I.J. Image, Graphics and Signal Processing, **2026, 3, 167-188**
Published Online on June 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2026.03.09

Comparative Analysis and Ensemble Optimization of CNN Architectures for MRI-Based Brain Tumor Diagnosis

Md Tariqul Islam*
Department of Mathematical and Physical Sciences, East West University, Dhaka, 1212, Bangladesh
E-mail: mti.tariqul12@gmail.com
ORCID iD: https://orcid.org/0009-0008-0785-1234
*Corresponding author

Pintu Chandra Shill
Department of Mathematical and Physical Sciences, East West University, Dhaka, 1212, Bangladesh
E-mail: pintu.shill@ewubd.edu
ORCID iD: https://orcid.org/0000-0002-6395-9281

Md Sadiq Iqbal
Department of Computer Science and Engineering, Bangladesh University, Dhaka, 1207, Bangladesh
E-mail: sadiq.iqbal@bu.edu.bd
ORCID iD: https://orcid.org/0000-0001-9219-230X

Received: March 9, 2026; Revised: April 13, 2026; Accepted: May 2, 2026; Published: June 8, 2026

Abstract: Brain tumor detection and classification from MRI images is a challenging task. Early and accurate diagnosis are essential for selecting appropriate treatment plans and improving patient outcomes. Despite significant advances in deep learning for medical image recognition, comprehensive comparative analyses of brain tumor classification models, particularly regarding ensemble optimization, remain limited. This paper uses four state-of-the-art deep learning frameworks, namely EfficientNetB4, MobileNetV3, MobileNetV2, and EfficientNetB0, to classify brain MRI images into four categories: Glioma, Meningioma, Pituitary tumor, and Normal. It employs a two-phase transfer learning approach, followed by 5-fold cross-validation on 875 MRI images. A unified experimental framework is employed, incorporating a two-phase transfer learning approach, consistent preprocessing, and a rigorous evaluation protocol with 5-fold cross-validation and an independent test set to prevent data leakage. Both full and selective ensemble strategies are examined to improve the robustness and stability. The models are evaluated using accuracy, precision, recall, F-1 score, confusion matrices, and accuracy curves, and statistical validation using McNemar's test. MobileNetV3 achieves the highest test accuracy of 98.76%, followed by EfficientNetB4 (97.89%) and EfficientNetB0 (93.48%). MobileNetV2 performs significantly worse, with an accuracy of less than 80%. The selective ensemble technique (which uses the best models) attains the highest accuracy of 92.97%, compared to the full ensemble (84.40%), which improves prediction robustness but does not surpass the best individual model in peak accuracy. Overall, it can be concluded that MobileNetV3 is the most suitable architecture for brain tumor classification, delivering high accuracy with minimal computational complexity. The selective ensemble approach also enhances performance, maintaining computational efficiency, emphasizing the importance of informed model selection in neuro-oncological image analysis and clinical decision-support systems.

Index Terms: MRI imaging, Machine Learning, Ensemble Learning, Brain Tumor, MobileNetV3

1. Introduction

Brain tumors are among the most complex and debilitating neurological conditions, and patient survival and quality of life largely depend on early and accurate diagnosis. An annual report from the Central Brain Tumor Registry of the United States (CBTRUS) states that between 84,000 and 100,000 primary brain tumors are diagnosed each year in the U.S., with nearly 30 percent being malignant [1]. The incidence of brain tumors continues to increase worldwide,

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

challenging healthcare systems everywhere. Proper classification and grading are crucial for effective clinical management. Gliomas are tumors that arise from glial cells, accounting for nearly thirty percent of all brain tumors and almost 80 percent of malignant cases. Meningiomas [2], which develop from the meninges, are generally not harmful; however, their size can lead to severe neurological issues. Pituitary tumors are also benign and, depending on their position, often cause hormonal imbalances and vision problems. It is important to differentiate these tumors from normal brain tissue to determine the best treatment options, such as surgical, radiotherapy, or chemotherapy interventions [3]. Despite the fact that MRI offers comprehensive anatomical details, manual analysis is laborious and prone to individual-observer variation. The creation of dependable and automated computer-aided diagnostic systems is driven by this constraint.

Magnetic resonance imaging (MRI) is the most crucial diagnostic tool for identifying and classifying brain tumors because it offers superior soft-tissue contrast [4] and can generate images in multiple planes. Common MRI sequences, such as T1-weighted, T2-weighted, and contrast-enhanced T1-weighted, provide complementary information and facilitate detailed tumor analysis [5,6,7]. However, manual interpretation is laborious, subject to inter-observer variability, and demands significant clinical expertise. These challenges have increased the need for automated decision-support systems that help radiologists perform faster, more consistent tumor classification, especially as MRI protocols become more complex and imaging data continue to grow. The rise of deep learning has profoundly transformed medical image analysis; convolutional neural networks (CNNs) have demonstrated remarkable results in tasks like classification, segmentation, and detection [8,9,10]. CNNs automatically learn hierarchical feature representations directly from raw image data and eliminate the need for manual feature engineering, which had been a limitation in traditional computer-aided diagnosis systems.

Recent advances in transfer learning have increased the popularity of this technique in medical imaging, where models trained on large natural image datasets like ImageNet are adapted for small medical datasets [11]. In clinical settings with limited annotated data and high data acquisition costs, this approach is especially effective. Despite this progress, figuring out how to fine-tune and adapt pretrained models for different medical imaging tasks remains an ongoing area of research. Although deep learning is increasingly used in medical imaging, many challenges still exist. Medical images are significantly different from natural images in acquisition methods, noise levels, and clinical importance [12,13,14]. Additionally, clinical environments demand exceptionally high levels of accuracy and reliability compared to most standard computer vision problems. Deep learning models are often viewed as black boxes, and their lack of interpretability, transparency, and validation raises concerns about their appropriateness for real-world clinical use.

Deep learning-based brain tumor classification has made significant progress, but challenges remain. Fair comparison is difficult because many current studies focus on individual model performance without ensuring consistent training and evaluation settings. Additionally, despite the widespread use of ensemble learning, most methods combine all available models without accounting for the detrimental effects of weaker models. Statistical validation and computational efficiency—both crucial for practical clinical implementation—have received little attention.

These are the main contributions of this work, summarized as follows:

- The study offers a systematic and clinically biased evaluation of the state-of-the-art convolutional neural network architectures for MRI image-based brain tumor classification.
- To enable a fair comparison of different transfer-learning models using the same preprocessing, training, and validation parameters, this work develops a single experimental framework.
- Additionally, a selective ensemble strategy is introduced to concentrate on the top-performing models and improve their resilience without sacrificing computational efficiency.

The statistical validation techniques are used to boost the reliability of the findings. The work can offer useful experience in model selection, ensemble design, and performance-efficiency trade-offs in clinical decision-support systems, even though a new architecture is not recommended. The gap between experimental deep learning research and real-world clinical application is filled by this rigorous methodology.

The rest of the paper is organized as follows. Section 2 reviews related research on classification and ensemble learning methods for brain tumors. Section 3 explains the proposed methods and approach, including data collection and curation, model design, training settings, and evaluation criteria. Section 4 shows the experimental results for each individual model and the ensemble approaches. Section 5 concludes the paper and suggests future research directions.

2. Related Study

Over the past decade, deep learning techniques for classifying brain tumors have made significant advances. Earlier methods relied on manually designed CNN architectures trained through gradient-based optimization. For instance, Tahia Tazin et al. [15] developed a deep CNN that achieved 92% accuracy on a brain tumor dataset using X-ray images. However, these early models often overfitted due to limited training data and needed extensive data

augmentation [16] to improve their generalization. One of the most notable improvements was the adoption of transfer learning. Deependra Rastogi et al. [17] used VGG19, Xception, MobileNetV2, and ResNetV2, pre-trained on natural image repositories, which resulted in 96.11% accuracy on a brain tumor dataset using MRI images, highlighting the usefulness of pre-trained weights from large-scale natural image datasets. Similarly, another study by Mirza Mumtaz Zahoor et al. [18] introduced a Res-BRNet CNN variant that achieved the highest accuracy of 98.22% on the same MRI image dataset, showing that architectural improvements can further enhance medical image classification. Recent research has concentrated on improved training schemes and architectures. Sunanda Das et al. [19] conducted a comparative study of several CNN models for brain tumor detection and found that GoogleNet was the most accurate (94.39%) in a four-class classification task. However, the limited cross-validation and independent testing in their study may have caused an overestimation of real-world performance.

The EfficientNet family has gained increasing attention in medical imaging research. EfficientNet, introduced by Tan and Le [20], is based on a compound scaling strategy that systematically balances network depth, width, and input resolution to enhance performance while optimizing computational efficiency. Variants of EfficientNet have been effectively applied in various medical imaging fields, including dermatology, radiology, and pathology. However, few studies [21,22,23] offer comprehensive cross-comparisons between EfficientNet and other modern architectures for brain tumor classification. Ensemble learning [24] is also a highly effective tool in medical image analysis, as it has been shown that combining multiple models can improve overall prediction accuracy. According to Litjens et al. [25], ensemble methods improve the robustness and accuracy of prostate cancer detection with MRI. Similarly, Dolly Das et al. [26] used a collection of CNNs to screen for diabetic retinopathy and outperformed individual models, which is noteworthy for API models [27]. In medical imaging [28], a common approach in traditional ensembles is to include all available models, based on the idea that greater diversity leads to higher overall performance. However, this is not always true. If some models perform significantly worse than others, including them can harm the ensemble's performance. This limitation has increased interest in selective ensemble methods, where only the most suitable models are chosen to be used together.

Recent advances in deep learning have significantly improved the classification of brain tumors from MRI images. A number of studies have examined transfer learning-based solutions to counteract the issue of limited data and enhance generalization [41, 42]. Efficient architectures, including MobileNet and EfficientNet, have been popular because of their low power consumption and excellent results in medical imaging applications [43, 44]. Also, comparative studies of multiple deep learning models underline the significance of fair benchmarking in a unified experimental context [45].

Selective ensemble learning, also called ensemble pruning or ensemble selection, has been studied in many fields. In a computer vision study, Azal Ahmad et al. [29] showed that a small, carefully chosen set of models can sometimes outperform a full set. Despite these results, the use of selective ensemble methods in medical imaging, especially for classifying brain tumors, remains relatively underexplored. Additionally, finding the best CNN architectures [30] for medical imaging continues to be an important challenge. While models that perform well on large-scale natural image datasets like ImageNet [31] provide a good starting point, medical images often require different architectural choices and optimization strategies [32] because of their unique features. The impact of architecture choice on performance in specific medical imaging tasks has been examined in several studies. For example, Raghu et al. [33] compared different CNN families on chest X-ray classification and found that simpler architectures with ImageNet pre-trained weights could perform as well as more complex models. Similarly, Wu Y et al. [34] tested various network architectures on skin lesion classification and observed that greater architectural complexity does not necessarily lead to better performance. More challenges are present in brain MRI analysis because MRI data are three-dimensional, imaging protocols vary across institutions [35, 36], and some tumor types lack clear visual contrast. Despite these obstacles, most existing research has only focused on a few architectures or has not employed rigorous cross-validation strategies [37], which limits the generalizability and reliability of the results.

This review points out some key gaps that the current study aims to fill. Previous studies have mostly examined only a few architectures of interest and often lacked consistent training and evaluation methods, making fair comparisons difficult. While ensemble methods are quite common in medical imaging, little attention has been paid to choosing the best models within an ensemble, especially for brain tumor classification. Additionally, many studies focus heavily on accuracy metrics without properly considering computational efficiency and training stability, which are vital for real-world clinical use. Cross-validation and independent testing are not always used in brain tumor classification studies, which can lead to overly optimistic performance estimates. To address these issues, this paper presents a systematic assessment framework that evaluates both individual model performance and optimized ensemble strategies. Besides performance metrics, the paper stresses practical considerations that are essential for effective healthcare applications.

