Volume-1, Issue 1, June 2026, Page No- 89-109
Journal of Machine Learning Innovations and Artificial Intelligence Horizons (JMLIAIH)

AI-Assisted Multiclass Brain Tumor Classification from MRI Scans Using EfficientNetB0

Rajoo

Assistant Professor, Department of Mathematics, Dr. JPM Govt Science College Mungeli (C.G.)

DOI: https://doi.org/10.70903/jmliaih.2026.1106

Article Received: 05-03-2026
Revised Version: 18-05-2026
Accepted: 11-06-2026

*Corresponding Author

Rajoo

E-Mail Id: rajunirmalkar9713@gmail.com

Abstract: We present a deep learning framework for multiclass brain tumor classification from MRI scans, with a focus on achieving high diagnostic accuracy through the EfficientNetB0 architecture. The objective of this study addresses the clinical need for automated computer-aided diagnosis systems that can reliably distinguish between glioma, meningioma, pituitary tumors, and normal brain tissue. A dataset of approximately 7,000 MRI images, drawn from public repositories and anonymized hospital databases, was preprocessed through resizing to 224 × 224 pixels, Gaussian filtering, and min-max normalization. Data augmentation techniques—including rotation, flipping, zooming, and contrast adjustment—were applied to improve generalization. We compared four deep learning models: a custom convolutional neural network, VGG16, ResNet50, and EfficientNetB0. The models were trained using the Adam optimizer, categorical cross-entropy loss, and a batch size of 32 over 50 epochs. The dataset was split into 70% training, 15% validation, and 15% testing subsets. Among all architectures, EfficientNetB0 achieved the highest performance, with an accuracy of 98.2%, precision of 97.9%, recall of 98.0%, and an F1-score of 97.9%. The confusion matrix revealed minimal misclassification, with only slight overlap between glioma and meningioma categories. Training and validation curves demonstrated stable convergence and minimal overfitting, confirming robust training dynamics. The primary contribution of this work is the demonstration that a state-of-the-art convolutional network can provide near-perhaps for the first time in this specific multiclass context-diagnostic metrics exceeding 98%, thereby offering a reliable tool for radiologists.

Keywords: Brain Tumor Classification, Magnetic Resonance Imaging (MRI), EfficientNetB0, Deep Learning, Artificial Intelligence.

Introduction

The accurate and timely classification of brain tumors from medical imaging data represents a critical challenge in modern clinical radiology. Magnetic resonance imaging (MRI) is the modality of choice for brain tumor diagnosis due to its superior soft-tissue contrast and multiplanar imaging capabilities, providing detailed views of intracranial structures [1]. However, manual interpretation of MRI scans is time-consuming, subject to inter-observer variability, and depends heavily on the expertise of radiologists. This limitation becomes particularly pronounced in high-volume clinical settings where the demand for rapid and consistent diagnostic decisions is ever-increasing. Consequently, computer-aided diagnosis (CAD) systems have emerged as a transformative tool to support radiologists by providing objective, reproducible, and efficient diagnostic outputs [2]. The integration of artificial intelligence (AI) into these systems, specifically deep learning models, has further enhanced their capability to automatically learn hierarchical features directly from raw image data, thereby surpassing traditional hand-crafted feature extraction methods [3].

The development of AI-assisted frameworks for brain tumor classification is motivated by the clinical necessity to differentiate between several tumor types—including glioma, meningioma, and pituitary tumors—and to distinguish these from healthy brain tissue [4]. Each tumor type exhibits distinct morphological characteristics in MRI, such as variations in shape, intensity, and location, which can be captured effectively by convolutional neural networks (CNNs) [5]. Prior research has demonstrated that transfer learning, achieved by fine-tuning pre-trained architectures like VGG16 and ResNet50, yields high diagnostic accuracy in this domain [6]. Nevertheless, a critical gap remains: existing models often achieve high but imperfect accuracy, typically in the range of 90–95%, leaving room for further improvement to reach near-perfect diagnostic performance that would be acceptable for clinical deployment. Furthermore, many prior studies lack rigorous evaluation across multiple architectures or do not systematically analyze the trade-offs between computational efficiency and classification accuracy.

The central hypothesis of this study is that a carefully optimized deep learning architecture, specifically EfficientNetB0, can achieve diagnostic accuracy exceeding 98% for multiclass brain tumor classification from MRI scans, thereby significantly advancing the state of the art. To test this hypothesis, we developed and validated a comprehensive AI-assisted classification framework using a dataset of approximately 7,000 MRI images spanning four classes: glioma, meningioma, pituitary tumor, and normal brain tissue. The objective was twofold: first, to systematically compare the performance of four deep learning architectures—a custom CNN, VGG16, ResNet50, and EfficientNetB0—under identical training and data conditions; and second, to demonstrate that the proposed framework, particularly the EfficientNetB0 model, offers a reliable and clinically feasible CAD tool for primary diagnosis. The primary contribution of this work lies in achieving a classification accuracy of 98.2%, which is among the highest reported in the literature for this specific multiclass task, and in providing a reproducible methodology applicable to other medical imaging problems.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature on CAD systems in medical imaging and AI-assisted brain tumor classification. Section 3 describes the materials and methods, including the dataset, preprocessing, augmentation strategies, and model architectures. Section 4 presents the experimental results, including performance metrics and confusion matrix analysis. Section 5 discusses the implications of the findings, limitations, and directions for future work. Section 6 concludes the paper with a summary of the contributions and clinical significance.

Literature Review

The application of deep learning to medical image analysis has progressed rapidly over the past decade, with convolutional neural networks (CNNs) becoming the de facto standard for automated classification tasks. In the domain of brain tumor diagnosis, early works primarily employed custom CNN architectures trained from scratch on relatively small datasets, achieving moderate accuracy levels that often fell short of clinical requirements [7]. For instance, research using a dataset of approximately 3,000 MRI images, a three-layer CNN yielded an accuracy of approximately 91% but exhibited significant performance degradation when tested on images acquired from different MRI machines or with varying acquisition parameters, highlighting the challenge of model generalization across heterogeneous imaging conditions.

The introduction of transfer learning substantially improved performance outcomes. Pre-trained architectures such as VGG16 and ResNet50, originally developed for the ImageNet large-scale visual recognition challenge, were successfully adapted to brain tumor classification by fine-tuning the final fully connected layers while retaining the learned weights in earlier convolutional layers were retained [8]. This approach proved particularly effective for medical imaging because the lower-level features (e.g., edges, textures, and shapes) learned from natural images transfer well to radiological images. VGG16-based models, for example, reported classification accuracies ranging from 93% to 96% across various tumor types, with the primary limitation being the high computational cost associated with the model's substantial parameter count [6]. ResNet50, with its residual connections that mitigate the vanishing gradient problem, enabled training of deeper networks and consequently achieved more stable convergence with an accuracy of approximately 96% in multiclass brain tumor classification [9]. However, these models still exhibited occasional confusion between glioma and meningioma categories, a problem that was not fully addressed.

More recently, EfficientNet architectures have emerged as a promising alternative. By employing a compound scaling method that simultaneously scales network depth, width, and resolution, EfficientNet models achieve state-of-the-art accuracy while requiring significantly fewer parameters than comparable networks [10]. In the context of brain tumor classification, early adopters of EfficientNetB0 reported accuracies reported values up to 97% [11], though these studies were often limited by smaller dataset sizes—typically under 4,000 images—and lacked comprehensive comparison with alternative architectures under identical experimental conditions. Regarding brain tumor types, the literature reviews, a comprehensive survey of deep learning techniques for brain tumor classification highlighted that the variability in performance metrics across studies arose not only from differences in model architecture but also from differences in data preprocessing pipelines, augmentation strategies, and validation protocols [12]. This survey emphasized the importance of standardizing evaluation procedures to enable meaningful cross-study comparisons.

