiveness, literature assessment, clinical significance, practical application, limits, and future research. Since reliability and interpretability are essential for medical image analysis systems, failure patterns are explored along with classification performance.

5.1. Key Findings

The experiment's findings show that transfer learning may help classify moderately-sized multi-class MR images of brain tumors. The performance was found to be very different in the architectures evaluated, however. The top three test accuracy values were achieved by the EfficientNetB0 (73.35%), soft-voting ensemble (70.81%) and VGG16 (70.56%), respectively. The chosen ResNet50 architecture was not very transferable to this MRI classification task and yielded a much lower accuracy of 40.10%.

The efficientNetB0 architecture also exhibited the most consistent training behaviours and the best overall balance of classes evaluated. This could be explained by its effective architecture, compound scaling and small parameter number compared to heavier CNN models [4, 8].

The model seemed to be a better balance between the representation of features and generalization, especially in meningioma, pituitary tumor and no tumor images. VGG16 was competitive overall, but its training curves suggested that it was overfitting and class-wise, its performance was poor, particularly in recalling gliomas. ResNet50 performed an unstable validation curve and poor test results, indicating that deeper residual networks might need to be carefully fine-tuned, have more data or be regularized differently for this particular application.

The most important result of the study is that the best individual model did not prove to be superior to the soft voting ensemble. The accuracy of the ensemble was determined as 70.81%, which is slightly above the VGG16 but below the EfficientNetB0. The result is particularly significant since it is commonly believed that ensemble learning leads to better robustness, whereas the present results indicate that, in the case of an under-performing base learner, it can be detrimental to robustness to rely on simple unweighted probability averaging. As ResNet50 was equally contributing to the ensemble, even though it was not very accurate and its predictions were unstable, the probability outputs of ResNet50 reduced the effectiveness of the combined model. This corroborates previous findings that the quality, calibration, and diversity of the individual models are important factors for ensemble performance [20, 22, 29, 30].

These are the consistently low glioma recalls across all models, the most clinically significant. Recall values were between 0.110 and 0.200, with ResNet50 being the lowest and VGG16 being the highest. EfficientNetB0 was 0.190, and the ensemble was 0.180. This means that under the current conditions, glioma is the most challenging class to detect. Low glioma recall is especially significant, because gliomas are clinically relevant tumors and can be clinically very heterogeneous regarding their shape, texture, boundary definition, enhancement pattern and anatomical spread. However, a model that has an acceptable overall accuracy may be clinically limited because of the lack of representation of many glioma cases. Hence, it is crucial that the sensitivity of the classes is taken into account, as well as the overall accuracy in medical image classification, and class-wise sensitivity.

Qualitative support from the Grad-CAM analysis was provided for the quantitative findings. In the meningioma and pituitary tumor cases, the activation maps tended to center on anatomically meaningful areas (periphery dural-based areas and sellar area, respectively). Glioma cases had less and weaker activation consistent with their low glioma recall. These observations suggest that the model picked up some clinically relevant spatial cues but that the sensitivity for glioma is poor, which means that richer contextual cues may be needed. Thus, the results indicate that Grad-CAM can be used as a complementary interpretability technique, but it should not be used to replace clinical validation [17].

5.2. Comparison with Recent Literature

Recent ensemble-based brain tumour MRI studies have reported their performance using different datasets and evaluation protocols. An accuracy of 98.70% was reported by Tummala et al. [12] with the ensemble of Vision Transformer models. Roy et al. [14] achieved a mean 10-fold cross-validation accuracy of 97.90% and 98.15% accuracy with the DeepEFE-SVM model. Talukder et al. [22] obtained 99.78% accuracy with the Genetic Algorithm-based Weight Optimization and 99.84% accuracy with the Grid Search-based Weight Optimization. On two public datasets, Almuhaimeed et al. [27] achieved 99.54% and 98.90% accuracy, respectively. Multiclass classification reached an accuracy of 99.69% by Velpula et al. [29] and 98.00% by Anand et al. [30] with a sensitivity of 98.70%, precision of 98.25%, and an F1 score of 98.50%.