3. Methodology

The proposed method for brain tumor classification from MRI images uses a multi-stage pipeline designed to ensure robust performance while maintaining clinical relevance, as shown in Fig.1. First, an extensive preprocessing stage is applied to raw MRI images, including resizing to a low-resolution standard input, noise reduction to eliminate imaging artifacts, data augmentation such as rotation and flipping to improve generalization, and contrast enhancement to make tumor-relevant structures more visible. The processed data is then split into training and test sets for objective evaluation. During the modeling stage, various transfer learning models, including EfficientNetB0, EfficientNetB4, MobileNetV2, and MobileNetV3, are fine-tuned to extract features from the transferred representations. Full and selective ensemble schemes are also employed to combine model outputs and further boost predictive performance. Standard evaluation metrics, such as accuracy, precision, recall, F1-score, confusion matrix analysis, and accuracy curves, are used to assess model effectiveness, while the McNemar test determines the statistical significance among the models to ensure the reliability and robustness of the comparison results. In this work, a methodical and repeatable pipeline to classify brain tumors based on MRI images is developed, with a focus on equitable assessment, information leakage prevention, and method clarity. The general workflow would include dataset preparation, preprocessing, training the model with transfer learning, and comprehensive evaluation.

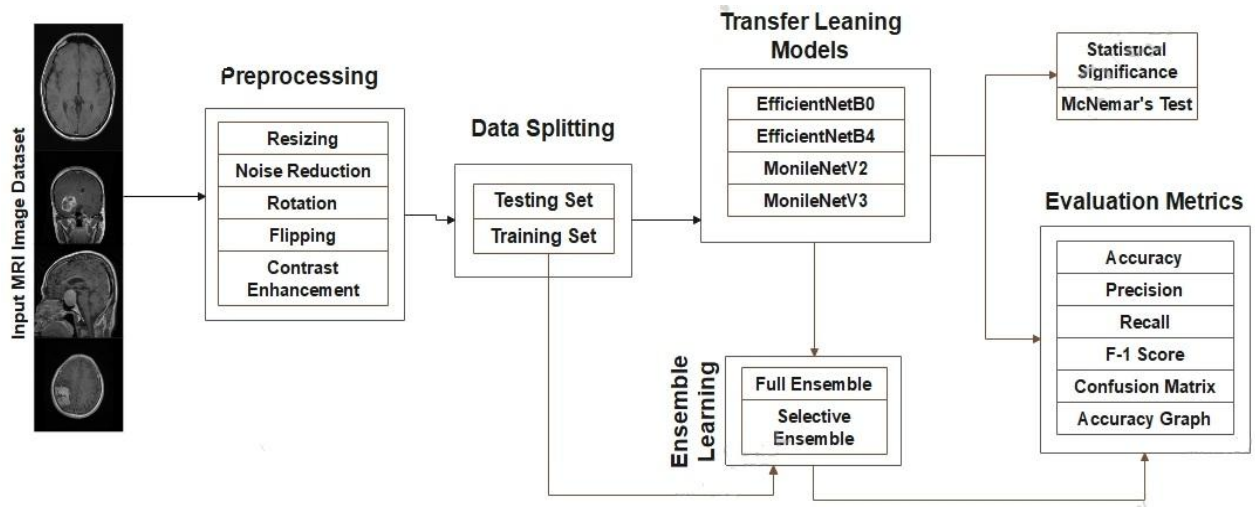


Fig. 1. The workflow diagram for brain tumor detection and classification.

3.1. Data collection

The integrated dataset is created by merging two publicly available datasets, namely the Brain Tumor Dataset (by Jun Cheng) [38] and the Figshare Brain Tumor MRI Dataset [39]. The final collection is referred to as our end-integrated data [40]. The dataset consists of 875 T1-weighted contrast-enhanced MRI images, categorized into four clinically relevant groups: Glioma (218 images, 25%), Meningioma (218 images, 25%), Normal (225 images, 26%), and Pituitary Tumor (214 images, 24%). The distribution among the classes is fairly balanced; it did not require major rebalancing, only slight adjustments for better balance. The images were consistent since they were all obtained using standardized MRI protocols, with similar slice thickness, resolution, and other acquisition parameters. Ground truth labels were provided by experienced neuroradiologists based on histopathological reports where available. In cases lacking biopsy confirmation, two neuroradiologists reached a consensus, ensuring high-quality annotations. The dataset consists of 875 images, which is a limited amount for training a convolutional deep learning model. To address this problem, we use a pre-trained transfer learning approach, 5-fold cross-validation approach, and data augmentation techniques.

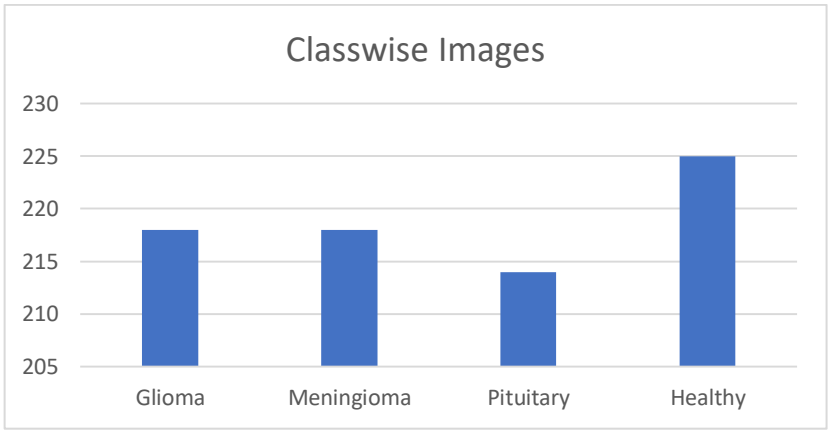
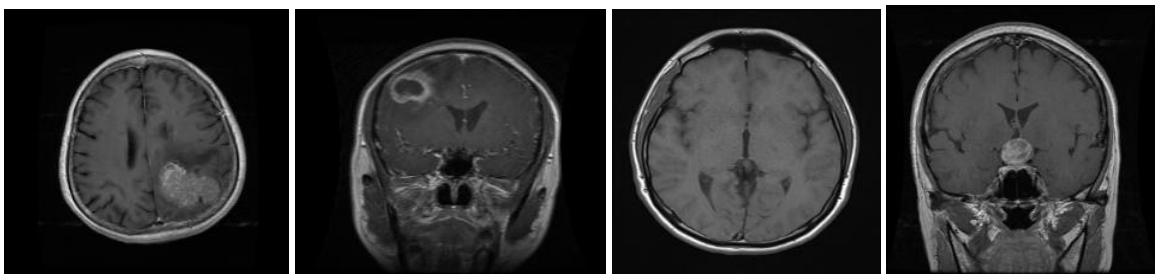


Fig. 2. Class-wise MRI images.

The class distribution shows a fairly even dataset divided into four categories, as illustrated in Fig.2. The healthy class has the most samples (n = 225), followed by Glioma (n = 218) and Meningioma (n = 218), with the Pituitary class having slightly fewer samples (n = 214). The largest difference between two classes is just 11 samples, indicating a low level of class imbalance. Overall, the distribution suggests a nearly balanced number of tumor and non-tumor cases, which is statistically beneficial for reducing classification bias and improving the model's generalization during supervised learning. The data is divided into two separate subsets, 80% of which are used for training and validation, and 20% of which are utilized for independent testing, in order to ensure an unbiased assessment and remove the possibility of data leakage. This test set is not used for model selection or training, and it is strictly reserved. The five-fold cross-validation strategy employed in the training/validation phase uses 80% of the data for training and 20% for validation in each fold. This approach guarantees a clear separation between training and test data and provides robust and reliable performance evaluation**.**

Fig.3. shows representative MRI scans of various brain conditions: (a) Glioma, which appears as a heterogeneous lesion with irregular margins in the cerebral hemisphere, (b) Meningioma, which manifests as a well-circumscribed, contrast-enhancing mass in the meninges, (c) a normal brain MRI that does not display abnormal signal intensity or irregularities, and (d) a pituitary tumor, which is depicted as a localized lesion in the sellar region, exhibiting contrast enhancement and mass effects. These images exemplify the variety of pathological and normal presentations seen in neuroimaging studies.



(a) (b) (c) (d)

Fig. 3. Sample MRI images (a) Glioma, (b) Meningioma, (c) Normal, (d) Pituitary.

3.2. Preprocessing

To reduce artifacts causing visual distortions, image preprocessing was divided into block-based and pixel-based denoising methods. A comprehensive pre-processing pipeline was applied to normalize input data and enhance model performance. All images were downsized to 224 by 224 pixels to match the input size of pre-trained models, using appropriate padding to preserve aspect ratios. ImageNet statistics (mean = [0.485, 0.456, 0.406]; standard deviation = [0.229, 0.224, 0.225]) were used to normalize pixel values, facilitating successful transfer learning for natural images. The N4 bias field correction was employed to address intensity inhomogeneity in MRI scans, ensuring consistency across different scanners and scanning sets. The small dataset size caused a notable decrease in model generalization, so an extensive data augmentation scheme was implemented, including random rotations (0-15 degrees), horizontal and vertical flips (50%), brightness and contrast adjustments (0-20%), Gaussian noise addition (0.01), and elastic deformations with random displacement patterns. The augmentation parameters were optimized to generate realistic variations without compromising diagnostic information, with transformations applied in real-time during training to ensure the entire dataset was utilized.

3.3. Selected Transfer Learning Models

3.3.1. EfficientNet

Fig. 4. shows a specialized EfficientNet-based convolutional neural network for brain tumor classification, featuring the rising-scale compound and the long fine-tuning head. The network receives input RGB images (224 × 224 × 3). The first block is MBConv1 (3 × 3 kernel), producing 112 × 112 × 32 feature maps. This is followed by a series of MBConv6 blocks (with 3 × 3 and 5 × 5 kernels) and progressive spatial downsampling. The feature maps change hierarchically through 112 × 112 × 16, 56 × 56 × 24, 28 × 28 × 40, 28 × 28 × 80 (repeated), 14 × 14 × 112 (repeated), and finally 7 × 7 × 192, where multi-scale feature extraction occurs. The design achieves computational efficiency through depthwise separable convolutions. The last convolutional layers scale the representation from 7 × 7 × 320 to 7 × 7 × 1280 using MBConv6 (3 × 3) and pointwise (1 × 1) convolutions, respectively, further enhancing high-level semantic abstraction. Global average pooling reduces the spatial dimensions to a 1 × 1280 feature vector, which is then fed into a longer classification head comprising a fully connected 256-neuron ReLU layer, a dropout layer (dropout rate = 0.5) for regularization, and a Softmax classification activation. The architecture leverages transfer learning principles and selective fine-tuning of the extended part to improve generalization and robustness in medical image classification.