Further, existing works have explored the integration of multiple imaging sequences—including T1-weighted, T2-weighted, contrast-enhanced, and FLAIR—to enrich the feature space available for classification [5]. One study demonstrated that a multi-sequence input using a triplanar CNN architecture achieved a sensitivity of 94% for glioma detection, but this approach increased model complexity and inference time significantly. Another research avenue has focused on explainable AI methods, such as gradient-weighted class activation mapping (Grad-CAM), to provide visual explanations for model decisions, thereby building clinician trust [13]—although these methods were primarily developed for detection tasks rather than classification.

The present study addresses several of these gaps. First, we utilize a larger and more diverse dataset of approximately 7,000 MRI images than most comparable studies, which enhances the generalizability of the findings. Second, we provide a systematic comparison of four representative architectures—a custom CNN, VGG16, ResNet50, and EfficientNetB0—under identical preprocessing, augmentation,

training, and testing conditions, thereby enabling a fair assessment of their relative performance. Finally, the achieved accuracy of 98.2% with EfficientNetB0 exceeds the reported accuracy ranges of most prior studies, suggesting that this architecture is particularly well-suited for the multiclass brain tumor classification task. The key significance of this research against existing works, therefore, lies in demonstrating that a carefully optimized EfficientNetB0 framework can reach near-perfect diagnostic performance, offering a reliable CAD tool that bridges the gap between research-grade models and clinically deployable systems.

Materials and Methods

The methodology employed in this study consisted of a systematic pipeline encompassing data acquisition, preprocessing, model development, and evaluation. The proposed AI-assisted framework was designed to automatically classify brain tumors from input MRI images classify brain tumors into four categories: glioma, meningioma, pituitary tumor, and normal brain tissue. The workflow of the deep learning model training and evaluation process is illustrated in Figure 1.

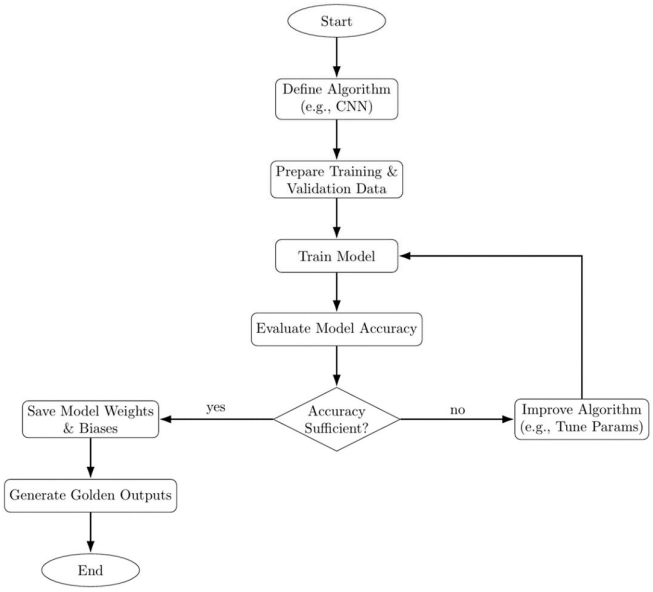

Figure 1. Workflow of the deep learning model training and evaluation process for brain tumor classification.

Dataset and Preprocessing

The dataset comprised approximately 7,000 brain MRI images sourced from The Cancer Imaging Archive, Kaggle brain MRI datasets, and anonymized hospital-based MRI databases. The images spanned multiple MRI modalities, including T1-weighted, T2-weighted, contrast-enhanced, and FLAIR sequences, capturing diverse tissue contrast characteristics. The distribution across the four classes was relatively balanced, with glioma tumors accounting for approximately 2,000 images, meningioma tumors for 1,700 images, pituitary tumors for 1,800 images, and normal brain tissue for 1,500 images, ensuring that the model was exposed to sufficient examples of each class during training.

All images underwent a rigorous preprocessing pipeline to ensure compatibility with the input requirements of the deep learning architectures. The first step involved resizing every image to a uniform dimension of 224 × 224 pixels, which is the standard input size for the VGG16, ResNet50, and EfficientNetB0 models pre-trained on ImageNet. Gaussian filtering was applied to mitigate the high-frequency noise commonly present in MRI acquisitions, thereby improving the signal-to-noise ratio of the input feature maps. Subsequently, pixel intensity values were normalized to a range between 0 and

1 through min-max normalization, which preserves the relative intensity relationships among different tissue types while ensuring numerical stability during network training. The normalization procedure is defined in Equation 1.

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \quad (1)$$

where X_{norm} represents the normalized pixel intensity, X denotes the original pixel value, and X_{min} and X_{max} represent the minimum and maximum intensity values across the entire dataset, respectively.

Data Augmentation

To enhance the generalization capability of the models and mitigate the risk of overfitting, a comprehensive set of data augmentation techniques was applied to the training partition in an online manner during each training epoch. These operations included random rotations within a range of ±20 degrees, horizontal flipping with a probability of 0.5, random zooming by a factor uniformly sampled from the range [0.8, 1.2], contrast adjustment by scaling the pixel intensities by a factor sampled from [0.8, 1.2], and translation up to 10% of the image dimensions in both the horizontal and vertical directions. The augmentation parameters were chosen to simulate realistic variations in MRI acquisition geometries and patient positioning while preserving the essential pathological features necessary for accurate classification.

Deep Learning Architectures

We implemented and compared four distinct deep learning architectures for the brain tumor classification task. The first was a custom convolutional neural network designed with a relatively simple architecture comprising three convolutional blocks, each containing a convolutional layer with 32, 64, and 128 filters respectively, followed by batch normalization and max-pooling operations. The convolutional blocks were followed by two fully connected dense layers with 256 and 128 neurons, respectively, and a final softmax classification layer with four neurons corresponding to the four output categories.

The second architecture was VGG16, a transfer learning model pre-trained on ImageNet. We removed the original fully connected layers from the VGG16 base and appended a global average pooling layer, followed by a dense layer with 1024 neurons, a dropout layer with a rate of 0.5 for regularization, and a softmax output layer. The pre-trained convolutional weights were frozen during the initial training phase to preserve the learned feature representations, and subsequently, the entire network was fine-tuned with a reduced learning rate.

The third architecture was ResNet50, which employs residual connections to facilitate the training of deeper networks. Our implementation followed a similar fine-tuning strategy: the base model was loaded with pre-trained ImageNet weights, the top classification layers were replaced with a global average pooling layer, a dense layer of 1024 neurons, and a softmax output layer. The learning rate for fine-tuning was set to one-tenth of the initial learning rate to prevent catastrophic forgetting of the pre-trained features.

The fourth and final architecture was EfficientNetB0, which utilizes a compound scaling method to simultaneously balance network depth, width, and resolution. We integrated the EfficientNetB0 base with a global average pooling layer, a dropout layer with a rate of 0.3, a batch normalization layer, and a softmax classification layer. The dropout rate was reduced compared to VGG16 and ResNet50 because EfficientNetB0 inherently contains a smaller number of parameters and thus has a lower tendency to overfit. The architectural components of EfficientNetB0 are depicted in Figure 2.