Although the reported accuracy values range from 98.00% to 99.84%, they should not be interpreted as directly comparable with the 73.35% accuracy of EfficientNetB0 and the 70.81% accuracy of the ensemble in the present study because the corresponding studies differ in dataset composition, number of classes, augmentation procedures, validation methods, and patient-wise separation.

The reported performance can be significantly influenced by differences in the source of the datasets, the number of classes, the image processing, augmentation, the splitting

strategy, the validation design, and the evaluation protocol. Some studies randomly split the images, perform cross-validation without patient-level separation, or combine data from multiple repositories. The training and testing sets may contain slices from the same patient or image, leaking data and possibly enhancing performance [5, 13, 14].

In contrast, the current study used a fixed patient-wise divide and only the independent test set for final evaluation. This more rigorous evaluation design might partially account for the lower, but more conservative, estimate of performance.

The other reason for the difference from the high-performance studies is the use of the no-tumor class and the setting of four classes. Some studies only consider the classification of tumour subtype or fewer classes, which may simplify the classification task. During the current study, the model had to discriminate between glioma, meningioma, pituitary tumor and no tumor seen. The difficulty of this setting is increased, since not only must the tumor be detected, but it must be discriminated between tumor types in the same system.Other findings include that general accuracy may be accompanied by clinically important deficiencies. There are some studies that focus on the aggregate accuracy without discussing the performance of classes in terms of their recall or misclassification patterns. Even if a model has reasonable accuracy, the detection of gliomas can be weak, as demonstrated by the present study. This is similar to the work to resolve the problem of the minority class or difficult class from underperformance by balancing classes, learning with imbalance issues or synthetic augmentation [14, 22, 27]. Thus, the present results highlight the importance of reporting class-wise precision, recall, F1-score, as well as clinical error patterns and confusion matrices, in addition to overall accuracy.

This reduced performance might also be due to the input resolution, which was 224 × 224 corners in the 2D T1-weighted MRIs. Increased resolution images, multi-sequence MRI, or volumetric modeling might be useful for the detection of gliomas, which tend to be diffuse, show edema, have heterogeneous enhancement and irregular borders. Other MRI sequences like T2 and FLAIR may be helpful in determining the extent of the tumour and the extent of the edema, and may not be appreciated in a single T1-weighted slice [13]. Thus, the current findings should be viewed as a minimum but clear-cut two-dimensional MRI classification experiment of transfer-learning models.

Table 6. Quantitative summary of previous ensemble-based brain tumour MRI classification studies and the proposed work under their respective experimental protocols

"Study	Methodology	Dataset/Classes	Explainability	Quantitative Performance	Key finding/ positioning
Tummala et al. [12]	Vision transformer ensembling	Brain tumor MRI classification	Not the primary focus	Accuracy: 98.70%	A three-class brain tumour classification problem was tested using an ensemble of Vision Transformer. The present patient-wise 4-class study yielded a different result compared to the one obtained from a different set of data and evaluation protocol.
Roy et al. [14]	Explainable ensemble with data balancing	Brain tumor MRI classification	Grad-CAM included	Accuracy: 98.15%; mean 10-fold cross-validation accuracy: 97.90%	Dual-GAN-based class balancing, DeepEFE feature extraction with SVM classification and Grad-CAM interpretation were used. The cross-validation protocol is not a patient-wise test protocol as is used in the present study.
Talukder et al. [22]	Deep ensemble with data balancing and fine-tuning	Brain tumor MRI classification	Limited explainability focus	GAWO accuracy: 99.78%; GSWO mean five-fold accuracy: 99.84%	The values reported are not identical to the present equal-weight patient-wise evaluation, as the ensemble was optimized using weight-optimization methods and tested using five-fold cross-validation.
Almuhaimeed et al. [27]	GAN-augmented data with	Public brain tumor MRI datasets	Limited explainability focus	Accuracy: 99.54% on Dataset I and	The proposed AE-cGAN-based synthetic augmentation and Swin Transformer

