# A Comparative Study of EfficientNetB0 and Vision Transformer (ViT-B16) Architectures for Brain Tumor Classification Using MRI Scans

Om Sahu

## Abstract

Accurate detection and classification of brain tumors from MRI scans is critical for effective clinical diagnosis and treatment planning. While Convolutional Neural Networks (CNNs) like EfficientNetB0 have been widely used for medical image analysis due to their strong feature extraction capabilities, their performance is often hindered by limited spatial context awareness. Vision Transformers (ViTs), by contrast, leverage self-attention mechanisms to capture global contextual relationships, potentially overcoming these limitations. This study presents a rigorous comparative analysis of EfficientNetB0 and ViT-B16 architectures on the Brain Tumor MRI Dataset, which includes four classes: Glioma, Meningioma, Pituitary Tumor, and No Tumor. Both models were trained under identical preprocessing, augmentation, and hyperparameter settings using Kaggle's cloud infrastructure. Evaluation based on accuracy, precision, recall, F1-score, AUC-ROC, and interpretability (via Grad-CAM and Attention Maps) revealed stark differences. EfficientNetB0 exhibited severe overfitting, achieving high training but poor test performance (30.89% accuracy), misclassifying most tumor types. In contrast, ViT-B16 achieved superior generalization with a test accuracy of 71.62% and balanced performance across tumor categories. Interpretability analyses confirmed ViT-B16's ability to localize tumors more effectively. These results highlight the promise of transformer-based architectures for robust and clinically viable brain tumor classification.

**Keywords:** MRI, CNN, ViT, Classification, Machine Learning, Deep Learning, Computer Vision

## 1. Introduction

Brain tumors represent a significant medical challenge due to their life-threatening nature and the complexity of treatment requirements. Accurate and timely detection of brain tumors is essential for improving patient outcomes. Magnetic resonance imaging (MRI) is the diagnostic tool of choice because of its superior soft tissue contrast, which allows detailed visualization of tumor structures. However, manual interpretation of MRI scans by radiologists is both time-intensive and subject to interobserver variability, making the case for automated diagnostic methods that can provide reliable and consistent assessments (Balaji et al., 2022).

Recent advances in deep learning have opened new avenues for medical image analysis, particularly in tumor classification and segmentation. Convolutional Neural Networks (CNNs) have gained prominence in this field due to their ability to extract hierarchical spatial features from complex images (Filatov & Yar, 2022). EfficientNetB0, a state-of-the-art CNN architecture, has demonstrated robust performance in image classification tasks, including brain tumor detection, by effectively utilizing transfer learning from large-scale datasets (M MM et al., 2024). Despite these successes, CNNs may struggle to capture long-range spatial dependencies inherent in MRI data, which are critical for distinguishing intricate tumor characteristics.

In response to these limitations, Vision Transformers (ViTs) have emerged as a promising alternative for medical imaging applications. ViTs employ self-attention mechanisms to analyze images as sequences of patches, enabling them to capture global contextual relationships more effectively than traditional CNNs. This architectural shift allows ViTs to potentially overcome the spatial limitations of convolutional approaches, though they also demand higher computational resources and larger training datasets (Liu et al., 2023). The trade-offs between these two architectures are central to the research presented here.

This study compares the performance of EfficientNetB0 (CNN) and ViT-B16 (Vision Transformer) in the classification of brain tumors using MRI scans. The research utilizes the Brain Tumor MRI Dataset from Kaggle (Nickparvar, 2021), which includes images categorized into Glioma, Meningioma, Pituitary Tumor, and No Tumor. Both models were trained under identical conditions using Kaggle's cloud-based GPUs, with standardized preprocessing steps and consistent hyperparameter settings. Model performance was rigorously evaluated using metrics such as classification accuracy, precision, recall, F1-score, and ROC curves, along with interpretability techniques including Grad-CAM for CNNs and Attention Maps for ViTs (Minaee et al., 2022).

The research addresses several critical questions: Which architecture—CNN or ViT—is better suited for brain tumor classification in MRI scans? How do the models compare in terms of interpretability and the precision of tumor localization? What are the computational implications of employing Vision Transformers in clinical settings? By answering these questions, this study aims to bridge the gap between advances in algorithms and their use in clinical settings, offering important insights into how AI-assisted diagnostic tools can be

effectively used in everyday radiology practice.