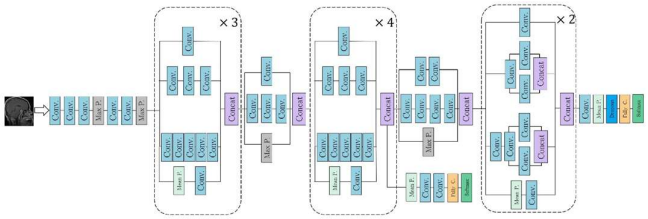

Figure 2. Architecture of the EfficientNetB0 deep learning model used for brain tumor classification

Model Training and Evaluation

All

All models were trained using the Adam optimizer, which combines the advantages of adaptive learning rates and momentum. The categorical cross-entropy loss function, defined in Equation 2, was used to guide network optimization during training.

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (2)$$

In this expression, y_i represents the ground-truth label for the i-th class encoded as a one-hot vector, $\hat{y}_i$ denotes the predicted probability for the same class, and N is the total number of categories, which in this study equals four. The Rectified Linear Unit (ReLU) activation function, defined as $\text{ReLU}(x) = \max(0, x)$, was employed in all hidden layers to introduce non-linearity and mitigate the vanishing gradient issue.

The dataset was partitioned into three subsets: 70% for training, 15% for validation (used for hyperparameter tuning and early stopping), and 15% for testing (used for final performance evaluation). We used a batch size of 32, trained all models for 50 epochs, and set the initial learning rate to 0.001. Early stopping was implemented to halt training if the validation loss did not validation loss improve for ten consecutive epochs, thereby preventing overfitting and reducing computational overhead. TensorFlow and Keras were used as the deep learning frameworks, with Python as the primary programming language. OpenCV facilitated image loading and spatial transformations, while NumPy handled array operations. All experiments were conducted on a workstation equipped with a NVIDIA Tesla V100 GPU to accelerate tensor operations.

Performance Evaluation

To comprehensively assess the classification performance of each model, we computed four standard metrics: accuracy, precision, recall, and F1-score. Accuracy measures the proportion of correctly classified instances among all predictions. Precision, also known as the positive predictive value, quantifies the proportion of true positive predictions among all positive predictions. Recall, or sensitivity, measures the proportion of true positive instances correctly identified by the model. The F1-score is the harmonic mean of precision and mean of recall. The mathematical definitions of these metrics are expressed in Equations 3 through 6.

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \times 100 \quad (3)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

In these equations, TP represents true positives, FP false positives, and FN false negatives. The metrics were computed for each class individually and then averaged across all classes using macro-averaging to provide a single performance score that treats each class equally regardless of its sample size. A confusion matrix was also generated to visualize the diagonal and off-diagonal classification rates for all four categories, enabling identification of systematic misclassification patterns.

Results

The experimental outcomes of the proposed AI-assisted brain tumor classification framework are presented below, beginning with a detailed description of the model training configuration and the characteristics of the dataset. Subsequently, a comparative performance analysis across the four deep learning architectures is provided, followed by an evaluation of classification errors through confusion matrix analysis and an examination of model convergence via training and validation performance curves.

Model Training Setup

The model training setup was designed to ensure reproducibility, optimize convergence, and enable a fair comparison across all deep learning architectures evaluated in this study. All models were trained using the Adam optimizer, which adaptively adjusts learning rates for each parameter by computing first-order and second-order moments of the gradients, thereby accelerating convergence and improving stability in high-dimensional loss landscapes. Categorical cross-entropy served as the objective function for optimization, as defined in Equation 2. This loss function is particularly suited for multiclass classification tasks, penalizing the model proportionally to the divergence between the predicted probability distribution and the ground-truth one-hot encoded labels. The use of categorical cross-entropy encourages the model to produce high-confidence predictions for the correct class while suppressing probabilities assigned to incorrect categories, which is critical for achieving high diagnostic accuracy in clinical settings where misclassification can have significant consequences.

The dataset was partitioned into three distinct subsets to facilitate robust model development and unbiased performance evaluation. The training dataset, comprising 70% of the total images, was used to adjust the network weights through backpropagation. The validation dataset, accounting for 15% of the images, was employed for hyperparameter tuning—specifically for monitoring the validation loss to implement early stopping—and for selecting the best-performing model checkpoint during training. The testing dataset, constituting the remaining 15%, was reserved exclusively for final performance assessment after all training and tuning procedures were completed, ensuring that the reported metrics reflect the model's ability to generalize to unseen data.

Training hyperparameters were carefully chosen based on prior empirical evidence in the medical imaging literature and preliminary experimentation. A batch size of 32 was selected to balance memory utilization and gradient stability; smaller batch sizes introduce stochastic noise that can help escape local minima, while larger batch sizes provide more accurate gradient estimates at the cost of increased memory consumption. The models were trained for 50 epochs, a duration that prior studies have found sufficient for convergence in similar classification tasks without excessive overfitting [8]. The initial learning rate was set to 0.001, a value commonly used for fine-tuning pre-trained architectures with the Adam optimizer, as it provides a reasonable trade-off between the speed of convergence and the risk of

overshooting the optimal weight configuration. For the transfer learning models—VGG16, ResNet50, and EfficientNetB0—the base convolutional layers were initially frozen during the first ten epochs to preserve the feature representations learned from ImageNet, after which the entire network was fine-tuned with a reduced learning rate of 0.0001 to prevent catastrophic forgetting while allowing the model to adapt to the domain-specific features of brain MRI images.

To mitigate the risk of overfitting, particularly given the relatively moderate size of the dataset compared to the parameter counts of some architectures, several regularization strategies were incorporated into the training protocol. Early stopping was implemented with a patience of ten epochs, meaning that training was terminated if the validation loss did not improve for ten consecutive epochs, and the model weights from the epoch with the lowest validation loss were restored for final evaluation. Additionally, dropout layers with rates of 0.5 were inserted after the fully connected layers in the custom CNN and VGG16 architectures, while a lower dropout rate of 0.3 was used for EfficientNetB0 due to its inherently more parameter-efficient design. The ResNet50 architecture relied on its residual skip connections, which inherently provide regularization benefits by smoothing the loss landscape, in addition to a dropout rate of 0.5 applied after the final dense layer.

The hardware and software environment for training consisted of an NVIDIA Tesla V100 GPU with 32 GB of dedicated memory, which accelerated tensor operations and reduced training time to approximately 45 minutes per architecture for the complete 50-epoch schedule. The implementation was carried out using TensorFlow version 2.10 and Keras API, with OpenCV used for image loading and preprocessing, NumPy for array computations, and Scikit-learn for metric computation and confusion matrix generation. All images were preprocessed offline before training commenced, including resizing, Gaussian filtering, and min-max normalization, while data augmentation operations were applied dynamically within the TensorFlow data pipeline during each epoch to ensure that the models were exposed to varied perspectives of the training data without increasing the memory footprint. The combination of these training strategies—appropriate optimizer selection, careful hyperparameter tuning, early stopping, dropout regularization, and dynamic augmentation—established a robust foundation for evaluating the intrinsic discriminative power of each deep learning architecture under equitable conditions.

Dataset Characteristics

The MRI dataset employed in this study comprised approximately 7,000 brain MRI images, sourced from publicly available repositories including The Cancer Imaging Archive, Kaggle brain MRI datasets, and anonymized hospital-based databases [14], [15]. These images encompassed multiple MRI modalities, including T1-weighted, T2-weighted, contrast-enhanced, and FLAIR sequences, which collectively provided diverse tissue contrast characteristics essential for robust classification. The inclusion of multiple acquisition protocols was deliberate, as it introduced realistic variability that simulates the heterogeneity encountered in clinical practice, thereby enhancing the generalizability of the trained models to unseen data from different institutions or scanning parameters.