	autoencoders and Swin Transformers			98.90% on Dataset II	classification were tested on two public datasets. There are differences between synthetic augmentation and the current study in terms of dataset composition and augmentation.
Velpula et al. [29]	CNN feature extraction with an ensemble voting classifier	Brain tumor MRI classification	Not the primary focus	Multiclass accuracy: 99.69%; binary accuracy: 100%	Multi-class and binary MRI datasets were used to evaluate ensemble voting. Unlike the fixed four-class patient-wise protocol used in the present work, the evaluation settings used in the binary and multiclass settings are different.
Anand et al. [30]	Weighted average ensemble using CNN/VGG features	Glioma MRI classification	Not the primary focus	Accuracy: 98.00%; sensitivity: 98.70%; precision: 98.25%; F1-score: 98.50%	A binary classification problem for brain tumour classification was evaluated using a weighted-average ensemble. The binary task is not directly comparable with the four-class classification task of the present study.
Proposed work	Equal-weight soft voting of VGG16, ResNet50, and EfficientNetB0	Four-class MRI: glioma, meningioma, no tumor, pituitary	Grad-CAM included	EfficientNetB0: 73.35%; ensemble: 70.81%; VGG16: 70.56%; ResNet50: 40.10%	In the patient-wise four-class evaluation, EfficientNetB0 demonstrated the highest accuracy of 73.35%, and the equal weighted ensemble had 70.81% accuracy. The analysis quantitatively proves that adding the weak ResNet50 learner lessens the gain of equal-weight probability fusion. No greater accuracy is claimed than that of previous studies.

Previous studies report accuracy values between 98.00% and 99.84%, while in the current study, the accuracy was 73.35% for EfficientNetB0 and 70.81% for the equal weight ensemble.

These findings do not directly imply any performance comparisons since the studies have varying compositions of datasets, different numbers of classes, synthetic augmentation, cross-validation procedures, and patient-wise separation. Hence, the present work does not promise to achieve greater accuracy but rather a patient-wise and class-sensitive quantitative analysis to show that equal weight ensemble fusion may not enhance the performance if an inferior constituent learner is added on.

5.3. Clinical and Practical Considerations

The most important requirement for practical clinical use is not only high overall accuracy but also that of clinically important tumor classes. The low recall for gliomas in the present study does not promote immediate clinical applicability. In a decision support environment, the failure to detect cases of glioma can have serious consequences, as gliomas often require timely diagnosis, treatment planning and follow-up. So, for any model to be practically used, it should be sensitive to high-risk tumor classes, which would also result in an increase in false positive results.

There are several strategies that may be useful to increase the sensitivity of gliomas. One way to boost the recall for the glioma class is by decreasing the decision threshold for that class. This could, however, lead to overly optimistic predictions of gliomas and should be used with caution. Imbalance-aware training techniques, such as focal loss, can also improve the performance of hard or underrepresented classes [14, 22]. Oversampling, class-specific augmentation, or synthetic image generation using a generative adversarial network can further aid the model to learn more patterns of gliomas [14]. Furthermore, localization-guided training with

tumor masks, bounding boxes, or attention constraints can be employed so that the network will focus on tumor-relevant areas rather than on background anatomy.

The ensemble results are also beneficial. The results show the impact of this unweighted soft voting, which will cause the performance to decrease if the weak model is included in the ensemble. Do not include a model in an ensemble in a clinical system simply for the reason that it is a variant of the model. Instead, the accuracy measure of each constituent model, along with class-wise recall, calibration, and error diversity measure, should be measured. However, averaging can be suboptimal and other methods like weighted voting or selective ensemble formation can be used to achieve better performance, particularly when one of the models is consistently worse [20, 22, 29, 30].