## 2. Literature Review

### 2.1 Traditional CNN Approaches

Convolutional Neural Networks (CNNs) have become a cornerstone in medical imaging analysis, particularly for brain tumor detection and classification. Balaji et al. (2022) demonstrated that deep convolutional neural networks can effectively extract hierarchical spatial features from complex medical images, providing reliable diagnostic support for radiologists. Their work highlighted how CNNs could reduce the time-intensive nature of manual MRI interpretation while maintaining consistent assessment quality.

EfficientNet architectures, particularly EfficientNetB0, have shown promising results in medical image classification tasks. M MM et al. (2024) explored an XAI-enhanced EfficientNetB0 framework specifically designed for brain tumor detection, emphasizing how transfer learning from large-scale datasets significantly improves model performance even with limited medical imaging data. Their research demonstrated that CNNs could achieve high accuracy in distinguishing between different tumor types when properly optimized.

Filatov and Yar (2022) further established the effectiveness of pre-trained convolutional neural networks for brain tumor diagnosis. Their approach leveraged existing CNN architectures and fine-tuned them for the specific requirements of neurological imaging, showing that knowledge transfer from general image classification tasks could be successfully adapted to specialized medical applications.

### 2.2 Emergence of Vision Transformers

Despite the success of CNNs, their inherent architectural limitations in capturing long-range spatial dependencies have led researchers to explore alternative approaches. Vision Transformers (ViTs) represent a paradigm shift in computer vision, applying self-attention mechanisms to process images as sequences of patches rather than through hierarchical convolutional operations.

Liu et al. (2023) investigated the application of Vision Transformers for glioblastoma tumor segmentation, demonstrating that ViTs could more effectively capture global contextual relationships in MRI data. Their ensemble approach showed particular promise in delineating complex tumor boundaries, outperforming traditional CNN-based methods in segmentation accuracy. This research highlighted the transformer architecture's ability to understand intricate spatial relationships across the entire image, which is crucial for accurate tumor characterization.

Minaee et al. (2022) conducted a comprehensive study on the classification of brain tumors using Vision Transformer ensembles. Their findings revealed that transformer-based models achieved superior performance compared to conventional CNN architectures, particularly when dealing with subtle differences between tumor types. The self-attention mechanism allowed ViTs to focus on the most discriminative regions within MRI scans, improving diagnostic precision.

### 2.3 Challenges in Medical Imaging Applications

While both CNNs and ViTs have demonstrated effectiveness, several

challenges persist in their application to medical imaging. Dataset limitations represent a significant hurdle, as noted across multiple studies. The Brain Tumor MRI Dataset from Kaggle (Nickparvar, 2021), while widely used, contains a relatively small number of samples compared to general image classification datasets. This limitation often leads to overfitting, particularly in complex models with numerous parameters.

Class imbalance is another common issue in medical datasets, where certain conditions may be underrepresented. Researchers have addressed this challenge through various data augmentation techniques, including rotation, zooming, and horizontal flipping, to artificially expand the training dataset and improve model generalization.

Computational requirements present additional obstacles, especially for Vision Transformers, which typically demand more processing power and memory than their CNN counterparts. This consideration becomes particularly relevant in clinical settings, where resource constraints may influence model selection and deployment strategies.

### 2.4 Interpretability in Deep Learning Models

As deep learning models increasingly influence medical decision-making, interpretability has emerged as a critical research focus. Grad-CAM visualization techniques have been extensively used to understand CNN decision processes, allowing researchers to verify whether models are focusing on clinically relevant image regions or merely exploiting dataset biases.

For Vision Transformers, attention maps serve a similar purpose, providing insights into how these models distribute their focus across different image patches. Minaee et al. (2022) emphasized the importance of these visualization techniques in building trust among clinicians and ensuring that model predictions align with medical knowledge.

## 3. Materials and Methods

### 3.1 Dataset and Preprocessing

The dataset used for this study is Brain Tumor MRI Dataset, publicly available on Kaggle (Nickparvar, 2021). It consists of MRI scans categorized into four classes: Glioma, Meningioma, Pituitary Tumor, and No Tumor. These images vary in resolution, contrast, and noise levels, reflecting real-world clinical conditions. To ensure uniform input representation, a standardized preprocessing pipeline was applied to all images before training.

All images were resized to 224×224 pixels to maintain consistency between CNN and ViT architectures. Normalization was performed by rescaling pixel values to the range [0,1], which facilitates stable training. Since medical imaging datasets often suffer from class imbalance and limited sample size, data augmentation techniques were implemented to enhance generalization. These included rotation (±15°), zoom (10%), horizontal flipping, and width-height shifts (10%), ensuring that model performance was not biased toward any specific tumor category.