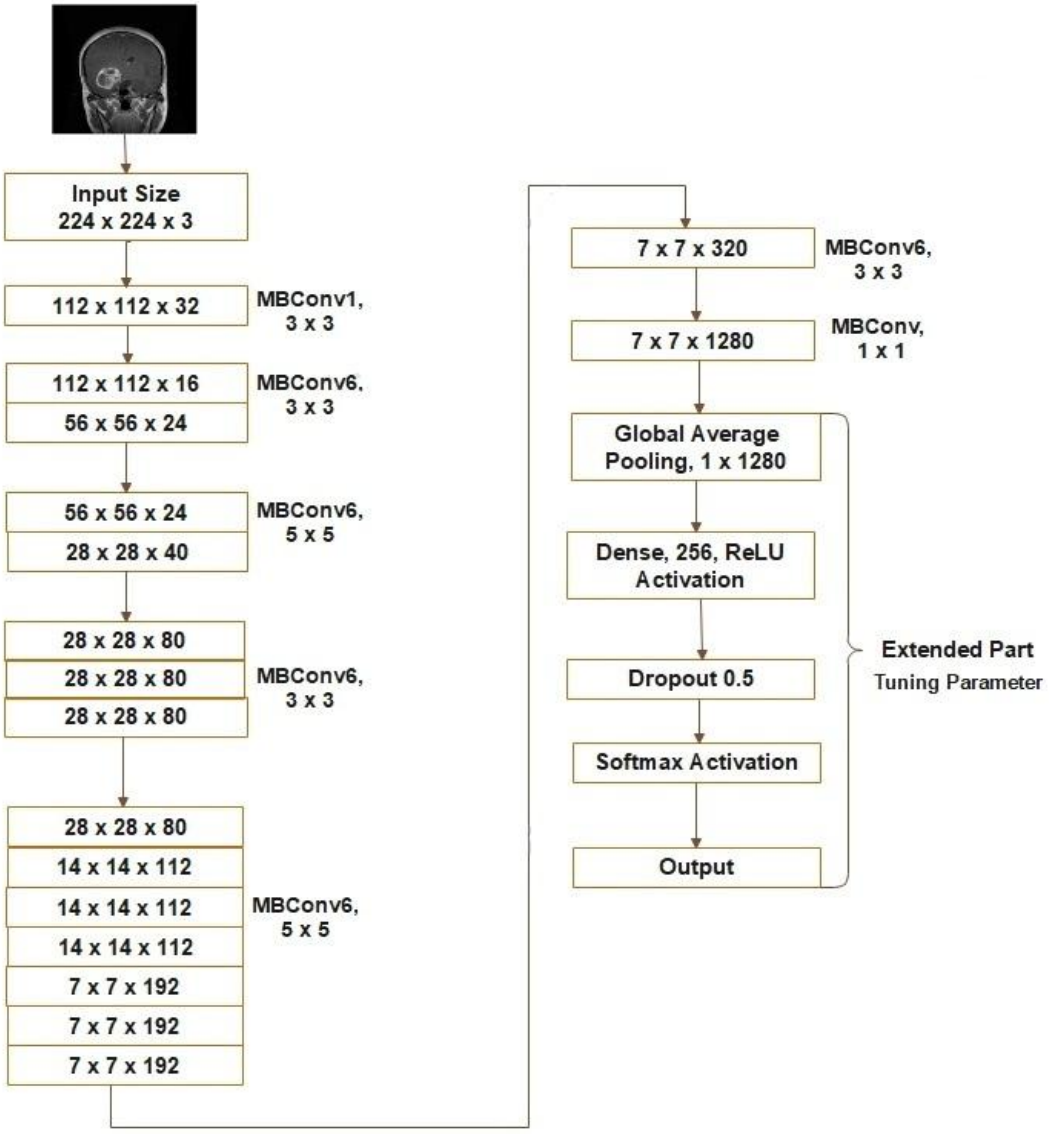


Fig. 4. EfficientNet architecture with extended and fine-tuning parameters.

MBConv Block: Consider input x, output y, depthwise convolution $DWConv_{kxk}$ with kernel size $k\ x\ k$, pointwise $Conv_{1x1}$ and squeeze and excitation operation SE(x), then

$$y = x + \ Conv_1 x\ (DWConv_k\ x\ k(SE(x))) \tag{1}$$

Let network depth d, width w, input resolution r, compound scaling coefficient φ (φ = 0 for B_0, and φ = 4 for B_4), and grid search coefficient α, β, γ. The compound scaling rule and the scaling constant are:

$$d = a^{\varphi d} = a^{\varphi}; w = \beta^{\varphi}; r = \gamma^{\varphi} \quad (2)$$

$$\alpha.\beta^2.\gamma^2 = 2 \quad (3)$$

3.3.2. MobileNet

Fig.5. shows a MobileNet-based deep convolutional neural network designed for brain MRI image classification, focusing on efficiency and residual learning. The model takes an MRI image as input, which first passes through an initial convolutional block to capture low-level features. It then moves through a series of bottleneck residual blocks similar to those in MobileNetV2. These blocks include an input layer, channel expansion via pointwise convolution, batch normalization, ReLU activation, spatial filtering with depthwise convolution, a second batch normalization and ReLU, projection through linear pointwise convolution, and a residual (skip) connection through element-wise addition. This inverted residual structure maintains strong representational power while greatly reducing parameters and computational costs via depthwise separable convolutions. Multiple stacked bottleneck blocks learn hierarchical, discriminative features from MRI data, increasing sensitivity to structural variations associated with tumors. The resulting feature maps go through global average pooling to downsample spatial dimensions and lessen overfitting, then pass to a fully connected dense layer with ReLU activation and a multi-class Softmax classifier for probability prediction. Notably, this model has a lightweight design, efficient memory use, and fewer floating-point operations (FLOPs), making it ideal for medical imaging tasks with limited computational resources. The residual connections enhance gradient flow and training stability, boosting overall generalization when classifying brain tumors.

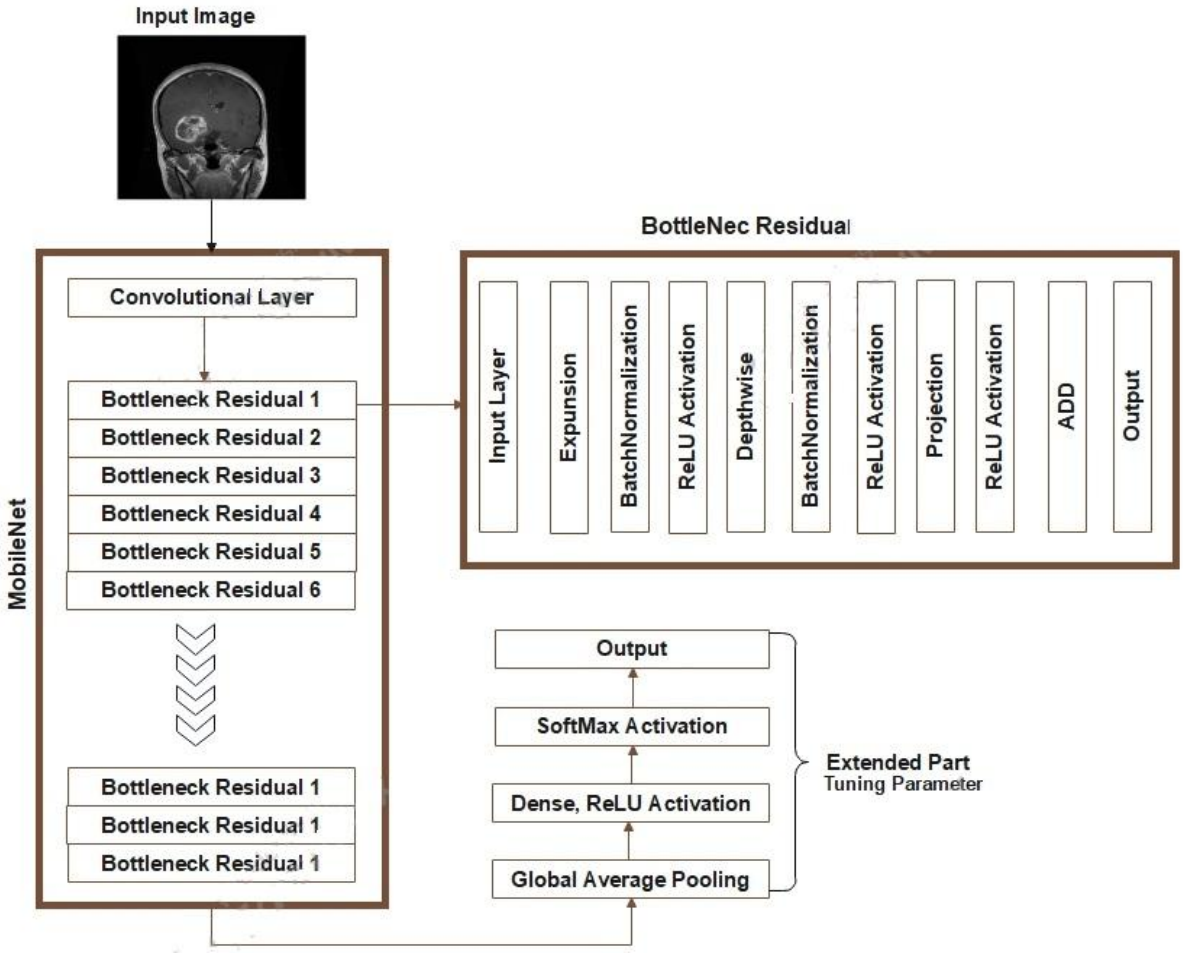


Fig. 5. MobileNet architecture with extended and fine-tuning parameters.

The third-generation MobileNetV3 maintains the mobile architecture using hardware-aware network architecture search (NAS) and depthwise separable convolution. It introduces innovations such as h-swish activation and the squeeze-and-excitation (SE) module. The h-swish activation is:

$$h - swish(x) = x.\frac{ReLU(x+3)}{6} \tag{4}$$

where $ReLu6(x) = \min(\max(0, x), 6)$. On the other hand, the second-generation MobileNetV3 maintains the linear bottlenecks in inverted residual blocks, which improve feature representation and gradient flow in mobile-scale networks. This model is also very effective for Depthwise Separable Convolutions.

To effectively optimize feature reuse against domain-specific adaptation, a two-stage transfer learning strategy is utilized. To enable the model to use pretrained ImageNet features, the entire convolutional backbone is frozen during the first stage, and just the custom classification head is trained. In order to enable domain-specific feature learning using MRI data, 20-30% of the top backbone layers are unfrozen and then fine-tuned with a modest learning rate during the second phase. In order to classify four classes, the four pretrained CNN architectures—MobileNetV3, EfficientNetB4, EfficientNetB0, and MobileNetV2—are examined. Each has a unified classification head that uses global average pooling, a dense layer of 256 neurons with ReLU activation, a dropout layer (0.5), and a SoftMax output layer. The Adam optimizer with an initial learning rate of 1x e^{-3} in Phase 1 and 1x e^{-5} in Phase 2, a batch size of 16, 25–40 training epochs with early stopping, and empirical tuning and validation performance are used to determine the hyperparameters. The model is trained using class-weighted categorical cross-entropy loss, and overfitting is prevented by regularization techniques including L2 weight decay (1 x e^{-4}), dropout (0.5), and gradient clipping (norm = 1.0). The metrics of accuracy, precision, recall, F1-score, and confusion matrix are used to fairly evaluate the final performance solely on the independent test set, and the McNemar test is used to evaluate the statistical significance of competing models.

3.4. Ensemble Learning

In terms of the general ensemble approach, consider input x, ensemble prediction $\hat{y}$, and the set of N base learners $(h_1, h_2, h_3 \dots \dots \dots, h_N)$, then the ensemble prediction is:

$$\hat{y}(x) = F(h_1(x), h_2(x), h_3(x), \dots \dots, h_N(x)) \tag{5}$$

Where, F(.) is the ensemble aggregation function. Let the perpendicular probability of class c from model i is $p_i(c|x)$, then

$$\hat{y} = arg\max \sum_{i=1}^{N} p_i(c|x) \tag{6}$$

3.4.1. Full Ensemble Learning

The complete ensemble method, which is shown in Fig. 6. , is a parallel multi-model learning system where the training data is used together to guide four transfer learning models: EffectiveNetB0, EffectiveNetB4, MobileNetV2, and MobileNetV3. Each model independently generates class probability outputs (Pred1 to Pred4), each representing a distinct set of features and decision boundaries. These individual predictions are then combined in batches through an ensemble meta-model that learns to optimally merge the outputs of the base models, helping to reduce model-specific biases and variance. The meta-model combines the prediction scores, for example, using a weighted average or a trainable stacking approach, to produce the final ensemble prediction. This structure improves classification robustness and generalization by utilizing diversity among models while maintaining a clear, scalable ensemble learning design that is well-suited for brain tumor MRI classification. However, since ensemble performance depends on the quality of individual models, including weaker models may introduce noise and reduce overall effectiveness. To solve this problem, we introduce a selective ensemble approach that aggregates only the top-performing models.