Another key factor to ensure clinical acceptance is explainability. Grad-CAM visualizations can verify that the model covers anatomically relevant areas. This can increase transparency and may be helpful in knowing if any suspicious behaviour of the model has been seen, such as if the model focused on any background areas or non-tumorous structures [17]. However, there are some important precautions that need to be taken when using Grad-CAM. The heatmap does not reflect the clinical accuracy of the prediction, and visually plausible activation can still be seen in incorrect predictions. Therefore, explainability should be included via expert review, quantitative validation and external testing.

The proposed system should be regarded as an assistive system instead of a replacement system for the radiologist. In a realistic workflow, this model could be deployed as a second-reader system to highlight suspicious MRI slices, facilitate prioritisation, or as a preliminary review system. Final diagnosis should be made by qualified clinical experts. The model would need to be validated on independent multi-center datasets, be scanner robust, be evaluated across different MRI acquisition protocols and be evaluated in a prospective clinical study prior to deployment. Practical application would also have to consider regulatory, ethical and medico-legal issues.

5.4. Limitations

There are several limitations of the study. The data set used was collected from public repositories and may not capture all of the variation in the clinical environment. Images in publicly available curated datasets may be relatively well labelled, and clinical MRI data can include motion artefacts, scanner variations, incomplete sequences, varying contrast quality, post-operative changes, and unusual tumour presentations. Hence, the conclusions with the results are not validated against the clinical data from hospitals.

Secondly, an analysis was done on a two-dimensional slice level. Full three-dimensional MRI volumes and multiple sequences are generally used to make a clinical diagnosis. A single 2D slice does not necessarily show the full spatial dimensions, infiltrative pattern, oedema or volumetric features of the tumour. This limitation is especially relevant in cases of glioma (brain tumor), which have indistinct tumor margins and where contextual information from multiple slices is more appropriate for classification. Slice-aggregation strategies, recurrent volume-level modeling or segmentation-based pipelines might be more clinically realistic, such as three-dimensional CNNs [5, 13].

Third, the study employed contrast-enhanced T1-weighted sections, and multi-sequence MRI data might generate more diagnostic information. Edema and infiltrative patterns can be seen on T2-weighted and FLAIR sequences and are especially important for the evaluation of gliomas [13]. Perhaps the lack of multi-modal imaging in the study accounts for the low glioma recall. There was no external validation performed. The models were tested on a held-out test set of the available data, but not independently tested on a different data set acquired at another institution, scanner or acquisition protocol. In the absence of validation by others, the model's robustness in the context of a shift in domains cannot be justified [22].

Fourth, Grad-CAM only offers post-hoc visual interpretation. Produces coarse scale heatmaps and does not explain the reasoning behind the model in a complete manner. However, the interpretation of the heatmaps can also be subjective and dependent on the model layer and preprocessing strategy chosen. Thus, the results of the Grad-CAM should not be used as definitive clinical evidence [17].

Fifth, when using an ensemble method, an equal-weight probability average method was used. This made it quite easy to check the transparency of this ensemble and reproduce it, and also opened the final prediction to weak learners. The ensemble would have been more effective if ResNet50 had performed well. More sophisticated approaches like weighted voting, class-wise weighting, selective model exclusion, stacking or fusion based on a calibration function can be used to lower this limitation [20, 22, 29, 30].

Sixth, it was not possible to conduct systematic and exhaustive hyperparameter optimization due to computational limitations. A limited validation search was performed, but further fine-tuning of learning rates, fine-tuning depth, dropout, loss function, and augmentation may give better performance. Last, radiologist-based validation of the heat maps generated by the Grad-CAM was not performed, which would be helpful to determine whether the highlighted regions are clinically significant.

5.5. Future Work

In the future, work should be directed towards further enhancing the sensitivity of gliomas and towards further

strengthening the clinical robustness. Given that glioma was the most challenging class in the current study, future models should focus on the class-wise recall instead of just the overall accuracy. To minimize the number of falsely negative predictions of glioma, possibly imbalance-aware loss functions like focal loss, class-weighted cross-entropy, and sensitivity-oriented threshold calibration can be used.