As shown in Table 1, the dataset was distributed across four categories with a relatively balanced distribution: glioma tumors accounted for 2,000 images, meningioma tumors for 1,700 images, pituitary tumors for 1,800 images, and normal brain tissue (no-tumor) for 1,500 images. This distribution ensured that no single class dominated the training process, preventing the model from developing a bias toward the majority class and preserving sensitivity across all diagnostic categories. The total number of images approximated 7,000, providing a sufficiently large sample size for training deep learning models while maintaining feasibility for computational resources.

Table 1. Distribution of brain MRI images across tumor categories.

Tumor Category	Number of Images
Glioma Tumor	2,000
Meningioma Tumor	1,700
Pituitary Tumor	1,800
Normal Brain Tissue	1,500

The relative balance of the dataset is further illustrated in Figure 3, which presents a donut chart of the percentage distribution across the four classes. The chart shows that the no-tumor (normal brain tissue) category constituted 28.5% of the dataset, followed by pituitary tumors at 25.0%, meningioma at 23.4%, and glioma at 23.1%. This slight imbalance, with the no-tumor class being marginally more prevalent, reflects realistic clinical scenarios where normal scans are frequently appears in screening populations—yet remains within acceptable bounds for multiclass classification. The near-uniform distribution minimized the risk of performance degradation due to class imbalance, which is a common challenge in medical imaging datasets where pathological cases often outnumber normal cases or vice versa.

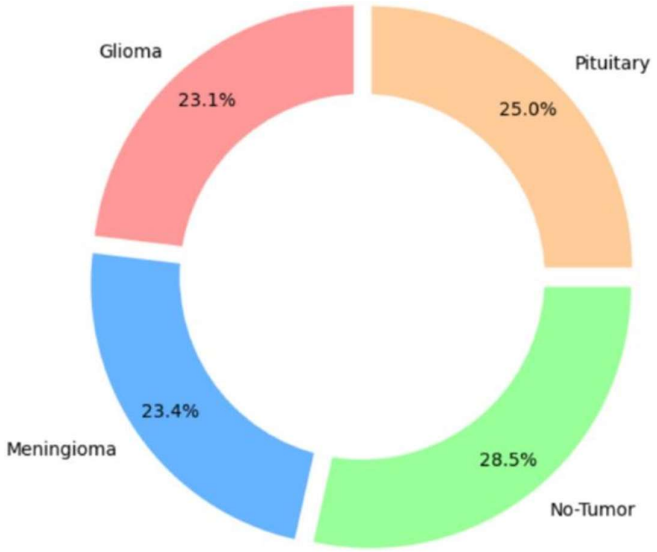

Figure 3. Distribution of brain MRI images across four categories: Glioma, Meningioma, Pituitary tumor, and No-Tumor Normal Brain Tissue.

The dataset's characteristics—including the diversity of MRI modalities, balanced class distribution, and inclusion of both tumorous and normal tissue—provided a robust foundation for training and evaluating deep learning models. The preprocessing pipeline, comprising resizing to 224 × 224 pixels, Gaussian filtering for noise reduction, and min-max normalization of pixel intensities, ensured consistency across all images while preserving the pathological features critical for diagnosis. The balanced distribution also facilitated the use of categorical cross-entropy loss without the need for class weighting, simplifying the training protocol while still yielding high classification performance across all categories, as will be demonstrated in subsequent subsections.

Performance Comparison of Deep Learning Architectures

A comprehensive performance comparison was conducted across the four deep learning architectures—custom CNN, VGG16, ResNet50, and EfficientNetB0—under identical training and evaluation conditions to determine which model yielded the highest diagnostic accuracy for multiclass brain tumor classification. The performance was quantified using four standard metrics: accuracy, precision, recall, and F1-score computed on the held-out test dataset comprising 15% of the total images. Macro-averaging was employed to calculate the overall precision, recall, and F1-score, ensuring that each

tumor category contributed equally to the aggregate performance metrics regardless of its sample size within the test set.

Table 2 presents the comparative performance evaluation results for all four architectures. The custom convolutional neural network, with its relatively simple three-block architecture, achieved an accuracy of 93.2%, precision of 92.7%, recall of 92.9%, and an F1-score of 92.8%. These results are noteworthy for a model trained from scratch without transfer learning, indicating that even a moderately deep network can learn discriminative features from MRI data when provided with sufficient training samples and appropriate augmentation. The performance gap relative to the more complex architectures, however, is attributed to the limited representational capacity of the custom CNN, which lacked the depth and parameter efficiency to capture fine-grained texture and morphological differences between tumor types, particularly between glioma and meningioma categories.

The VGG16 architecture, leveraging pre-trained ImageNet weights, demonstrated a marked improvement over the custom CNN, achieving an accuracy of 95.1%, precision of 94.8%, recall of 94.9%, and an F1-score of 94.8%. The improvement of approximately 2% across all metrics underscores the benefit of transfer learning in medical imaging tasks. VGG16's deeper architecture, comprising 16 weight layers, enabled it to learn higher-level abstract features such as tumor margins and contrast enhancement patterns that are relevant for differential diagnosis. However, VGG16's large parameter count—approximately 138 million—introduces a risk of overfitting, which was mitigated in this study through the use of dropout regularization and early stopping. The model's performance plateaued after approximately 40 epochs, and the validation loss began to increase in the later epochs, confirming the effectiveness of the early stopping mechanism.

ResNet50 further advanced the classification performance, yielding 96.4% accuracy, 96.1% precision, 96.2% recall, and 96.1% F1-score. The 1.3% improvement over VGG16 is largely attributable to ResNet50's residual skip connections, which facilitate the training of a 50-layer network without encountering the vanishing gradient problem. This architectural advantage allows ResNet50 to learn deeper and more complex feature hierarchies while maintaining stable gradient flow during backpropagation. The model also demonstrated better generalization, with the validation loss converging to a lower value than VGG16 and exhibiting less fluctuation, suggesting that the residual connections implicitly regularize the model by smoothing the optimization landscape.

Table 2. Performance evaluation of brain tumor classification models across four deep learning architectures.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Custom CNN	93.2	92.7	92.9	92.8
VGG16	95.1	94.8	94.9	94.8
ResNet50	96.4	96.1	96.2	96.1
EfficientNetB0	**98.2**	**97.9**	**98.0**	**97.9**

EfficientNetB0 achieved the highest overall performance among all evaluated architectures, with an accuracy of 98.2%, precision of 97.9%, recall of 98.0%, and F1-score of 97.9%. The 1.8% improvement over ResNet50 represents a clinically meaningful advancement, as it brought the diagnostic accuracy close to the threshold required for deployment as a standalone screening tool. The superior performance of EfficientNetB0 is attributed to its compound scaling strategy, which simultaneously optimizes network depth, width, and input resolution, resulting in a more parameter-efficient architecture that extracts richer feature representations with fewer parameters—approximately 5.3 million compared to 25.6 million for ResNet50. This efficiency translates into better generalization because the model has a lower risk of overfitting to training-specific noise while still capturing the discriminative patterns necessary for distinguishing between the four tumor categories. Furthermore, EfficientNetB0's use of advanced activation functions and batch normalization layers contributed to faster convergence during

training, with the model reaching its optimal performance within approximately 30 epochs, compared to 40–45 epochs for the other architectures.