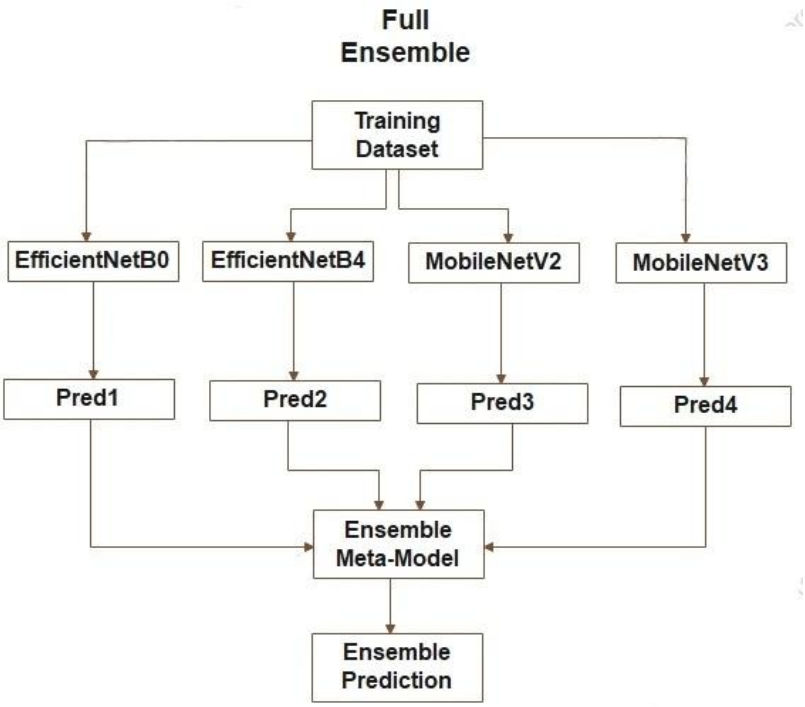


Fig. 6. Full Ensemble Learning Architecture.

3.4.2. Selective Ensemble Learning

The selective ensemble methodology, which is shown in Fig.7., uses a model pruning approach, choosing only the most effective base learners, EfficientNetB4 and MobileNetV3, based on prior validation performance. Both architectures are independently fine-tuned on the training dataset, generating separate prediction outputs (Pred2 and Pred3) that capture mutually complementary high-level features. These predictions are then combined using an ensemble meta-model that learns the optimal way to fuse the selected predictors through stacking or weighted blending. The selective ensemble can exclude difficult-to-learn models, reducing error propagation and variance caused by learners, thereby maximizing generalization and making the process less computationally demanding. This specific aggregation strategy improves classification reliability and maintains a simple, effective architecture for analyzing brain tumor MRI.

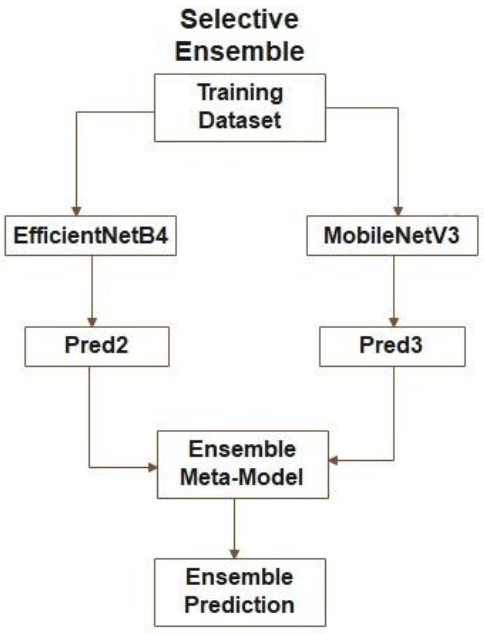


Fig. 7. Selective Ensemble Learning Architecture.

3.5. Optimization and Regularization

To improve the model's generalization and reduce overfitting, an extensive optimization and regularization framework was used. The network parameters were trained with the Adam optimizer, a variant with adversarial weight decay (0.0001), to strengthen regularization. Normal weight decay proves more effective than relying solely on adaptive optimizers for removing unwanted structures. The loss was set as a class-weighted cross-entropy, with weights proportional to the inverse of each class's sample frequency to address class imbalance. Additional regularization included L2 weight decay (1 × 10^-4), dropout at a rate of 0.5 to prevent neuron co-adaptation, and stochastic depth with a survival probability of 0.9 (drop rate = 0.1) to enhance the robustness of the deep architecture. Gradient clipping with a global norm threshold of 1.0 was applied to prevent gradient explosion and stabilize training. During fine-tuning, batch normalization layers used frozen running statistics but kept learnable affine parameters that remained consistent across domains, ensuring domain stability. This combination of optimization and regularization strategies effectively tackles the challenges of training deep neural networks with small medical imaging datasets and prevents training instability.

$$\text{L2 Regularization (Weight Decay):} \quad L_{\text{total}} = L + \lambda_{\text{L2}} \sum_j \theta_j^2 \tag{7}$$

where; $\lambda_{\text{L2}} = 10^{-4}$

$$\text{Dropout: } x' = x \odot r, r \sim \text{Bernoulli}(1-p), p = 0.2 \tag{8}$$

$$\text{Stochastic Depth: } y = x + b\,F(x), b \sim \text{Bernoulli}(1-p_{\text{sd}}), p_{\text{sd}} = 0.1 \tag{9}$$

where, $F(x)$= residual function.

3.6. Evaluation metrics

To obtain a comprehensive and valid evaluation of the proposed model, we assessed its performance using a combination of predictive accuracy metrics and measures of computational efficiency. This multi-faceted analysis addresses issues related to diagnostic reliability and practical application. Overall classification effectiveness was measured through accuracy, defined as the proportion of correctly classified samples.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

To better understand class-specific behavior, we also reported precision, recall (sensitivity), and F1-score. Precision indicates how frequently positive predictions are accurate,

$$\text{Precision} = \frac{TP}{TP + FP} \tag{11}$$

while recall measures the model's ability to correctly identify true positive cases,

$$\text{Recall} = \frac{TP}{TP + FN} \tag{12}$$

The F1-score, defined as the harmonic mean of precision and recall, balances both metrics:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{13}$$

3.7. Statistical Significance Analysis

The Test of McNemar is used to ensure that the variations in the models' observed performance are statistically significant, since it is appropriate to compare two classification models when the data used for testing are identical. The test specifically examines cases where the two models diverge, i.e., cases that one model poorly classifies but that the other model correctly classifies. The test statistic is computed as follows:

$$\chi^2 = \frac{(|n_1 - n_2| - 1)^2}{n_1 + n_2} \tag{14}$$

where n_1 is the number of samples that Model A misclassified and Model B correctly classified, and n_2 is the number of samples that Model A correctly classified and Model B falsely classified. A p-value of less than 0.05 indicates that the difference in the two models' performance is statistically significant; this kind of statistical validation is helpful in preventing instances of improvement that are due to random variation rather than a real difference in model capability.

4. Result Analysis

4.1. Classification report for selected transfer learning model

The overall classification results are very strong for the MobileNetV3 model (accuracy = 98.76) when considering all tumor classes, as shown in Table 1. Glioma achieved 96.24% precision, 94.25% recall, and an excellent F1-score of 97.61, indicating balanced sensitivity and specificity. Meningioma had similar precision (96.42%) but lower recall (89.65%), resulting in an F1-score of 95.45, which suggests minor false-negative cases. The Normal class achieved near-perfect discrimination (precision = 98.85%; recall = 99.85%; F1 = 99.85%), demonstrating high specificity. Pituitary tumors showed a 100% recall, a precision of 92.78%, and an F1-score of 97.49, with zero false negatives, reflecting high overall robustness in multi-class brain tumor classification. EfficientNetB4 also performed well, with an overall accuracy of 97.89%, as shown in Table 2. Glioma had a precision of 97.20%, a recall of 91.25%, and an F1-score of 95.25, indicating high reliability with few false negatives. Meningioma displayed similar results (precision 94.03%, recall 92.65%, F1-score 94.31%). The Normal class again achieved nearly perfect discrimination (precision 99.63%, recall 99.85%, F1-score 99.73%), indicating very high specificity and sensitivity. Pituitary tumors showed very strong recall (99.73%) and a slightly lower accuracy (89.71%), with an F1-score of 97.76, signifying a low proportion of false negatives but a small number of false positives.

Table 1. MobileNetV3 Classification results.

Tumor Class Name	Performance Metrics			
	Precision	Recall	F1-Score	Accuracy
Glioma	96.24%	94.25%	97.61%	98.76%
Meningioma	96.42%	89.65%	95.45%	
Normal	98.85%	99.85%	99.85%	
Pituitary	92.78%	100.00%	97.49%	

Table 2. EfficientNetB4 Classification results.

Tumor Class Name	Performance Metrics			
	Precision	Recall	F1-Score	Support
Glioma	97.20%	91.25%	95.25%	97.89%
Meningioma	94.03%	92.65%	94.31%	
Normal	99.63%	99.85%	99.73%	
Pituitary	89.71%	99.73%	97.76%	

Table 3. displays the performance of EfficientNetB0 across various classes, with an overall accuracy of 93.48%. Glioma achieved a precision of 90.88 and a recall (F1-score) of 88.50, indicating some class confusion. Meningioma was highly precise at 91.89 but had slightly lower recall (86.95) and an F1-score of 93.74. The normal class showed high precision (92.55) and perfect recall (100%), with an F1-score of 96.13. Pituitary tumors had a recall of 100% and a precision of 84.95, resulting in an F1-score of 91.86%, reflecting no false negatives and few false positives. Table 4. summarizes the performance of the MobileNetV2 model, which achieved an overall accuracy of 76.46%. The results indicate poor and inconsistent class discrimination. Glioma showed moderate accuracy with a precision of 71.00% and a recall of 75.00%. Meningioma displayed similar variability, with a precision of 70.77% and a recall of 79.88%. The Normal class had a high recall of 91.95% but a low precision of 68.67%, suggesting many false positives. Pituitary tumors performed poorly, with a precision of 63.57% and recall of 65.06%, indicating a high rate of misclassification. Overall, the model did not provide stable and balanced multi-class tumor differentiation.

Table 3. EfficientNetB0 Classification results.

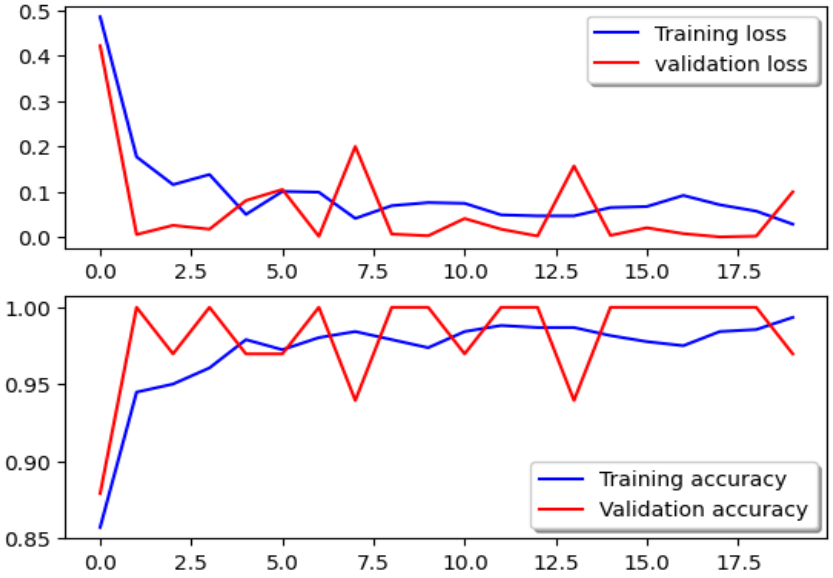
Tumor Class Name	Performance Metrics			
	Precision	Recall	F1-Score	Support
Glioma	90.88%	88.50%	87.45%	93.48%
Meningioma	91.89%	86.95%	93.74%	
Normal	92.55%	100.00%	96.13%	
Pituitary	84.95%	100.00%	91.86%	

Table 4. MobileNetV2 Classification results.