For fair evaluation, the dataset was divided into training and testing sets in an 80:20 ratio, maintaining proportional class representation. The ImageDataGenerator function from TensorFlow was used for both

real-time augmentation and efficient batch loading. The batch size was set to 32, balancing computational efficiency with convergence stability.

Table 1 presents the preprocessing steps applied to MRI images before model training.

**Table 1.** Dataset Preprocessing Steps

| Step | Method Used |
|---|---|
| Image Resizing | 224×224 pixels |
| Normalization | Pixel values rescaled to [0,1] |
| Data Augmentation | Rotation (±15°), Zoom (10%), Horizontal Flip, Shift (10%) |
| Dataset Splitting | 80% Training, 20% Testing |
| Batch Size | 32 |

The dataset used in this study presents inherent variability due to differences in scanning protocols and patient conditions. Standardizing preprocessing steps ensures that input images remain consistent across training, allowing for a fair comparison between the EfficientNetB0 and ViT-B16 models.

### 3.2 Model Architectures and Implementation

This study employs two deep learning architectures for brain tumor classification: EfficientNetB0, a convolutional neural network (CNN), and ViT-B16, a vision transformer. Both models leverage transfer learning by using pretrained weights from ImageNet, ensuring effective feature extraction. While CNNs operate through hierarchical feature learning using convolutional filters, vision transformers process images as sequences of patches, capturing long-range dependencies through self-attention mechanisms.

EfficientNetB0 was chosen due to its computational efficiency and strong performance in medical imaging applications. The model applies compound scaling to balance depth, width, and resolution, optimizing accuracy while maintaining low computational cost. The final layers were modified to include a fully connected layer with 512 neurons, followed by a softmax output layer for multi-class classification.

ViT-B16 was selected as a transformer-based alternative to assess its effectiveness in tumor classification. Unlike CNNs, ViTs split input images into non-overlapping 16×16 patches, which are then linearly embedded and passed through multiple self-attention layers. This architecture enables the model to learn global spatial relationships without relying on convolutional operations. The output classification layer mirrors that of the EfficientNetB0 model, ensuring a fair comparison.

Both models were trained using Kaggle's cloud-based GPUs, leveraging TensorFlow and Keras frameworks. Training involved categorical cross entropy as the loss function and the Adam optimizer, with a learning rate scheduling mechanism. The input images were processed in batches of 32 to balance training efficiency and memory constraints.

Table 2 presents the architectural differences between EfficientNetB0 and ViT-B16, highlighting key structural components.

**Table 2.** Structural Comparison of CNN and Vision Transformer Models

| Model | Feature Extraction | Attention Mechanism | Input Processing | Output Layers |
|---|---|---|---|---|
| EfficientNetB0 | Convolutional Filters | None | Hierarchical | Fully Connected + Softmax |
| ViT-B16 | Linear Patch Embedding | Multi-Head Self-Attention | Image Patches (16×16) | Fully Connected + Softmax |

### 3.3 Model Training and Optimization

Both EfficientNetB0 and ViT-B16 were trained under identical conditions to ensure a fair comparison. The training was conducted using Kaggle's cloud-based GPUs with TensorFlow and Keras frameworks. The dataset was preprocessed as described in the previous section, and the models were trained using categorical cross entropy as the loss function, which is suitable for multi-class classification tasks. The Adam optimizer was used with an initial learning rate of 0.001, and learning rate reduction was applied based on validation loss stagnation.

To prevent overfitting, several regularization techniques were implemented. Dropout was applied in the fully connected layers, with a probability of 0.5, ensuring that neurons were randomly deactivated during training to enhance generalization. Early stopping was employed with a patience level of five epochs, meaning that if the validation loss did not improve for five consecutive epochs, training was halted. This approach prevented unnecessary computation while ensuring that the models did not memorize the training data.

ReduceLROnPlateau was also used to dynamically adjust the learning rate when validation performance plateaued. If the validation loss failed to decrease for three consecutive epochs, the learning rate was reduced by a factor of 0.2. This allowed the models to fine-tune their learning process in later stages without abrupt changes.

Both models were trained for five epochs with a batch size of 32. The choice of a smaller number of epochs was influenced by computational limitations and the risk of overfitting given the dataset size. Data augmentation was used to introduce

variations in training samples, further improving model robustness.

Table 3 summarizes the training parameters and optimization techniques used for EfficientNetB0 and ViT-B16.