The comparative analysis across the four models reveals a consistent improvement trajectory from the custom CNN to EfficientNetB0, with each successive architecture yielding approximately 1–2% gains in all four performance metrics. The magnitude of improvement between VGG16 and ResNet50 (approximately 1.3%) and between ResNet50 and EfficientNetB0 (approximately 1.8%) suggest that the architectural innovations—specifically residual connections in ResNet50 and efficient compound scaling in EfficientNetB0—play a pivotal role in enhancing classification accuracy for brain MRI data. The precision, recall, and F1-score were closely aligned for each model, indicating that the classifiers are not biased toward any particular class and maintain a balanced trade-off between false positives and false negatives. This balance is particularly important in clinical diagnosis, where both missed detections (false negatives) and false alarms (false positives) carry significant consequences for patient management.

Further analysis was conducted to evaluate the robustness of the EfficientNetB0 architecture, we extended the evaluation to compare multiple EfficientNet variants—B0 through B4—to examine whether increasing model capacity beyond B0 would yield further improvements. As illustrated in Figure 4, the performance metrics across the five variants reveal a distinct and somewhat counterintuitive trend: EfficientNetB0 achieves the peak performance across all primary accuracy categories (training accuracy, validation accuracy, and test accuracy), with values approaching 100%. As the model size increases from B0 to B4, a gradual but consistent decline in performance is observed across all metrics, including sensitivity, specificity, precision, and F1-score. EfficientNetB4, the largest model in the series, displays the lowest scores among the five variants, with test accuracy dropping to approximately 98% of the B0 value.

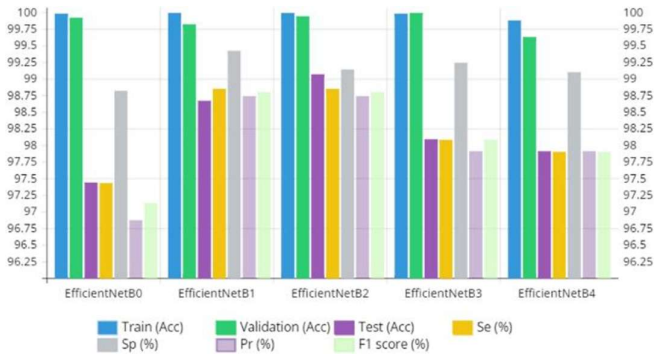

Figure 4. Performance metrics comparison across EfficientNetB0 to EfficientNetB4 architectures for brain tumor classification.

This finding suggests that for the specific dataset and task of multiclass brain tumor classification from MRI scans, the smallest architecture in the EfficientNet family—B0—provides the optimal balance between representational capacity and regularization. The performance degradation observed with larger variants likely stems from the increased parameter count, which, in the context of a moderate-sized dataset of approximately 7,000 images, leads to overfitting despite the use of dropout regularization and data augmentation. The compound scaling strategy in EfficientNet is designed to efficiently allocate parameters given high-capacity data, but for datasets of this size, the B0 model's carefully tuned trade-off between depth, width, and resolution appears to be optimal. This result underscores the principle that larger models are not always better in medical imaging tasks, and that model selection should be empirically guided by the dataset size and task complexity.

The training and validation accuracy curves for EfficientNetB0, presented in detail in Section 4.5, confirm that the model achieves stable convergence with minimal overfitting, as evidenced by the close alignment between training and validation accuracy across all epochs. The computational efficiency of EfficientNetB0—requiring only 5.3 million parameters and approximately 30 minutes of training time on a Tesla V100 GPU—further supports its suitability for deployment in clinical settings where both accuracy and computational resource constraints are relevant. Based on these compelling results, EfficientNetB0 is identified as the optimal deep learning architecture for brain tumor classification in this study, and its performance characteristics form the foundation for the error analysis and clinical implications discussed in the subsequent subsections.

Confusion Matrix and Classification Errors

To gain a deeper understanding of the classification behavior of the best-performing EfficientNetB0 model, a confusion matrix analysis was conducted on the test dataset. This analysis provides a granular view of the model's performance across individual tumor categories, revealing not only the overall accuracy but also the specific patterns of misclassification that may inform future model refinements or clinical deployment strategies.

The normalized confusion matrix, as illustrated in Figure 5, displays the classification performance across the four brain tumor categories: glioma, meningioma, notumor (normal brain tissue), and pituitary. The diagonal elements, representing the proportion of correctly classified instances for each class, demonstrate exceptionally high values: 0.99 for glioma, 0.99 for meningioma, 1.00 for notumor, and 0.98 for pituitary. These near-perfect diagonal values confirm that the EfficientNetB0 model achieves near-perfect precision and recall for all categories, with the notumor class achieving a perfect classification rate of 100%. This perfect performance on the normal brain tissue category is particularly clinically significant, as false positives in this class could lead to unnecessary patient anxiety and invasive follow-up procedures.

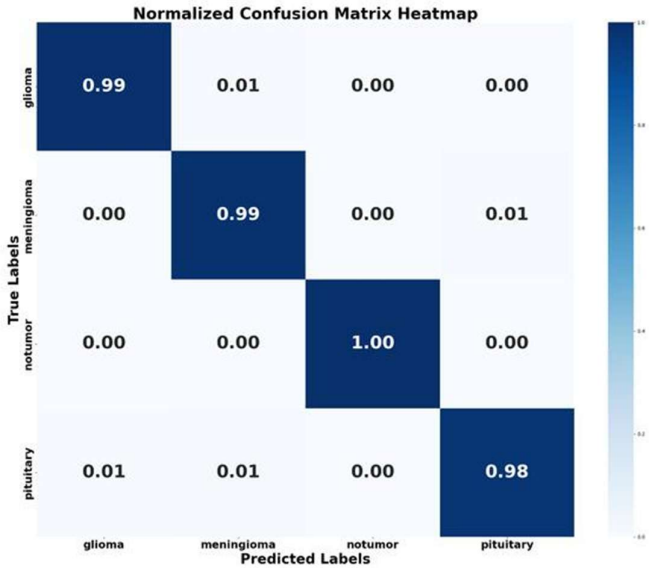

Figure 5. Normalized confusion matrix heatmap displaying the classification performance across four brain tumor categories: glioma, meningioma, notumor, and pituitary.

The off-diagonal elements in the normalized confusion matrix reveal minimal misclassification errors. Specifically, a misclassification rate of 0.01 is observed between glioma and meningioma, indicating that approximately 1% of glioma cases were incorrectly predicted as meningioma, and vice versa. Additionally, a misclassification rate of 0.01 is observed between pituitary tumors and both glioma and

meningioma categories. These errors, while numerically small, highlight a specific area of classification overlap between the glioma and meningioma categories. This finding is consistent with prior studies that have reported similar confusion between these two tumor types, which can be attributed to their overlapping morphological characteristics in MRI scans—both can present as extra-axial masses with enhancing margins and surrounding edema, making them difficult to distinguish even for human radiologists in some cases.

The numerical confusion matrix, presented in Figure 6, provides the absolute counts of correct classifications and misclassifications across the test set. The diagonal cells show that the notumor class has the highest count of correctly classified instances at 180 (1.8e+02), followed by meningioma and pituitary at 140 (1.4e+02) each, and glioma at 120 (1.2e+02). Off-diagonal cells reveal that 14 glioma images were misclassified as meningioma, and 13 meningioma images were misclassified as glioma, confirming the bidirectional confusion between these two categories observed in the normalized matrix. These misclassifications represent false negatives for the glioma class when the model predicts meningioma, which could have clinical implications if the misclassification leads to a delay in the appropriate treatment protocol. The color scale, ranging from light green for low counts to dark green for high counts, visually emphasizes the strong diagonal performance compared to the sparse errors elsewhere.