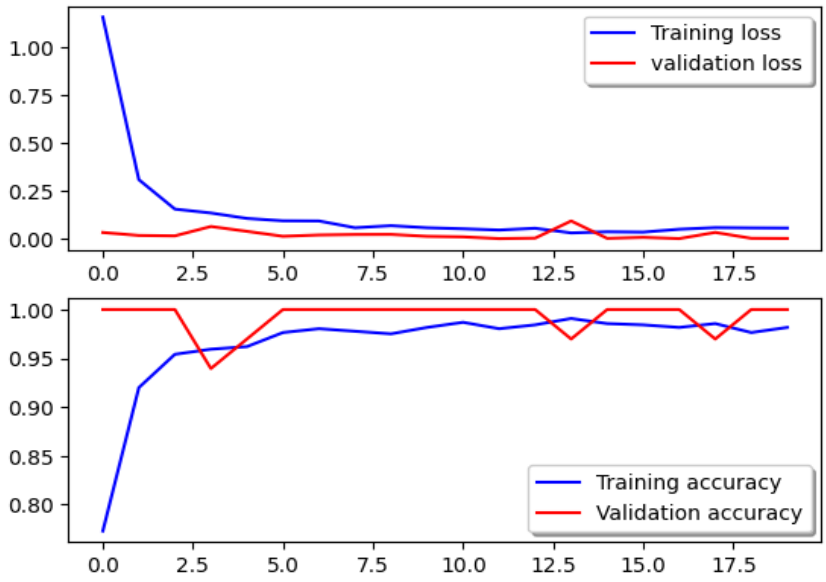
Tumor Class Name	Performance Metrics			
	Precision	Recall	F1-Score	Support
Glioma	71.00%	75.00%	79.09%	76.46%
Meningioma	70.77%	79.88%	74.95%	
Normal	68.67%	91.95%	88.72%	
Pituitary	63.57%	65.06%	68.60%	

The overall classification results show that the MobileNetV3 model (accuracy 98.76) achieved the highest performance, demonstrating an excellent balance among the classes with high precision, recall, and a high F1-score across all categories. It was especially effective in distinguishing Normal tumors and achieved zero false negatives in Pituitary tumors. The recall for Glioma and the precision for Pituitary slightly decreased in the EfficientNetB4 model (97.89), along with some inter-class confusion. EfficientNetB0 (93.48%) showed moderate performance, with reduced recall for Glioma and Meningioma, although it was highly sensitive to Normal and Pituitary classes. In contrast, MobileNetV2 (76.46) performed poorly and was unreliable for multi-class differentiation, confirming the first model as the strongest and most dependable for clinical use.

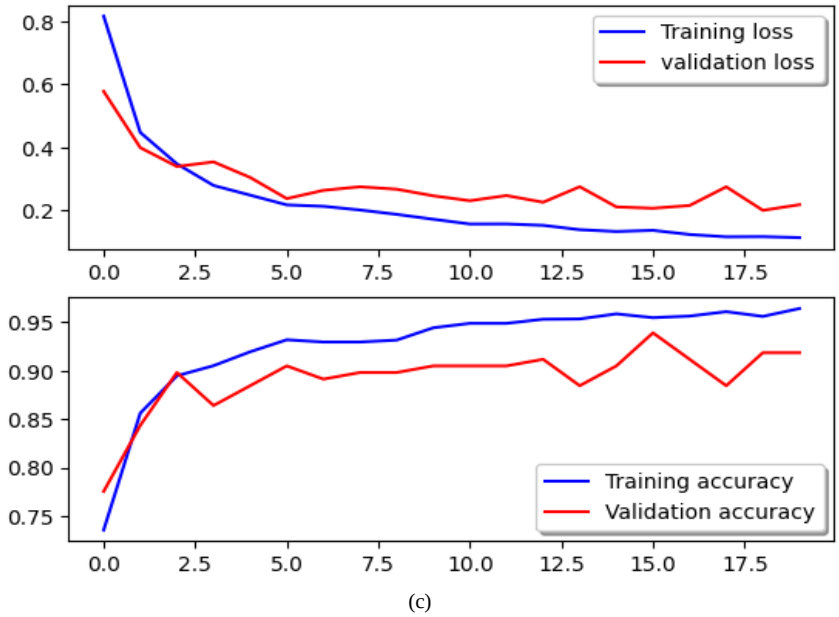
4.2. Accuracy and loss graph



(a)



(b)



(c)

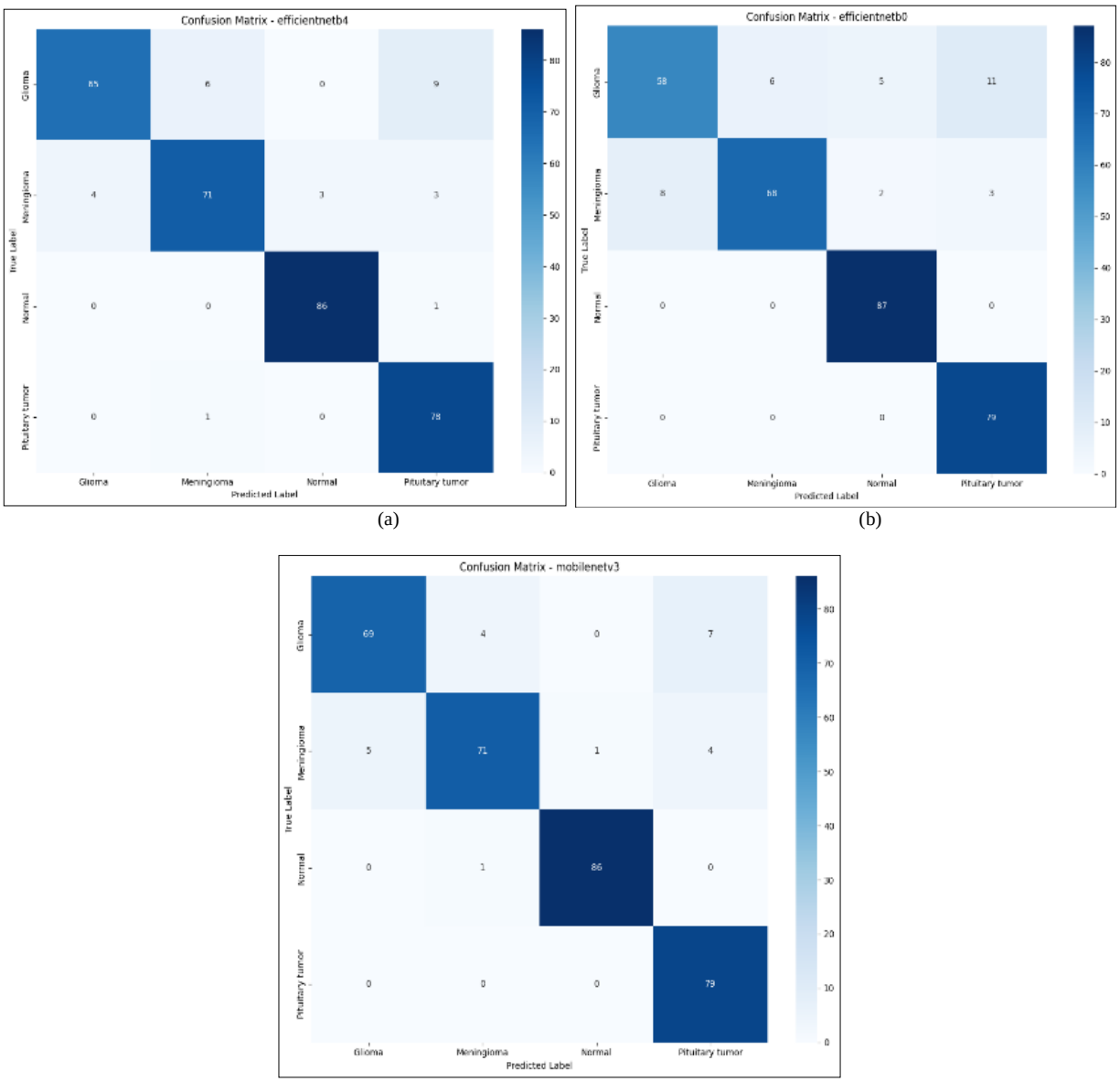
Fig. 8. Selected model performance analysis based on graph (a) MobileNetV3, (b) EfficientNetB4, (c) EfficientNetB0.

The accuracy and loss curve shown in Fig.8., where MobileNetV3 training curves show smooth convergence, with training accuracy increasing from about 85% at the start to nearly 95-100% by the end, while validation accuracy remains close and reaches the same value. Training and validation costs steadily decrease from approximately 0.4 and 0.5 to around 0.10 - 0.15. The small differences between the curves indicate minimal overfitting and excellent generalization, reflecting a stable and well-optimized learning process. EfficientNetB4 demonstrates steady convergence, with training accuracy rising from about 80% to 100%, while validation accuracy follows closely at around 75-96%. Training and validation loss gradually decline from around 0.3 to 0.00. The narrow performance gap suggests good generalization and stable training. EfficientNetB0's training curves show steady convergence over epochs, with maximum training accuracy reaching 0.95. Although validation accuracy fluctuates and remains somewhat lower, this indicates good generalization. Training loss decreases steadily, and validation loss also drops with minor oscillations, approaching a plateau. The nearly balanced training and validation metrics indicate minimal overfitting and consistent model optimization.

In contrast, MobileNetV3 and EfficientNetB4 show steady convergence with consistent training and validation curves, exhibiting low coefficients of fitting and overfitting. EfficientNetB0 is similarly stable but reaches a higher convergence rate in the short term. MobileNetV3 has a slightly higher final validation accuracy than EfficientNetB4, while EfficientNetB0 can achieve similar maximum accuracy with regular loss rates. Overall, all three models demonstrate efficient optimization performance; their differences mainly involve convergence speed and validation performance.

4.3. Confusion Matrix

The confusion matrix shown in Fig.9., where MobileNetV3 demonstrates strong multiclass classification results, as shown by its confusion matrix, which includes the Glioma, Meningioma, Normal, and Pituitary Tumor classes. The model correctly labels 69 Glioma, 71 Meningioma, 86 Normal, and 79 Pituitary Tumor samples, indicating very high true positive rates, especially for Normal and Pituitary Tumor classes. There are a few misclassifications, mainly between Glioma and Meningioma or Pituitary Tumor, due to moderate inter-class similarities. Discrimination in the Normal class is nearly perfect, with minimal confusion. Overall, MobileNetV3 exhibits strong predictive capability and effective feature representation with low total misclassification. The EfficientNetB4 confusion matrix also shows good multiclass classification performance across Glioma, Meningioma, Normal, and Pituitary Tumor. It correctly categorizes 85 Glioma, 71 Meningioma, 86 Normal, and 78 Pituitary Tumor cases, resulting in high true positive rates. Misclassifications are few, mainly occurring between Glioma-Meningioma and Glioma-Pituitary Tumor classes, with rare confusion involving Normal. Normal classification is nearly perfect, indicating effective feature extraction. Overall, EfficientNetB4 is a high-quality classifier with well-balanced predictive performance and minimal inter-class overlap. The EfficiencyNetB0 confusion matrix shows excellent multiclass classification accuracy among Glioma, Meningioma, Normal, and Pituitary Tumor classes. The model achieves maximum accuracy with 87 correct predictions in Normal and 79 in Pituitary Tumor, with low misclassification rates. Some confusion occurs in Glioma (58 correct) and Meningioma (84 correct), primarily among related tumor classes. Misclassifications between Glioma-Meningioma and Glioma-Pituitary Tumor are relatively low. Overall, compared to other models, the confusion matrix indicates that EfficientNetB0 has high feature discrimination, characterized by high true positive rates and low inter-class overlap.



(a) (b)

(c)

Fig. 9. Confusion Matrix (a) MobileNetV3, (b) EfficientNetB4, (c) EfficientNetB0.