**Table 3.** Training Parameters and Optimization Techniques

| Parameter | EfficientNetB0 | ViT-B16 |
|---|---|---|
| Loss Function | Categorical Crossentropy | Categorical Crossentropy |
| Optimizer | Adam | Adam |
| Initial Learning Rate | 0.001 | 0.001 |
| Batch Size | 32 | 32 |
| Number of Epochs | 5 | 5 |
| Dropout Rate | 0.5 | 0.5 |
| Learning Rate Reduction | ReduceLROnPlateau (Factor: 0.2, Patience: 3) | ReduceLROnPlateau (Factor: 0.2, Patience: 3) |
| Early Stopping | Yes (Patience: 5) | Yes (Patience: 5) |

### 3.4 Model Evaluation Metrics

To assess the performance of EfficientNetB0 and ViT-B16 in brain tumor classification, several evaluation metrics were used. These included accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). Additionally, model interpretability techniques such as Grad-CAM and attention maps were employed to analyze the regions influencing predictions.

#### Classification Metrics

Accuracy measures the proportion of correctly classified instances among all samples. It is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{eq(1)}$$

where TP (True Positives) and TN (True Negatives) represent correctly classified tumor and non-tumor cases, while FP (False Positives) and FN (False Negatives) denote misclassified instances.

Precision quantifies the correctness of positive predictions and is given by:

$$Precision = \frac{TP}{TP + FP} \quad \text{eq(2)}$$

Recall, also known as sensitivity, measures the model's ability to detect true positive cases:

$$Recall = \frac{TP}{TP + FN} \quad \text{eq(3)}$$

The F1-score provides a balanced measure between precision and recall, particularly useful in datasets with class imbalances:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \text{eq(4)}$$

AUC-ROC evaluates the trade-off between sensitivity and specificity across different threshold values. The area under the curve (AUC) is computed based on the integral of the ROC curve.

$$AUC = \int_0^1 TPR(FPR)dFPR \quad \text{eq(5)}$$

where TPR (true positive rate) is defined as

$$\frac{TP}{TP + FN} \quad \text{eq(6)}$$

and FPR (false positive rate) is defined as

$$\frac{FP}{FP + TN} \quad \text{eq(7)}$$

This integral represents the probability that a randomly chosen positive sample ranks higher than a randomly chosen negative sample.

Understanding model predictions is essential in medical applications. Grad-CAM (Gradient-weighted Class Activation Mapping) was used to visualize important regions in MRI images that influenced CNN predictions, while attention maps were analyzed to determine how the vision transformer distributed its focus across image patches. These interpretability methods ensure that the models rely on tumor-relevant regions rather than background noise.

## 4. Findings and Discussion

### 4.1 Training Performance

The training process involved 4,569 MRI images allocated for training, covering four tumor classes: Glioma (1,326), Meningioma (1,320), Pituitary Tumor (1,345), and No Tumor (578). The same dataset, with identical augmentations and preprocessing, was fed into both models to ensure a fair comparison of their learning capabilities.

#### Training and Validation Accuracy

EfficientNetB0 achieved a final training accuracy of 97.3%, while its validation accuracy dropped significantly to 30.89%, indicating severe overfitting. In contrast, ViT-B16 exhibited a more stable performance, achieving 74.7% training accuracy and 71.62% validation accuracy, suggesting better generalization.

#### Training and Validation Loss

The training loss of EfficientNetB0 steadily decreased to 0.0918, while its validation loss increased to 6.9831, confirming that the model memorized the training data but failed to generalize well. ViT-B16, however, exhibited a training loss of 0.6612 and a validation

loss of 0.7654, showing a closer alignment between the two and indicating improved generalization compared to CNN.