The use of multi-sequence MRI data should also be explored. The combination of T1, Contrast-Enhanced (CE) T1, T2 and FLAIR sequences might be able to offer more anatomical and pathological information than any single sequence. This could be especially valuable in the detection of glioma since the edema, infiltration, necrosis and enhancement patterns may appear in distinctly different MRI sequences [13].

Future research can also be directed towards prediction at the patient level or volume level rather than the slice level. 3D CNN, volumetric transformer, recurrent aggregation of slices and segmentation guided classification may be more appropriate for modeling the spatial context. The use of tumor localization or segmentation masks could be beneficial to the classifier in directing it towards neoplastic regions of interest and preventing background structures from being misclassified.

The ensemble approach should also be enhanced. In the future, validation-based weighted voting, class-specific weighting, selective exclusion of not-so-good models, stacking, and meta-learning-based fusion can be used rather than equal-weight soft voting. Prior to incorporating into an ensemble, each base learner should be assessed for class-wise performance, calibration and error diversity.

This can avoid reducing the ensemble accuracy due to the weak models. In future, it is necessary to incorporate Grad-CAM, Grad-CAM++ with occlusion sensitivity and LIME, SHAP and attention-based visualization to make the explainability. Expert radiologists should also analyze the heatmaps to determine if the areas that are highlighted are clinically relevant [15, 16, 31]. If available, masks that contain tumors may be added to quantitative localization measures.

Future research should be tested for external validation. Models need to be validated on independent data sets from other institutions, scanners, MRI protocols and patient populations. The validation will help to decide whether the model is able to extrapolate beyond the data set on which it was developed. There should also be a clinical evaluation done prior to any practical deployment.

Lastly, future research should consider issues of reproducibility, such as the release of the data preprocessing pipeline, train-validation-test split information, model configurations, and evaluation scripts, where allowed by the dataset license. This would enable other researchers to repeat the results and compare their techniques under the same experimental conditions.

6. Conclusion

In this paper, a transfer learning method is presented to classify brain tumor MRI into four classes using VGG16, ResNet50, EfficientNetB0 and probabilistic soft-voting ensemble. The four classes were: glioma, meningioma, pituitary tumor, and no tumor.

The classes were glioma, meningioma, pituitary tumor, and no apparent tumor. Conventional performance evaluation was combined with visual explanations using Grad-CAM to assess if the model predictions were supported by anatomically meaningful image regions.

The best individual model is the fine-tuned EfficientNetB0, which has an overall test accuracy of 73.35%. VGG16 also had a competitive accuracy of 70.56%, but it demonstrated a tendency to overfit during training. ResNet50 is not well-generalized under the training configuration selected, and obtained an accuracy of 40.10%.

The soft-voting ensemble had 70.81%, which is higher than ResNet50, marginally higher than VGG16, but lower than EfficientNetB0. The result indicates that the simple unweighted ensemble fusion do not necessarily lead to better classification accuracy, particularly when a constituent learner is not accurate. The results of the class-wise analysis revealed that the tumour class glioma was the most challenging to identify. The model with the best performance for gliomas was EfficientNetB0, but the recall was still not very high.

This is clinically relevant as undiagnosed gliomas could diminish the clinical utility of an automated decision support system. Thus, it suggests that future systems should consider class-wise sensitivity, especially for clinically important tumor classes, rather than overall accuracy. The visualizations of the Grad-CAM were found to be helpful for qualitative information on the model behaviour.

Generally, the heatmaps displayed anatomically plausible patterns of activation for the meningioma and pituitary tumor cases, with reduced and less focal activation in the glioma cases. These results validate the value of explainable AI for a visual inspection of the model attention. However, Grad-CAM is considered an assistive interpretability tool, and not as a clinical validation.Finally, this study shows that EfficientNetB0 is a relatively stable and effective transfer learning architecture for the mentioned brain MRI classification problem. Concurrently, the results highlight some key drawbacks of unweighted ensemble fusion, slice-level classification, and limited glioma sensitivity. The proposed approach should thus be considered as an assistive

research system at this point and not as a clinically deployable diagnostic system. Several additional steps would be needed for practical deployment, including external validation on independent multi-center data, using full volumetric and multi-sequence MRI data to evaluate, reviewing explainability outputs with radiologists, and better strategies to reduce false-negative glioma predictions.