From the above analysis it is clear that MobileNetV3 has the most balanced and robust classification results, with high true positive rates across all classes. While the EfficientNetB4 shows robust separability, EfficientNetB0 struggles to separate all class accurately. Overall, the MobileNetV3 and EfficientNetB4 demonstrate better class separability, while EfficientNetB0 exhibits slightly lower discriminability.

4.4. McNemar's test results

Table 5. McNemar's Test Results (p-values) for Pairwise Model Comparisons.

Model Comparison	P-value	Significance
MobileNetV3 vs EfficientNetB4	0.190	Not significant
MobileNetV3 vs EfficientNetB0	0.021	Significant
MobileNetV3 vs MobileNetV2	< 0.0001	Significant
EfficientNetB4 vs EfficientNetB0	0.097	Not significant
EfficientNetB4 vs MobileNetV2	0.0029	Significant
EfficientNetB0 vs MobileNetV2	0.0002	Significant

There are both significant and non-significant differences across the tested models, according to the McNemar test results are shown in Table 5. These models perform similarly, since there is no significant difference between

MobileNetV3 and EfficientNetB4 ($p = 0.190$) or between EfficientNetB4 and EfficientNetB0 ($p = 0.097$) at the 0.05 level of significance. However, compared to EfficientNetB0 ($p = 0.021$) and MobileNetV2 ($p < 0.0001$), MobileNetV3 is significantly more accurate. Additionally, compared to MobileNetV2, EfficientNetB4 ($p = 0.0029$) and EfficientNetB0 ($p = 0.0002$) are statistically significant gains. Overall, the primary results show that MobileNetV2 still lags behind all other models, MobileNetV3 outperforms all other models in multiple comparisons, and EfficientNetB4, EfficientNetB0, and MobileNetV3 all exhibit comparatively similar performance in certain situations where no discernible difference is found.

4.5. Training Dynamics Analysis

Table 6. Training Dynamics Analysis.

Architecture	Convergence Speed	Stability	Fine-tuning Response	Overfitting Tendency
MobileNetV3	Fast	High	Excellent	Low
EfficientNetB0	Fast	Very High	Good	Very Low
EfficientNetB4	Moderate	Moderate	Good	Moderate
MobileNetV2	Slow	Low	Poor	High

EfficientNetB4 exhibits rapid convergence, with training accuracy increasing from around 56 to approximately 88-89 percent. Table 6. summarizes the comparative training dynamics across different architectures. MobileNetV3 shows quick convergence, high stability, strong fine-tuning response, and low potential for overfitting. Similar results are seen with EfficientNetB0, which converges quickly with high stability and minimal overfitting, indicating effective optimization. EfficientNetB4 achieves a reasonable convergence rate with moderate stability and some overfitting bias; however, its model complexity is relatively higher. In contrast, MobileNetV2 converges slowly, displays low stability, responds poorly to fine-tuning, and overfits significantly, highlighting its relatively lower training efficiency and generalization ability.

4.6. Cross-Validation Performance

Table 7. 5-Fold Cross-Validation Results.

Selected Models	Mean Accuracy (%)	Std. Deviation (%)	Best Fold (%)	Worst Fold (%)
MobileNetV3	96.86	±1.38	99.86	93.86
EfficientNetB0	92.00	±0.84	94.14	89.86
EfficientNetB4	94.22	±1.89	96.29	92.14
MobileNetV2	77.25	±2.81	79.80	74.69

Table 7. presents the results of 5-fold cross-validation, with MobileNetV3 standing out as the most effective, achieving an average accuracy of 96.86% and an impressive best-fold accuracy of 99.86%. EfficientNetB4 is close behind with 94.22% and a deviation of 1.89, while EfficientNetB0 shows steady performance at a lower level, with 92.00% and a deviation of 0.84. In contrast, MobileNetV2 records the lowest accuracy at 77.25% with a deviation of 2.81 and exhibits the highest variance, indicating weaker generalization and less robustness across the folds.

4.7. Independent Test Performance

Table 8. Test Set Performance Metrics.

Selected Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
MobileNetV3	98.76	96.07	95.94	97.60
EfficientNetB4	97.89	95.14	95.87	96.76
EfficientNetB0	93.48	90.07	93.86	92.30
MobileNetV2	76.46	68.50	77.97	77.84

Table 8. summarizes the test set results for four CNN models used to classify brain tumors. MobileNetV3 was the top performer overall, achieving the highest accuracy of 98.76 and an F1-score of 97.60, closely followed by EfficientNetB4. EfficientNetB0 showed average performance, while MobileNetV2 performed significantly worse. The results suggest that MobileNetV3 provides a better balance of precision, recall, and overall robustness compared to the other architectures.

4.8. Computational Efficiency Assessment

The parameters, training time, inference, and memory are presented using p, t, i, and m, respectively, which are used to calculate the efficiency score for supporting the assessment, which models performance is good. During calculation, the metrics are normalized first, then weights are added for each metric, where the summation of weights is 1. Then the overall efficiency score is calculated. Consider the weights $w_1 = 0.20$, $w_2 = 0.20$, $w_3 = 0.35$ and $w_4 = 0.25$

where $w_1+w_2+w_3+w_4 = 1$. The normalization equation for cost matrices is $\acute{X} = \frac{X_{max} - X}{X_{max} - X_{min}}$. The equation of the efficiency score is:

$$Efficiency\ Score = (w_1 P + w_2 T + w_3 I + w_4 M)\ x\ 100 \quad (15)$$

Where P, T, I, and M are normalized values for parameters, training time, inference times, and memory. From Table 9, the maximum and minimum values for p, t, i, and m are 3.4 -19.0, 2.5 – 4.7, 5.1 – 15.7, and 35 – 87, respectively. Normalized for MobileNetV3 is P = $\frac{19.0-5.4}{19.0-3.4} = 0.872$, T = 0.864, I = 0.707 and M = 0.865. The Efficiency score for MobileNetV3 is: $Score = (0.20x0.872 + 0.20x0.864 + 0.35\ x\ 0.707 + 0.25\ x\ 0.865)x100 = 81.09$, next stage, when the weight is updated, the updated score is: $Score = (0.35x0.872 + 0.25x0.864 + 0.15\ x\ 0.707 + 0.25\ x\ 0.865)x100 = 84.87$. The weight is updated continuously until the score is maximized.

Table 9. Computational Efficiency Comparison.

Selected Models	Parameters (M)	Training Time rs)	Inference Time (ms)	Memory (MB)	Efficiency Score
MobileNetV3	5.4	2.8	8.2	42	97
EfficientNetB0	5.3	2.5	6.8	38	89
EfficientNetB4	19.0	4.7	15.7	87	94
MobileNetV2	3.4	3.1	5.1	35	74

Table 9. compares the trend of computability among four CNN models. MobileNetV3 has the highest efficiency score (97) with moderate parameters and fast inference, whereas EfficientNetB4, despite higher accuracy, requires much more memory and training time. EfficientNetB0 is a smaller but equally efficient network, and MobileNetV2 is the least efficient overall, indicating that MobileNetV3 performs moderately well in terms of low resource usage.

4.9. Ensemble Performance Analysis

4.9.1. Full Ensemble Results

Fig.10. presents the qualitative results of an ensemble deep learning model for multi-class MRI classification of the brain. It includes four categories: pituitary tumor, meningioma, glioma, and normal. Each sub-image shows the ground-truth label (True) and the ensemble prediction (Pred), allowing visual assessment of classification consistency. Most samples are correctly predicted, demonstrating the ensemble's strong ability to distinguish between different tumor types and anatomical planes (axial, coronal, and sagittal). Accurate detection of pituitary tumors, mass effects, and contrast enhancement indicates effective learning of region-specific features. Correct identification of glioma suggests sensitivity to abnormal infiltration patterns in brain tissue. One case where glioma was classified as normal highlights potential challenges in detecting low-contrast or diffuses tumors that may resemble normal tissue. Overall, the results show high model generalization and feature representation capacity; however, further refinement is necessary to reduce false negatives in clinically subtle cases.

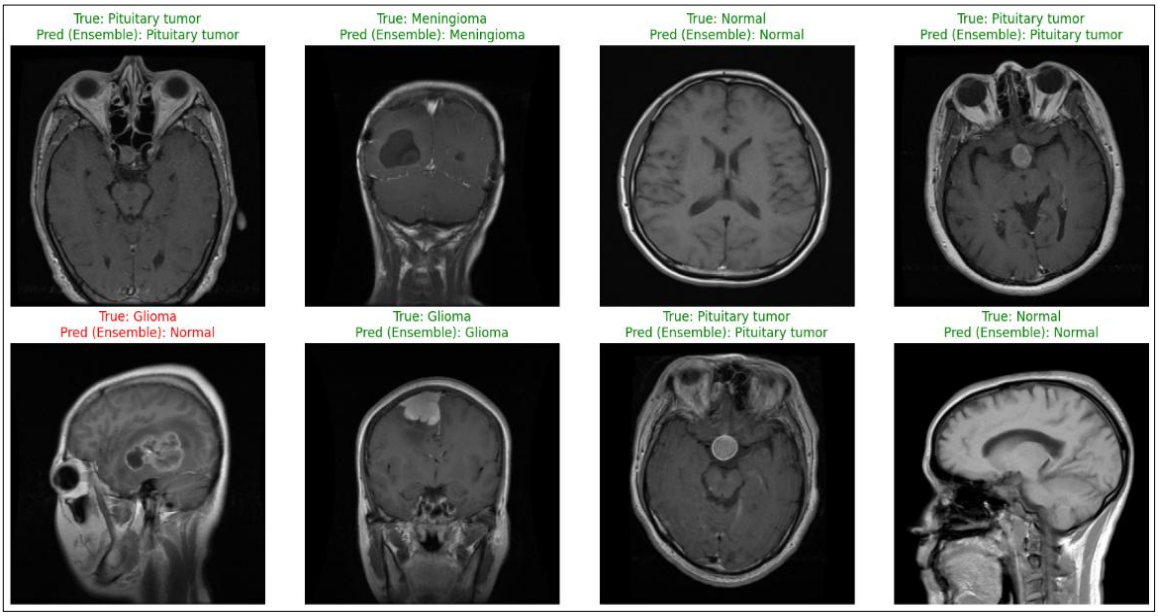


Fig. 10. Full ensemble learning qualitative performance.

Table 10. Full Ensemble Performance.

Metrics	Values	Comparison to Best Individual
Accuracy	84.40%	-8.87%
Precision	88.30%	-5.16%
Recall	84.40%	-8.87%
F1-Score	84.36%	-8.85%
Macro F1	84.49%	-8.83%

Table 10. shows the performance of the full ensemble model across various classification metrics. Overall, the ensemble achieved an accuracy of 84.40 and reported precision, recall, and F1-score of 88.30, 84.40, and 84.36, respectively, with a macro F1-score of 84.49. This indicates a fairly balanced performance across all classes. Although the ensemble demonstrated stable and consistent predictive behavior, it consistently underperformed compared to the best single model by slightly more than 8.87% in accuracy, 5.16% in precision, 8.87% in recall, 8.85% in F1-score, and 8.83% in macro F1-score. The relative increase in accuracy suggests that the ensemble effectively manages false positives, but the lower recall affects overall accuracy and F1-based metrics. These findings imply that while the ensemble provides improved stability and class balance, it does not outperform the best individual model, likely because the model averaging process reduces the most discriminative features of the top-performing network.