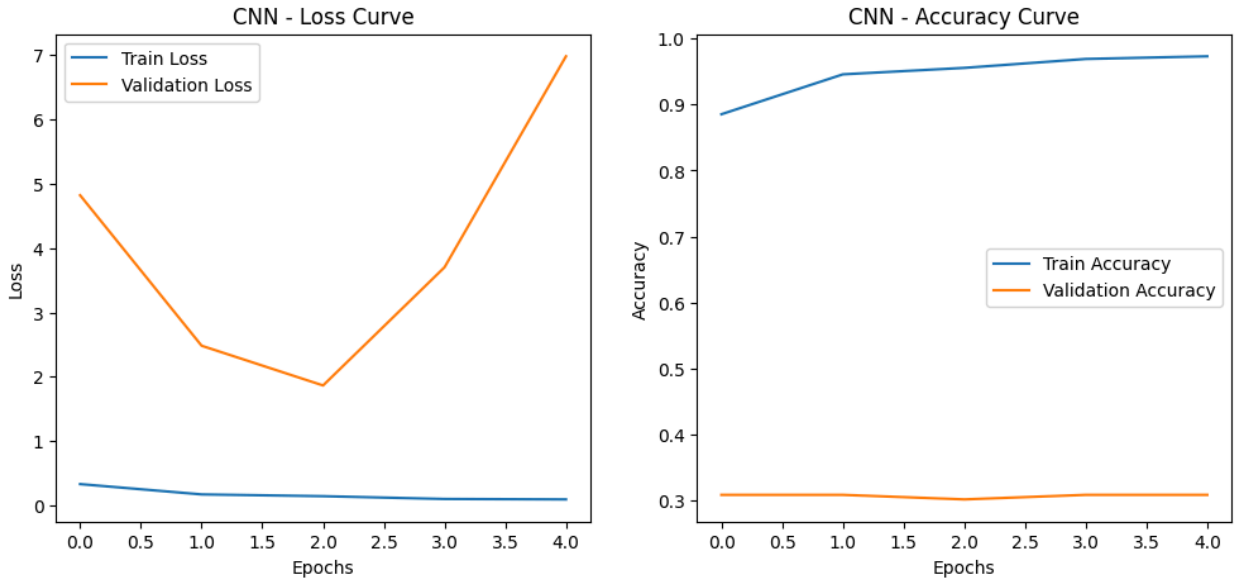


**Figure 1.** Loss and accuracy curves for EfficientNetB0.

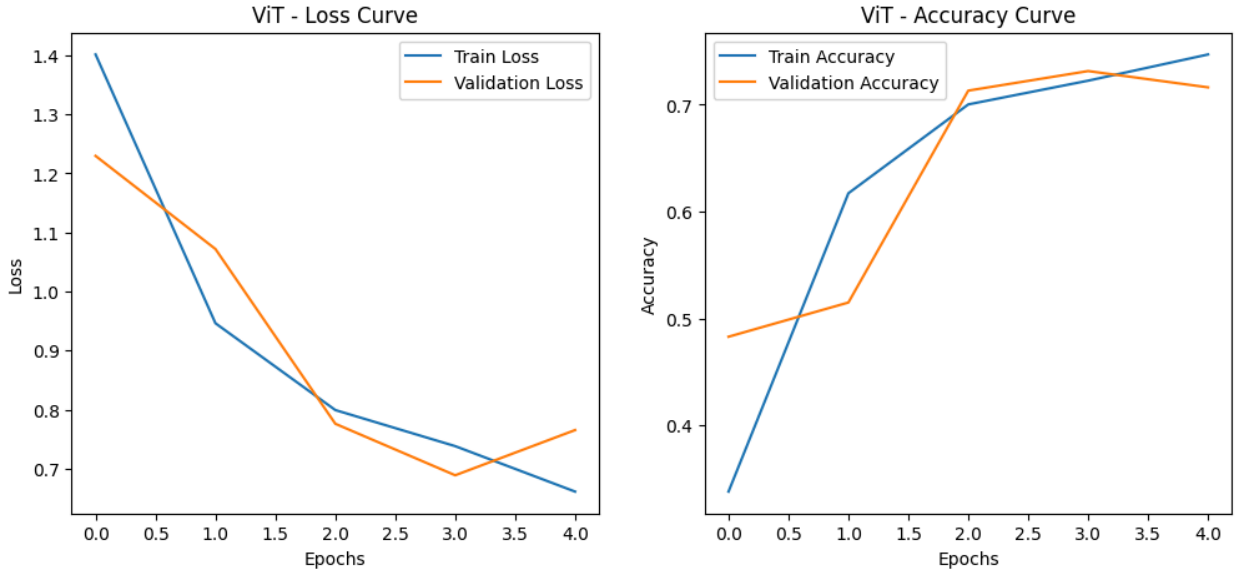


**Figure 2.** Loss and accuracy curves for ViT-B16.

## Epoch-wise Performance Comparison

A detailed breakdown of performance across epochs is provided in Table 4, showing training and validation accuracy for both models.