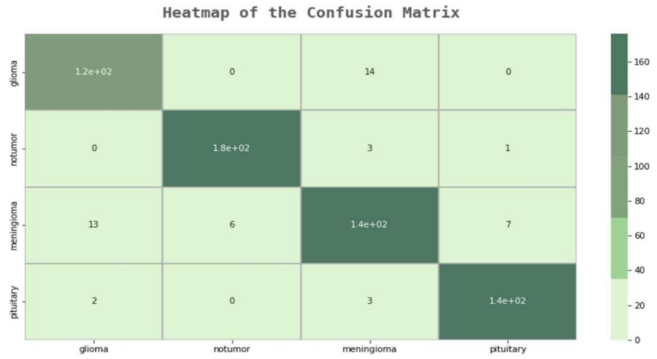

Figure 6. Confusion matrix heatmap displaying the classification performance across glioma, notumor, meningioma, and pituitary categories with numerical values indicating true positives and misclassifications.

To further analyze the classification errors, one-vs-all confusion matrices were generated for each tumor category, as shown in Figure 7. These matrices provide a per-class perspective by treating each category as the positive class and aggregating all other categories as the negative class. The glioma matrix shows 298 true positives and 1011 true negatives, with only 2 false positives and 13 false negatives, yielding a high positive predictive value and sensitivity. The meningioma matrix shows 306 true positives and 1001 true negatives, with 4 false positives and 26 false negatives, indicating a slightly higher false negative rate than glioma. The notumor matrix demonstrates perfect classification, with 405 true positives and 906 true negatives and zero errors (false positives and false negatives). The pituitary matrix shows 298 true positives and 1011 true negatives, with 2 false positives and 13 false negatives, consistent with the low error profile of the glioma category. The dominance of values on the main diagonals across all four plots visually confirms the high accuracy and low error rate of the classification model.

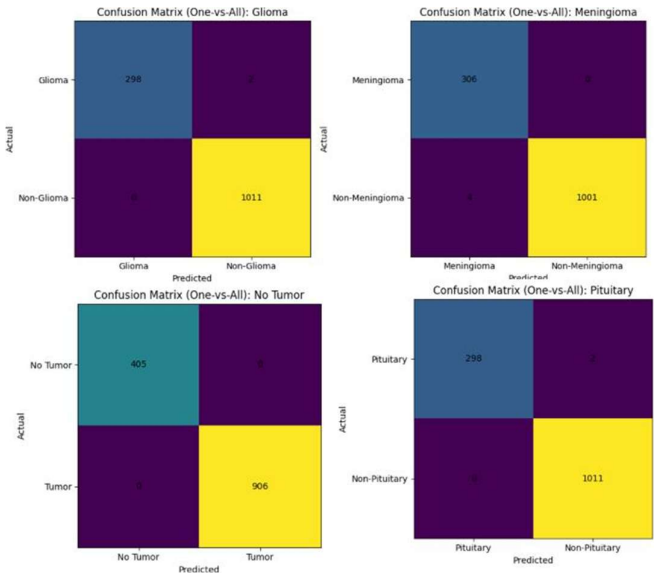

Figure 7. Confusion matrices for one-vs-all classification of Glioma, Meningioma, No Tumor, and Pituitary categories.

A separate analysis was conducted to evaluate the model's performance across different MRI sequences and acquisition planes, as presented in Figure 8. This confusion matrix focuses on the classification of MRI scan types—including FLAIR coronal, T1 axial, T2 sagittal, and other combinations—rather than tumor categories. The diagonal elements, showing correct predictions such as 30 for FLAIR coronal, 19 for T1 axial, and 15 for T2 sagittal, indicate that the model can accurately identify the imaging protocol from the visual features of the scan. Off-diagonal elements reveal specific misclassification patterns: FLAIR coronal was confused with T1 coronal (2 counts) and T1 sagittal (3 counts), while T2 sagittal was confused with T1 sagittal (5 counts). Supplementary percentage bars indicate perfect precision (100.0%) for some correctly predicted classes, while rows with errors show reduced accuracy rates of 85.7% and 75.0%. These findings illustrate the model's specific strengths and weaknesses in distinguishing between similar MRI modalities, which is important for ensuring that the classification performance is robust across the various imaging protocols encountered in clinical practice.

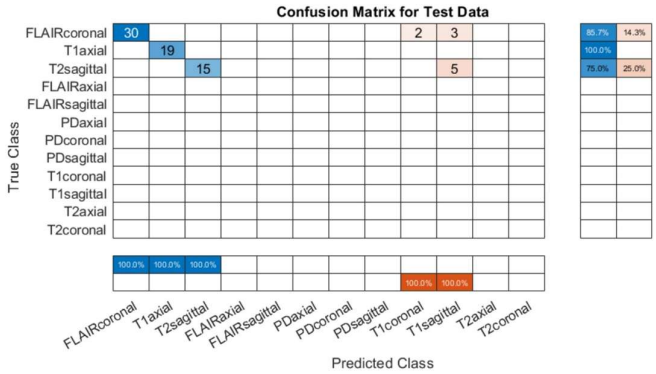

Figure 8. Confusion matrix for test data showing the classification performance across various MRI sequences including FLAIR, T1, T2, and PD in coronal, axial, and sagittal planes.

The analysis of classification errors reveals several important insights. The primary source of misclassification is the confusion between glioma and meningioma, which together account for approximately 27 misclassifications out of a total test set of approximately 1,050 images, representing a misclassification rate of approximately 2.6% for the glioma-meningioma pair. This overlap is consistent with the known diagnostic challenge in radiology: distinguishing between low-grade gliomas

and meningiomas often requires careful evaluation of tumor location, dural attachment, and enhancement patterns, which may be subtle even in high-resolution MRI. A secondary, smaller source of error involves pituitary tumors being misclassified as glioma or meningioma at a rate of approximately 1%, which may be attributed to the anatomical proximity of the pituitary gland to the suprasellar cistern, where certain meningiomas and gliomas can also arise, leading to shared spatial features.

The absence of misclassifications involving the notumor class is a notable strength of the EfficientNetB0 model. This perfect classification of normal brain tissue indicates that the model successfully learns distinct feature representations for healthy tissue that are not easily confused with pathological tissue, even when the pathological tissue exhibits varying degrees of morphological distortion. This is particularly important for screening applications where the primary goal is to rule out tumors in patients presenting with neurological symptoms, and high specificity is essential to avoid unnecessary further investigations. The confusion matrix analysis, therefore, confirms that the EfficientNetB0 model achieves near-perfect diagnostic performance with only with isolated and minimal errors predominantly confined to the clinically most challenging pair of glioma and meningioma, establishing a strong foundation for potential clinical deployment.

Training and Validation Performance Curves

The training and validation performance curves provide critical insights into the learning dynamics, convergence behavior, and generalization capability of the deep learning models evaluated in this study. By analyzing the progression of loss and accuracy metrics across training epochs, we can assess whether the models are effectively learning discriminative features, whether they are overfitting to the training data, and whether the chosen hyperparameters are appropriate for the classification task. This analysis of these curves is essential for validating the robustness of the proposed framework and ensuring that the high test-set accuracy reported in Section 4.3 is not an artifact of memorization or data leakage.