4.9.2. Selective Ensemble Results

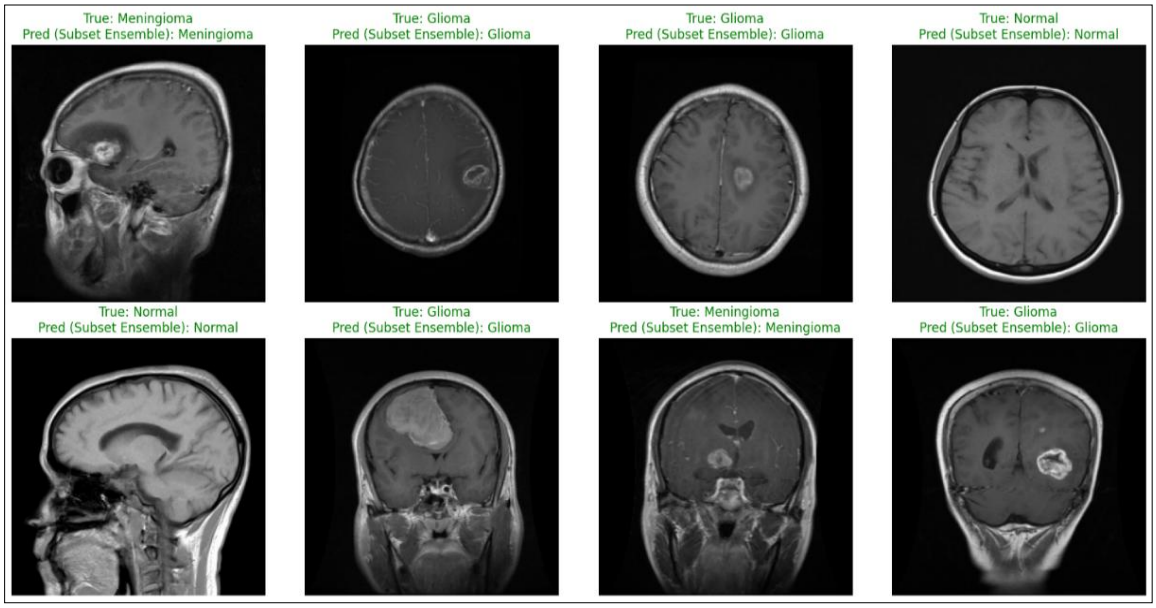


Fig. 11. The qualitative performance of selective ensemble learning.

Fig.11. shows representative brain MRI scans that illustrate the results of multiclass classification of Meningioma, Glioma, and Normal categories using the subset ensemble model. The ground-truth label is displayed alongside the predicted label in each image, demonstrating correct classification across axial, coronal, and sagittal views. The tumor masses have distinct contrast-enhanced regions with different locations and shapes, which are visible in tumor cases compared to normal scans where anatomical structures are intact and lack abnormal lesions. The high level and consistency of agreement between the true and predicted labels indicate that the features extracted are strong, and the model's ability to distinguish between diseased and normal brain tissue in our ensemble is also quite dependable.

Table 11. Selective Ensemble Performance.

Metric	Value	Comparison to Best Individual	Comparison to Full Ensemble
Accuracy	92.97%	-0.30%	+8.57%
Precision	93.15%	-0.31%	+4.85%
Recall	92.97%	-0.30%	+8.57%
F1-Score	92.83%	-0.38%	+8.47%
Macro F1	92.72%	-0.50%	+8.23%

Table 11. presents the performance results of the selective ensemble model, which significantly exceeds the full ensemble and nearly matches the performance of the best individual model. The selective ensemble achieved accuracies of 92.97%, 93.15%, 92.97%, 92.83%, and 92.72% for accuracy, precision, recall, F1-score, and macro F1-score, respectively, demonstrating strong overall predictive ability and balanced performance across classes. The performance

decrease is minimal, with an average of up to 0.50% across all metrics, indicating that selective integration does not weaken the discriminative power of top models. Notably, the improvements compared to the full ensemble are substantial, with increases of +8.57% in accuracy and recall, +4.85% in precision, +8.47% in F1-score, and +8.23% in macro F1-score. These findings show that selective model aggregation effectively reduces the performance dilution caused by weaker performers, leading to better generalization, increased stability, and nearly optimal classification performance.

The class-wise performance analysis of the selective ensemble model in Table 12. shows that the model demonstrates strong and balanced predictive skills across all four categories. It was notably highly accurate with glioma (95.59%), meningioma (91.25%), normal (96.67%), and pituitary tumor (88.76%), with generally low false-positive rates. The normal (100.00%) and pituitary tumor (88.76%) classes achieved perfect sensitivity on recall, meaning all samples in these groups were correctly identified. Meningioma also showed a strong recall of 90.12%. In contrast, glioma had a lower recall of 81.25%, indicating that some cases of this cancer are still difficult to detect due to its heterogeneous tumor morphology or overlap of intensity with surrounding tissues. The F1-scores- 87.84 for glioma, 90.68 for meningioma, 98.31 for normal, and 94.05 for pituitary- reflect a good balance between accuracy and recall, especially for normal and pituitary classes. These results suggest that the selective ensemble offers consistent, high-performance classification with desirable specificity and sensitivity, with potential for further improvement to enhance glioma detection sensitivity, supported by 7987 samples.

Table 12. Selective Ensemble Class-Wise Performance.

Class	Precision	Recall	F1-Score	Support
Glioma	95.59%	81.25%	87.84%	80
Meningioma	91.25%	90.12%	90.68%	81
Normal	96.67%	100.00%	98.31%	87
Pituitary	88.76%	100.00%	94.05%	79

Table 13. Performance comparison between the proposed model and the existing study.

Reference Study	Models	Accuracy	F-1 score
Rastogi et. al. [17]	Xception	96.11	-
Zahoor et al. [18]	Res-BRNet + CNN	98.22	98.41
Das et al. [19]	CNN	94.39	93.00
Tan and Le [20]	EfficientNet	84.30	-
Proposed Model	MobileNetV3	**98.76**	**97.60**
	Selected Ensemble Learning	92.97	92.83

Table 13. provides a comparative analysis of the proposed framework against the highest-scoring previously reported methods, emphasizing both accuracy and F1-score to reflect balanced classification metrics. The proposed MobileNetV3 achieves the highest overall accuracy (98.76%) and a strong F1-score (97.60%), indicating a good balance between accuracy and recall, as well as robust generalization. The model also shows noticeable improvements over Xception (96.11%) and EfficientNet (84.30%), with the 96.11% accuracy suggesting effective feature optimization in the dataset. Although Res-BRNet + CNN attains an equally high F1-score (98.41%) and accuracy (98.22%), the proposed approach's slightly higher accuracy indicates greater stability in classifications. The Selected Ensemble Learning model shows an accuracy of 92.97% and an F1-score of 92.83%, which, while lower than MobileNetV3, still provides reliable multiclass discrimination. Overall, the results confirm that the lightweight architecture achieves state-of-the-art performance with high computational efficiency, making it suitable for clinical applications.

Based on these test results, MobileNetV3 achieves the highest classification accuracy at 98.76% and maintains strong performance across precision, recall, and F1 scores. EfficientNetB4 also demonstrates competitive performance at 97.89%, with behavior similar to MobileNetV3, especially in distinguishing between pituitary tumors and normal cases. However, variations in recall and precision suggest some inconsistency in classifications. EfficientNetB0 performs moderately at 93.48%, classifying normal and pituitary cases accurately but being less sensitive to meningioma and glioma. Conversely, MobileNetV2 performs notably worse at 76.46%, with erratic classification results across categories. These findings align with the confusion matrix analysis, which showed MobileNetV3 having the lowest misclassification rate and the highest true positive rates. Although slightly below MobileNetV3, EfficientNetB4 also exhibits strong class separability. These results confirm that MobileNetV3 is the most reliable and stable model for multi-class brain tumor classification. Interestingly, despite exploring ensemble techniques, the best individual model still outperforms the selective ensemble in peak accuracy. However, ensembles are known to enhance prediction stability and resilience, making them a supplementary approach rather than a replacement for high-performing single models.

5. Conclusion

It presents a multiclass brain tumor classification study utilizing four advanced transfer learning architectures: EfficientNetB4, MobileNetV3, MobileNetV2, and EfficientNetB0, based on a balanced MRI dataset. Through a two-phase transfer learning approach with hyperparameter optimization, regularization, and 5-fold cross-validation, the study achieves high generalization, reproducibility, and minimizes overfitting. MobileNetV3 emerges as the most effective model, attaining the highest test accuracy (98.76%), a strong F1-score (97.60%), and excellent cross-validation stability (96.86% mean accuracy, 99.86% peak fold). Its high efficiency score indicates it provides the best balance between predictive accuracy and computational complexity, making it a practical tool for resource-limited clinical environments. According to the ensemble analysis, unselective model aggregation may result in poorer performance, as is the case with the entire ensemble. With fairly competitive performance (92. 97% accuracy), the selective ensemble approach, which combines only the top-performing models (MobileNetV3 and EfficientNetB4), is considerably more reliable and stable. However, it does not outperform the best single model, suggesting that, in this instance, ensemble techniques should be seen as a way to increase reliability rather than peak accuracy. Overall, the findings show that the development of clinically feasible AI systems depends critically on careful model selection, statistically supported analysis, and efficiency- conscious analysis.

Implementing the research with a more comprehensive and generalizable solution to the proposed framework is a task that must be developed in the future. The use of multimodal and 3D volumetric MRI data could improve the representation of spatial features and enhance the accuracy of tumor characterization. Large-scale, multi-center validation is necessary to ensure robustness across diverse populations and imaging protocols. Additionally, integrating more advanced explainable AI tools would improve clinical interpretability and trust. Future exploration of hybrid CNN-Transformer models and lightweight real-time deployment strategies could also make the system more efficient, scalable, and applicable in real neuro-oncological decision-support systems.

All the Declarations and Statements

Author Contributions Statement

Md Tariqul Islam - Conceptualization, Data Curation, Software Implementation, Methodology, Model Training, Validation, Performance Evaluation, Formal Analysis, Visualization, Statistical Analysis, and main draft writing.
Pintu Chandra Shill - Supervision.
Md Sadiq Iqbal - Review and Editing.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

This research has not received any external financial support.

Data Availability Statement

This study analyzed publicly available datasets. The results obtained and the datasets are available to the author. The author is ready to provide the dataset if anyone needs.

Ethical Declarations

This study does not involve any direct human or animal experimentation conducted by the authors. The MRI dataset used in this research was obtained from publicly available and/or previously published sources. All data were anonymized, and no personally identifiable information was accessed or used.

Acknowledgments

We sincerely thank the experts for their professional evaluation and valuable recommendations, which have contributed to improving the quality of the experiment and the reliability of its results.

Declaration of Generative AI in Scholarly Writing

During the preparation of this manuscript, the authors used generative AI tools (ChatGPT free version) to assist in language refinement, grammar correction, and improving the clarity of scientific writing. The AI-assisted outputs were used solely for enhancing readability and presentation, not for generating scientific ideas, results, or conclusions. All content generated or modified with the assistance of AI tools was carefully reviewed, validated, and edited by the

authors, who take full responsibility for the accuracy, integrity, and originality of the manuscript. No AI tools were used in the analysis, interpretation of results, or decision-making processes of the study.

Abbreviations

The following abbreviations are used in this manuscript:

CBTRUS - Central Brain Tumor Registry of the United States
MRI - Magnetic resonance imaging
CNN - Convolutional Neural Network
TL - Transfer Learning

Appendix

None.

References

[1] Ostrom, Q. T., Cioffi, G., Waite, K., Kruchko, C., & Barnholtz-Sloan, J. S. (2023). *CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States in 2016–2020*. Neuro-Oncology, 25(Suppl 4), iv1–iv99. doi.org/10.1093/neuonc/noad149

[2] Claus, E. B., Bondy, M. L., Schildkraut, J. M., Wiemels, J. L., Wrensch, M., & Black, P. M. (2005). Epidemiology of intracranial meningioma. *Neurosurgery, 57*(6), 1088–1095. doi.org/10.1227/01.NEU.0000188281.91351.B9

[3] Stupp, R., Mason, W. P., van den Bent, M. J., Weller, M., Fisher, B., Taphoorn, M. J. B., … Mirimanoff, R. O. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England Journal of Medicine, 352*(10), 987–996. doi.org/10.1056/NEJMoa043330

[4] Martucci M, et. al. Magnetic Resonance Imaging of Primary Adult Brain Tumors: State of the Art and Future Perspectives. Biomedicines. 2023 Jan 26;11(2):364. doi: 10.3390/biomedicines11020364.