Future research directions involve using weighted ensembles, selective ensembles, augmentation for gliomas, segmentation-aware learning, higher resolution inputs, 3D modelling and external clinical validation. It is hoped that these enhancements will further contribute to the robustness, interpretability, and clinical reliability of brain tumor classification in MRI images.

References

[1] Quinn T. Ostrom et al., "CBTRUS Statistical Report: Primary Brain and other Central Nervous System Tumors Diagnosed in the United States in 2012–2016," *Neuro-Oncology*, vol. 21, no. 5, pp. 1-100, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[2] Monika Agarwal et al., "Deep Learning for Enhanced Brain Tumor Detection and Classification," *Results in Engineering*, vol. 22, pp. 1-12, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Geert Litjens et al., "A Survey on Deep Learning in Medical Image Analysis," *Medical Image Analysis*, vol. 42, pp. 60-88, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[4] Md. Manowarul Islam et al., "BrainNet: Precision Brain Tumor Classification with Optimized Efficientnet Architecture," *International Journal of Intelligent Systems*, vol. 2024, no. 1, pp. 1-24, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] Md. Alamin Talukder et al., "An Efficient Deep Learning Model to Categorize Brain Tumor using Reconstruction and Fine-Tuning," *Expert Systems with Applications*, vol. 230, pp. 1-16, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Sandeep Kumar Mathivanan et al., "Employing Deep Learning and Transfer Learning for Accurate Brain Tumor Detection," *Scientific Reports*, vol. 14, no. 1, pp. 1-15, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Yuki Wong et al., "Brain Tumor Classification using MRI Images and Deep Learning Techniques," *PLOS One*, vol. 20, no. 5, pp. 1-19, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[8] Mingxing Tan, and Quoc Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, pp. 6105-6114, 2019. [Google Scholar] [Publisher Link]

[9] Kaiming He et al., "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016. [Google Scholar] [Publisher Link]

[10] Gao Huang et al., "Densely Connected Convolutional Networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 2261-2269, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[11] Alexey Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," *arXiv preprint*, pp. 1-22, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Sudhakar Tummala et al., "Classification of Brain Tumor using Vision Transformers Ensembling," *Current Oncology*, vol. 29, no. 10, pp. 7498-7511, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] Bjoern H. Menze et al., "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993-2024, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[14] Priyanka Roy, Fahim Mohammad Sadique Srijon, and Pankaj Bhowmik, "An Explainable Ensemble Approach for Advanced Brain Tumor Classification Applying Dual-GAN Mechanism and Feature Extraction Techniques Over Highly Imbalanced Data," *PLOS One*, vol. 19, no. 9, pp. 1-41, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Ramprasaath R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618-626, 2017. [Google Scholar] [Publisher Link]

[16] Aditya Chattopadhay et al., "Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks," *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, NV, USA, pp. 839-847, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[17] Wandile Nhlapho et al., "Bridging the Gap: Exploring Interpretability in Deep Learning Models for Brain Tumor Detection and Diagnosis from MRI Images," *Information*, vol. 15, no. 4, pp. 1-21, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[18] S. Deepak, and P.M. Ameer, "Brain Tumor Classification using Deep CNN Features via Transfer Learning," *Computers in Biology and Medicine*, vol. 111, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[19] Thomas G. Dietterich, "Ensemble Methods in Machine Learning," *Multiple Classifier Systems*, Springer, Berlin, Heidelberg, vol. 1857, pp. 1-15, 2000. [CrossRef] [Google Scholar] [Publisher Link]