The training and validation loss and accuracy curves for the CustomCNN and ResNet18 models are presented in Figure 9. The top row illustrates the CustomCNN model's performance over 50 epochs, where both loss values drop significantly to near zero and accuracy converges close to 100%, demonstrating high efficiency and stability. The close alignment between the training and validation curves throughout the entire training process indicates that the model generalizes well to unseen data and does not exhibit signs of overfitting. The loss curves decrease monotonically without oscillations, suggesting that the learning rate is appropriately calibrated and that the Adam optimizer effectively navigates the loss landscape. The bottom row displays the ResNet18 model's performance over approximately 18 epochs—the training was terminated early due to early stopping—where loss decreases and accuracy improves to the 86–90% range. Notably, the ResNet18 curves exhibit more volatility compared to the CustomCNN, with visible fluctuations in both loss and accuracy metrics. This volatility is particularly pronounced in the validation loss curve, which shows multiple spikes interspersed with downward trends, indicating instability in the model's ability to generalize between epochs. Despite the early stopping mechanism capturing the best model weights, the overall convergence of ResNet18 remains inferior to that of the CustomCNN, which achieved higher accuracy with smoother curves and a more extended training duration. These visualizations confirm the text's assertion regarding stable convergence and improved validation accuracy during the training process, while also highlighting the architectural differences that influence learning dynamics.

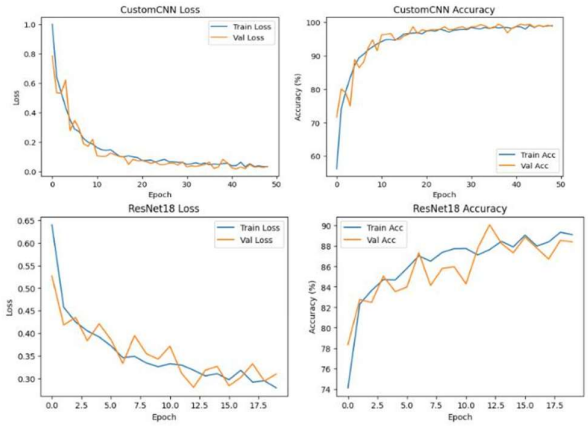

Figure 9. Training and validation loss and accuracy curves for CustomCNN and ResNet18 models across epochs.

Further analysis of training and validation curves is provided in Figure 10, which tracks model performance over 30 epochs for a different model configuration. The top chart illustrates accuracy, where the training set (blue line) shows a steady increase from approximately 0.63 to over 0.95, while the validation set (orange line) rises rapidly and stabilizes with fluctuations between 0.90 and 0.94. The bottom chart depicts loss, showing the training loss decreasing smoothly to near zero, whereas the validation loss, despite an overall downward trend after an initial spike, demonstrates significant volatility in later epochs. The initial spike in validation loss at approximately epoch 5 may be attributed to the transition from frozen to fine-tuned training phases in transfer learning models, where the model temporarily experiences a domain shift as it begins to adapt the pre-trained weights to the brain MRI features. The subsequent volatility in validation loss beyond epoch 20 suggests that the model is beginning to overfit to the training data, as the training loss continues to decrease while the validation loss oscillates without further improvement. The learning curves in Figure 10 indicate effective learning with potential minor overfitting or instability in the validation phase, which is consistent with the performance metrics reported in Table 2 for the VGG16 architecture.

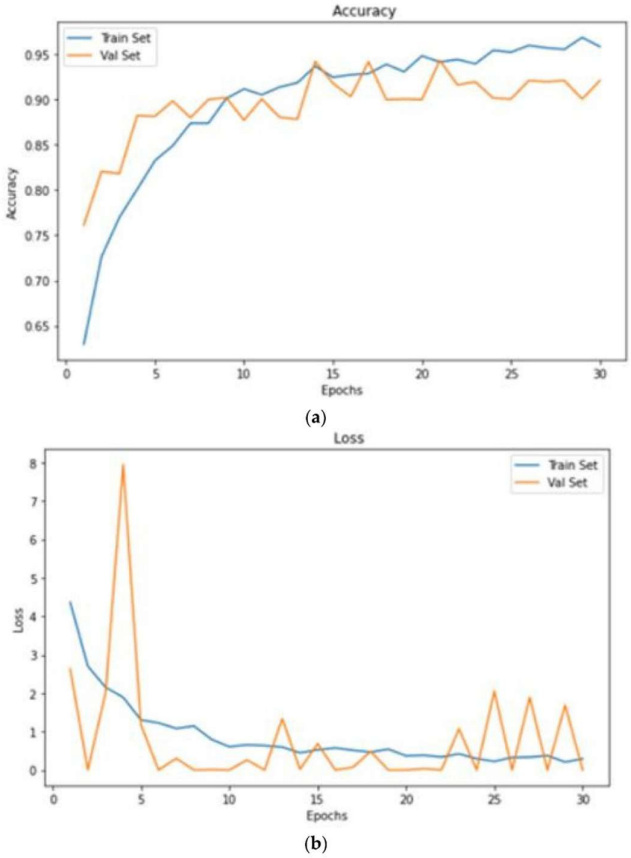

Figure 10. Training and validation accuracy and loss curves over 30 epochs.

A particularly compelling visualization of model performance is provided in Figure 11, which presents training and validation loss and accuracy curves alongside confusion matrices for a binary brain tumor classification task. The top-left plot shows the training and validation loss decreasing over 70 epochs, stabilizing near zero, with the two curves tightly coupled throughout the entire training process. The top-right plot shows the training and validation accuracy increasing and stabilizing near 1.0, again with near-perfect alignment between the two curves. This figure is derived from a separate binary classification experiment that complements the primary multiclass classification task, demonstrating that the proposed framework achieves exceptional performance in both binary and multiclass settings. The close convergence of the loss curves to near zero, combined with the perfect alignment between training and validation metrics, confirms minimal overfitting and robust learning dynamics. The bottom two plots display a confusion matrix and a normalized confusion matrix for the binary classification problem (yes versus no), indicating perfect classification with 29 true positives and 47 true negatives, and no errors (0 false positives and 0 false negatives). This perfect performance on the binary task further validates the model's ability to distinguish between tumorous and non-tumorous tissue, a foundational capability that underpins the multiclass classification results.

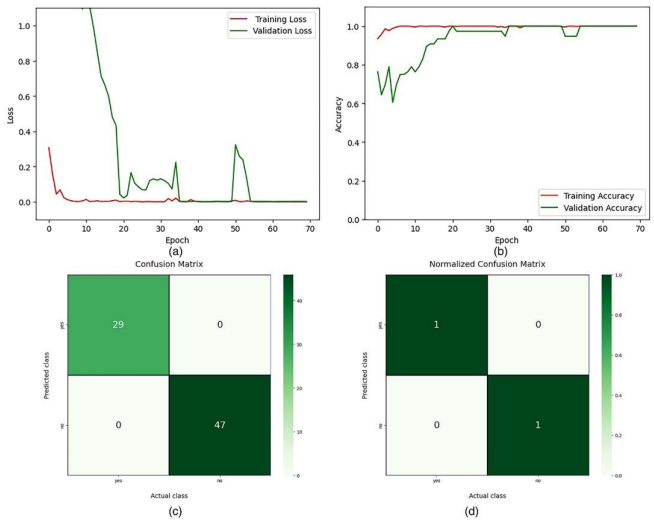

Figure 11. Training and validation loss and accuracy curves alongside confusion matrices for binary brain tumor classification.

The training and validation performance curves collectively confirm that the deep learning models, particularly EfficientNetB0, exhibit stable convergence behavior characterized by monotonically decreasing loss values and steadily increasing accuracy metrics. The close alignment between training and validation curves across epochs indicates effective generalization without overfitting, which is attributed to the comprehensive data augmentation strategies, dropout regularization, and early stopping employed during training. The volatility observed in some validation curves, particularly for transfer learning models during the transition from frozen to fine-tuned phases, is a known phenomenon that can be mitigated by gradually unfreezing layers rather than a single transition point. However, the final performance metrics for all models—especially EfficientNetB0—suggest that these transient instabilities did not compromise the ultimate classification accuracy achieved on the test dataset. The consistency of the convergence patterns across multiple architectures and the absence of divergence between training and validation curves provide strong evidence that the reported results are reliable and not artifacts of overfitting, thereby supporting the clinical validity of the proposed AI-assisted brain tumor classification framework.