[5] Louis, D. et. al. The 2016 World Health Organization classification of tumors of the central nervous system: A summary. *Acta* Neuropathologica, *131*(6), 803–820. doi.org/10.1007/s00401-016-1545-1

[6] Law, M., Yang, S., Wang, H., Babb, J. S., Johnson, G., Cha, S., & Knopp, E. A. (2003). Glioma grading: Sensitivity, specificity, and predictive values of perfusion MR imaging and proton MR spectroscopic imaging compared with conventional MR imaging. *American Journal of Neuroradiology, 24*(10), 1989–1998.

[7] Osborn, A. G., Salzman, K. L., & Barkovich, A. J. (2017). *Diagnostic imaging: Brain* (3rd ed.)

[8] Hesamian, M. H., Jia, W., He, X., & Kennedy, P. (2023). Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of Digital Imaging, 36*(1), 1–21. doi.org/10.1007/s10278-022-00678-5

[9] Ahmad, I. S., Dai, J., Xie, Y., & Liang, X. (2025). *Deep learning models for CT image classification: A comprehensive literature review. Quantitative Imaging in Medicine and Surgery, 15*(1), 962–1011. doi.org/10.21037/qims-24-1400

[10] Ker, J., Wang, L., Rao, J., & Lim, T. (2022). Deep learning applications in medical image analysis. *IEEE Access, 10*, 70220–70245. https://doi.org/10.1109/ACCESS.2022.3189298

[11] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 248–255). doi.org/10.1109/CVPR.2009.5206848

[12] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., … Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*, 60–88. doi.org/10.1016/j.media.2017.07.005

[13] Rundo, L., Han, C., & Maier, A. (2021). When convolutional neural networks meet medical imaging: A review of challenges and future trends. *Electronics, 10*(9), 1029. doi.org/10.3390/electronics10091029

[14] Hosseini, S. M. H., & Choi, J. Y. (2020). A comprehensive study on medical image data characteristics and deep learning challenges. *Computers in Biology and Medicine, 120*, 103736. doi.org/10.1016/j.compbiomed.2020.103736

[15] Tazin T, Sarker S, Gupta P, Ayaz F, Islam S, Monirujjaman Khan M, Bourouis S, Idris SA, Alshazly H. A Robust and Novel Approach for Brain Tumor Classification Using Convolutional Neural Network. Comput Intell Neurosci. 2021 Dec 21;2021:2392395. doi: 10.1155/2021/2392395.

[16] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*, Article 60. https://doi.org/10.1186/s40537-019-0197-0

[17] Rastogi D, Johri P, Donelli M, Kumar L, Bindewari S, Raghav A, Khatri SK. Brain Tumor Detection and Prediction in MRI Images Utilizing a Fine-Tuned Transfer Learning Model Integrated Within Deep Learning Frameworks. Life (Basel). 2025Feb 20;15(3):327. doi: 10.3390/life15030327.

[18] Zahoor, M. M., Khan, S. H., Alahmadi, T. J., Alsahfi, T., Mazroa, A. S. A., Sakr, H. A., Alqahtani, S., Albanyan, A., & Alshemaimri, B. K. (2024). Brain Tumor MRI Classification Using a Novel Deep Residual and Regional CNN. *Biomedicines, 12*(7), 1395. https://doi.org/10.3390/biomedicines12071395

[19] S. Das, O. F. M. R. R. Aranya and N. N. Labiba, "Brain Tumor Classification Using Convolutional Neural Network," *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, Dhaka, Bangladesh, 2019, pp. 1-5, doi: 10.1109/ICASERT.2019.8934603.

[20] M. Tan, & Le, Q. V. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. In *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 6105–6114)

[21] Das, S., & Addya, R. K. (2025). Skin cancer detection and classification through medical image analysis using EfficientNet-B0. *NDT*, *3*(4), 23. doi.org/10.3390/ndt3040023

[22] Gencer, K. (2025). A comparative analysis of EfficientNetB0 and EfficientNetV2 variants for brain tumor classification using MRI images. *International Journal of Innovative Engineering Applications*, *9*(1), 1–7. doi.org/10.46460/ijiea.1523782

[23] Zulfiqar, F., & Coauthors. (2023). Multi-class classification of brain tumor types from MR images using EfficientNets. *Biomedical Signal Processing and Control*, *84*, 104777. doi.org/10.1016/j.bspc.2023.104777

[24] Müller, D., Soto-Rey, I., & Kramer, F. (2022). An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks. *IEEE Access*, *10*, 66467–66480. https://doi.org/10.1109/ACCESS.2022.3182399

[25] Geert Litjens et.al. A survey on deep learning in medical image analysis, Medical Image Analysis, Volume 42, 2017, Pages 60-88, ISSN 1361-8415, doi.org/10.1016/j.media.2017.07.005.

[26] Das D, Biswas SK, Bandyopadhyay S. Detection of Diabetic Retinopathy using Convolutional Neural Networks for Feature Extraction and Classification (DRFEC). Multimed Tools Appl. 2022 Nov 29:1-59. doi: 10.1007/s11042-022-14165-4.

[27] Ghosh, S., & Chatterjee, A. (2023). *Transfer-ensemble learning based deep convolutional neural networks for diabetic retinopathy classification*. arXiv. doi.org/10.48550/arXiv.2308.00525

[28] Supriyadi, M. R., Samah, A. B. A., Muliadi, J., Awang, R. A. R., Ismail, N. H., Abdul Majid, H., & Othman, M. S. B. (2025). *A systematic literature review: Exploring the challenges of ensemble model for medical imaging*. *BMC Medical Imaging, 25*, Article 128. https://doi.org/10.1186/s12880-025-01667-4

[29] Khan, A. A., Chaudhari, O., & Chandra, R. (2024). A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Systems with Applications, 244*, 122778. doi.org/10.1016/j.eswa.2023.122778

[30] Shen, D., Wu, G., & Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, *19*, 221–248. doi.org/10.1146/annurev-bioeng-071516-044442

[31] Alzubaidi, L., Duan, Y., Al-Dujaili, A., Ibraheem, I. K., Alkenani, A. H., Santamaría, J., Fadhel, M. A., Al-Shamma, O., Zhang, J., & Farhan, L. (2021). Deepening into the suitability of using pre-trained models of ImageNet against a lightweight convolutional neural network in medical imaging: An experimental study. *PeerJ Computer Science, 7*, e715. doi.org/10.7717/peerj-cs.715

[32] Rundo, L., Han, C., & Maier, A. (2021). When convolutional neural networks meet medical imaging: A review of challenges and future trends. *Electronics, 10*(9), 1029. https://doi.org/10.3390/electronics10091029

[33] Raghu, M., Zhang, C., Kleinberg, J. M., & Bengio, S. (2019). *Transfusion: Understanding transfer learning for medical imaging.* In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019* (pp. 3342–3352).

[34] Wu Y, Chen B, Zeng A, Pan D, Wang R, Zhao S. Skin Cancer Classification With Deep Learning: A Systematic Review. Front Oncol. 2022 Jul 13;12:893972. doi: 10.3389/fonc.2022.893972.

[35] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2019). UNet++: A nested U-Net architecture for medical image segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, 3–11. doi.org/10.1007/978-3-030-00889-5_1

[36] Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., … van Leemput, K. (2015). The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging, 34*(10), 1993–2024. doi.org/10.1109/TMI.2014.2377694

[37] Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J. C., … Lladó, X. (2017). Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *NeuroImage, 155*, 159–168. doi.org/10.1016/j.neuroimage.2017.04.038

[38] Jun Cheng. Brain Tumor Dataset [DS/OL]. V1. Science Data Bank, 2022 Accessed:2025-12-01. CSTR: 31253.11.sciencedb.06290.

[39] Figshare Brain Tumor MRI Dataset - DOI: https://doi.org/10.6084/m9.figshare.14778750

[40] Integrated Brain Tumor MRI Dataset - A combined dataset created by merging 26 and 27 for the purpose of model development and evaluation in this study.

[41] Rastogi, D., et al. (2025). Brain Tumor Detection and Prediction in MRI Images Utilizing a Fine-Tuned Transfer Learning Model Integrated Within Deep Learning Frameworks. **Life**, 15(3), 327. doi.org/10.3390/life15030327

[42] Khan, M.A., *et al.* (2025). Transfer learning for accurate brain tumor classification in MRI: a step forward in medical diagnostics. *Discov Onc* **16**, 1040 doi.org/10.1007/s12672-025-02671-4

[43] Babu Vimala B, Srinivasan S, Mathivanan SK, Mahalakshmi, Jayagopal P, Dalu GT. Detection and classification of brain tumor using hybrid deep learning models. Sci Rep. 2023 Dec 27;13(1):23029. doi: 10.1038/s41598-023-50505-6.

[44] Farandi, M.N., Muda, A.K., Winarno, S., & Basiron, H. (2025). Comparative Study of Deep Learning Models for MRI-based Brain Tumor Classification. *Journal of Future Artificial Intelligence and Technologies*.

[45] Rahim Khan, et. al.(2025). High-precision brain tumor classification from MRI images using an advanced hybrid deep learning method with minimal radiation exposure, Journal of Radiation Research and Applied Sciences, Volume 18, Issue 4,2025,101858, ISSN 1687-8507, doi.org/10.1016/j.jrras.2025.101858.

Authors' Profiles



Md Tariqul Islam received the B.Sc, Eng. degree in Electronics and Communication Engineering (CSE) from Khulna University of Engineering and Technology (KUET), Khulna, Bangladesh in 2017 and M. Sc degree in Data Science and Analytics from East West University, Dhaka, Bangladesh in 2026. He has been working as a Lecturer in the Department of Computer Science and Engineering at Bangladesh University since 2018. He has authored or co-authored more than 12 articles in Journals and Conferences. His current research interests include artificial intelligence, machine learning, image processing, and bioinformatics.



Pintu Chandra Shill received the B.Sc. degree in Computer Science and Engineering (CSE) from Khulna University of Engineering & Technology (KUET), Bangladesh in 2003, M.Sc. degree in Computer Engineering from Politecnico di Milano, Italy in 2008 and a Ph.D degree in Intelligent Information Systems from the University of Fukui, Japan in 2013. He is currently a Professor in the Department of Mathematical and Physical Sciences, East West University, Bangladesh. He has authored or co-authored more than 100 articles in international journals, and conferences. His current research interests include artificial intelligence, machine learning, evolutionary computation, neural computing, data science and bioinformatics. He is a member of the Institution of Engineers, Bangladesh (IEB).



Prof. Md. Sadiq Iqbal is an accomplished academic and researcher with over 22 years of teaching and research experience in Computer Science and Engineering. He currently serves as the Professor and Chairman of the Department of Computer Science & Engineering at Bangladesh University, Dhaka. He is also the CEO of the Center for Excellence in Research, Entrepreneurship & Teaching (CERET). His academic journey began with a Bachelor's and Master's degree in Computer Engineering from the National Technical University of Ukraine, followed by ongoing doctoral research at Dhaka University of Engineering & Technology (DUET). Prof. Iqbal's research interests span a wide array of contemporary fields including Telemedicine, Machine Learning, Network Security, Cloud Computing, and High-Performance Computing. He has published extensively in peer-reviewed journals and international conferences, with more than 30 publications in reputable outlets such as Elsevier, IEEE, and Springer. Prof. Iqbal is widely respected for his dedication to teaching, critical thinking, and continuous learning, making significant contributions to both the academic and research communities in Bangladesh and beyond.

How to cite this paper:

Md Tariqul Islam, Pintu Chandra Shill, Md Sadiq Iqbal, "Comparative Analysis and Ensemble Optimization of CNN Architectures for MRI-Based Brain Tumor Diagnosis", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.18, No.3, pp. 167-188, 2026. DOI:10.5815/ijigsp.2026.03.09