[20] Ludmila I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, Wiley, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[21] Shaimaa E. Nassar et al., "A Robust MRI-based Brain Tumor Classification Via a Hybrid Deep Learning Technique," *Journal of Supercomputing*, vol. 80, no. 2, pp. 2403-2427, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[22] Md. Alamin Talukder et al., "A Deep Ensemble Learning Framework for Brain Tumor Classification using Data Balancing and Fine-Tuning," *Scientific Reports*, vol. 15, no. 1, pp. 1-23, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[23] Daniel Reyesa, and Javier Sánchezb, "Performance of Convolutional Neural Networks for the Classification of Brain Tumors using Magnetic Resonance Imaging," *Heliyon*, vol. 10, no. 3, pp. 1-22, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[24] Muhammad Sajjad et al., "Multi-Grade Brain Tumor Classification using Deep CNN with Extensive Data Augmentation," *Journal of Computational Science*, vol. 30, pp. 174-182, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[25] Jun Cheng et al., "Enhanced Performance of Brain Tumor Classification via Tumor Region Augmentation and Partition," *PLOS One*, vol. 10, no. 10, pp. 1-13, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[26] K. Afnaan et al., "Boosting Brain Tumor Detection with an Optimized ResNet and Explainability Via Grad-CAM and LIME" *Brain Informatics*, vol. 12, no. 1, pp. 1-19, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[27] Abdullah Almuhaimeed et al., "Brain Tumor Classification using GAN-Augmented Data with Autoencoders and Swin Transformers," *Frontiers in Medicine*, vol. 12, pp. 1-26, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[28] Amjad Rehman et al., "Microscopic Brain Tumor Detection and Classification using 3D CNN and Feature Selection Architecture," *Microscopy Research and Technique*, vol. 84, no. 1, pp. 133-149, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[29] Vijaya Kumar Velpula et al., "Enhanced Brain Tumor Classification using Convolutional Neural Networks and Ensemble Voting Classifier for Improved Diagnostic Accuracy," *Computers and Electrical Engineering*, vol. 123, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[30] Vatsala Anand et al., "Weighted Average Ensemble Deep Learning Model for Stratification of Brain Tumor in MRI Images," *Diagnostics*, vol. 13, no. 7, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[31] Krishan Kumar, and Kiran Jyoti, "Recent Advancements in Grad-CAM and variants: Enhancing Brain Tumor Detection, Segmentation, and Classification," *Systematic Review*, pp. 1-18, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[32] Ibrar Babar et al., "Unifying Genetics and Imaging: MRI-based Classification of MGMT Genetic Subtypes using Visual Transformers," *2023 18th International Conference on Emerging Technologies (ICET)*, Peshawar, Pakistan, pp. 122-128, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[33] Rezuana Haque et al., "NeuroNet19: An Explainable Deep Neural Network Model for the Classification of Brain Tumors using Magnetic Resonance Imaging Data," *Scientific Reports*, vol. 14, no. 1, pp. 1-22, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[34] Wei Tan, Yifeng Geng, and Xuansong Xie, "FMViT: A Multiple-Frequency Mixing Vision Transformer," *27th European Conference on Artificial Intelligence*, Santiago de Compostela, Spain, pp. 1-4857, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[35] Jia Deng et al., "ImageNet: A Large-Scale Hierarchical Image Database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, pp. 248-255, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[36] Karen Simonyan, and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arxiv*, pp. 1-14, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[37] Hafiz Muhammad Tayyab Khushi et al., "Improved Multiclass Brain Tumor Detection via Customized Pretrained EfficientNetB7 Model," *IEEE Access*, vol. 11, pp. 117210-117230, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[38] Muthulakshmi K., and Jayalakshmi M, "Rethinking Model of Efficientnet-B9 for Brain Tumor Classification: A High-Precision Deep Learning Approach," *Results in Engineering*, vol. 28, pp. 1-14, 2025. [CrossRef] [Google Scholar] [Publisher Link]