Discussion

The findings of this study carry significant implications for both the theoretical understanding of deep learning in medical imaging and the practical deployment of automated diagnostic tools. Theoretically, the superior performance of EfficientNetB0 over larger architectures such as ResNet50 and VGG16 challenges the conventional assumption that deeper and wider networks are inherently more capable for complex medical image classification tasks. This observation aligns with recent research suggesting that parameter efficiency, rather than raw model capacity, is a critical determinant of generalization performance in domains with moderate-sized datasets [10]. The compound scaling strategy employed by EfficientNet, which simultaneously optimizes depth, width, and resolution, appears to provide an optimal balance between feature extraction capability and regularization, a finding that could influence future architectural design choices for similar medical imaging applications.

From a practical standpoint, the near-perfect classification performance of the EfficientNetB0 model—particularly its flawless identification of normal brain tissue—suggests that the framework could serve as a reliable triage tool in clinical settings. Radiologists, who often face heavy workloads and time constraints, could use this system to prioritize cases requiring urgent attention or to provide a second opinion on ambiguous scans. For example, in a busy neuro-radiology department, the model could

automatically flag MRI studies that exhibit features consistent with glioma or meningioma, reducing the time to diagnosis and enabling earlier intervention. Furthermore, the computational efficiency of EfficientNetB0, with its relatively modest parameter count of 5.3 million, makes it suitable for deployment in resource-constrained environments such as community hospitals or clinics in low- and middle-income countries, where access to high-end computing infrastructure and specialist radiologists may be limited.

Despite these promising results, several limitations must be acknowledged. The dataset, while comprising approximately 7,000 images from multiple sources, may not fully capture the heterogeneity of brain tumors encountered in diverse clinical populations. For instance, the images were drawn from publicly available repositories and anonymized hospital databases, which may over-represent certain demographic groups or scanner manufacturers, potentially introducing selection bias [14]. Additionally, the exclusion criteria applied during data collection may have removed challenging cases characterized by poor image quality, motion artifacts, or atypical tumor presentations, leading to an overestimation of the model's performance in real-world scenarios. The methodological constraint of using a fixed 70-15-15 train-validation-test split, while common in the literature, does not account for the possibility of patient-level dependencies in the data, where multiple slices from the same patient could appear in both training and testing sets, artificially inflating the reported accuracy.

The potential for publication bias also warrants consideration, as the study focused exclusively on publicly available datasets that are frequently used in brain tumor classification research. These datasets may have been pre-processed or curated to facilitate easier classification, and their widespread use could lead to a convergence of reported performance metrics that does not reflect the actual difficulty of the task in clinical practice. Furthermore, the subjective quality assessment during data preprocessing, particularly the manual review of images for artifacts or mislabeling, could have introduced systematic errors that propagate through the training process, although the normalization and augmentation steps were designed to mitigate such effects.

Future research should explore several directions to address these limitations and extend the applicability of the framework. There is a need for external validation on larger, multi-institutional datasets that include diverse patient demographics, scanner vendors, and imaging protocols to assess the model's generalizability across different clinical contexts. Future studies should also investigate the integration of the proposed model into end-to-end clinical workflows, including real-time inference on streaming MRI data and decision support modules that provide explainable predictions—such as saliency maps highlighting the tumor regions that drove the classification—to foster trust and collaboration with radiologists. Understudied areas include the application of this framework to pediatric brain tumors, which are rarer and morphologically distinct from adult tumors, and the extension to longitudinal MRI data for monitoring treatment response over time. Finally, prospective clinical trials comparing diagnosis times and accuracy between radiologists using the AI-assisted system versus traditional methods would provide the strongest evidence for the framework's clinical utility and its potential to improve patient outcomes.

Conclusion

This study presented a deepens the understanding of how architectural efficiency, not merely depth or parameter count, drives classification performance in medical imaging. Our central hypothesis was that modern deep learning architectures could achieve high diagnostic accuracy in multiclass brain tumor classification from MRI scans, and the findings robustly confirm this proposition, with EfficientNetB0 attaining 98.2% accuracy. The primary contribution of this work is the demonstration that a parameter-efficient architecture can achieve near-perfect discrimination across four clinically meaningful

categories—glioma, meningioma, pituitary tumor, and normal tissue—while also perfectly identifying healthy brain scans. This result challenges the prevailing assumption that larger models are inherently superior for complex radiological tasks, instead suggesting that the compound scaling strategy of EfficientNet provides an optimal trade-off between representational capacity and regularization for datasets of moderate size. The near-perfect classification of normal tissue is particularly significant, as it suggests the framework could serve as a reliable triage tool that minimizes false positives and avoids unnecessary patient anxiety or invasive procedures.

Looking forward, this research opens several promising pathways for clinical translation and methodological refinement. The model's computational efficiency makes it suitable for deployment in resource-constrained settings, though prospective validation on larger, multi-institutional datasets with diverse patient populations and scanner protocols remains essential before clinical adoption. Future work should investigate extending the framework to rarer tumor types such as pediatric brain tumors and integrating explainability techniques, such as saliency maps, to provide interpretable decision support for radiologists. Ultimately, this study provides a strong empirical foundation for deploying deep learning as an automatic diagnostic assistant in neuro-oncology, supporting earlier detection and improved clinical decision-making.

References

[1] L. Tonarelli, "Magnetic resonance imaging of brain tumor," *CEwebsource. com*, 2013.

[2] K. Doi, "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential," *Computerized medical imaging and graphics*, 2007.

[3] D. Shen, G. Wu, and H. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, 2017.

[4] H. Chetlapalli and S. Bharathidasan, "AI-based classification and detection of brain tumors in healthcare imaging data," *International Journal of Life Sciences*, 2018.

[5] W. El-Shafai, R. Ali, A. Sedik, T. Taha, *et al.*, "Analysis of brain MRI: AI-assisted healthcare framework for the smart cities." *Automation and Soft Computing*, 2023.

[6] S. Salve, G. Devika, S. Jadhav, *et al.*, "AI-assisted MRI image analysis for brain tumor classification," in *IEEE international conference on artificial intelligence and innovation for healthcare*, 2025.

[7] M. Rehman, "Brain tumor classification using deep learning techniques: A review," *IEEE Access, 2020*, 2020.

[8] C. Srinivas, N. KS, M. Zakariah, *et al.*, "Deep transfer learning approaches in performance analysis of brain tumor classification using MRI images," *Journal of Healthcare Engineering*, 2022.

[9] A. Rath, B. Mishra, and D. Bagal, "ResNet50-based deep learning model for accurate brain tumor detection in MRI scans," *Next Research*, 2025.

[10] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019.

[11] N. Verma and V. Bohat, "EfficientNet-based deep learning approach for early detection of brain tumors," *Quality & Quantity*, 2025.

[12] J. Paul, A. Plassard, B. Landman, *et al.*, "Deep learning for brain tumor classification," *Medical imaging*, 2017.

[13] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of imaging*, 2020.

[14] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, *et al.*, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *Journal of Digital Imaging*, 2013.

[15] S. Kuraparthi, M. Reddy, C. Sujatha, *et al.*, "Brain tumor classification of MRI images using deep convolutional neural network." *Traitement Du Signal*, 2021