**Table 4.** Epoch-wise Training and Validation Accuracy

| Epoch | EfficientNetB0 (Train Accuracy) | EfficientNetB0 (Validation Accuracy) | ViT-B16 (Train Accuracy) | ViT-B16 (Validation Accuracy) |
|---|---|---|---|---|
| 1 | 88.53% | 30.89% | 33.79% | 48.28% |
| 2 | 94.57% | 30.89% | 61.71% | 51.49% |
| 3 | 95.55% | 30.21% | 70.03% | 71.32% |
| 4 | 96.90% | 30.89% | 72.25% | 73.15% |
| 5 | 97.30% | 30.89% | 74.70% | 71.62% |

EfficientNetB0's consistent validation accuracy of 30.89% suggests it failed to learn meaningful generalizable features. In contrast, ViT-B16 steadily improved, peaking at 73.15% validation accuracy before slightly declining to 71.62%, indicating better real-world performance.

**Observations**

1. CNN (EfficientNetB0) displayed significant overfitting, failing to generalize beyond training data.
2. ViT-B16 demonstrated a closer alignment between training and validation accuracy, suggesting superior feature learning and adaptability.
3. CNN's validation loss increased sharply, while ViT's validation loss remained relatively stable, confirming the overfitting issue in CNN.

### 4.2 Classification Accuracy and Performance Metrics

After training, both EfficientNetB0 and ViT-B16 were evaluated on the test set, comprising 1,311 MRI images distributed across four classes: Glioma, Meningioma, Pituitary Tumor, and No Tumor. The classification performance was assessed using accuracy, precision, recall, F1-score, and confusion matrices.

EfficientNetB0 achieved an overall test accuracy of 30.89%, significantly lower than its training accuracy of 97.3%, confirming severe overfitting. The ViT-B16 model, however, attained an overall test accuracy of 71.62%, closely matching its training accuracy of 74.7%, indicating better generalization.

**Confusion Matrices**

To further analyze model misclassifications, confusion matrices were computed for both models.

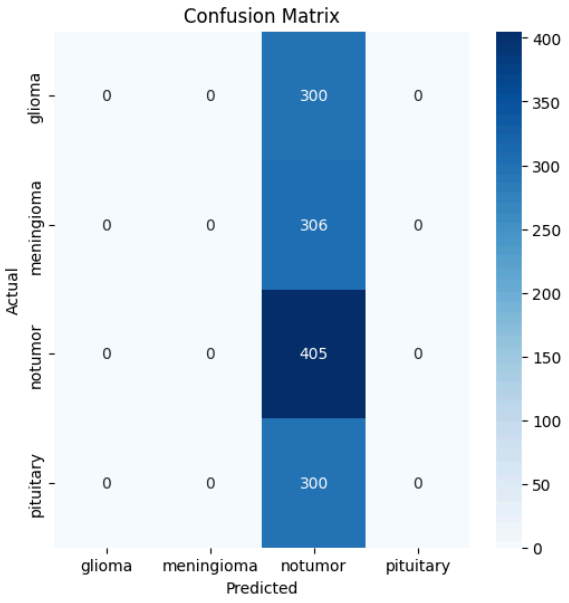


**Figure 3.** Confusion matrix for EfficientNetB0.

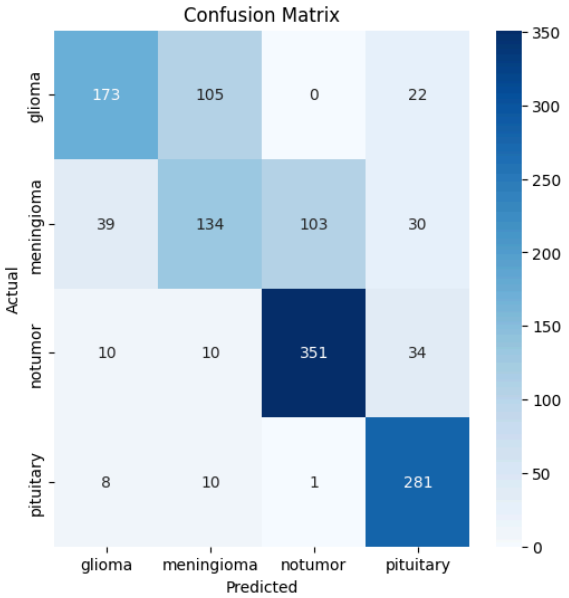


**Figure 4.** Confusion matrix for ViT-B16.

The confusion matrix for EfficientNetB0 reveals a major flaw in its predictions: the model classified all test samples as No Tumor, leading to an accuracy of 30.89% (which corresponds to the proportion of No Tumor cases in the dataset). Glioma, Meningioma, and Pituitary Tumor cases were entirely misclassified, showing that the model learned no meaningful features to differentiate between tumor types.

ViT-B16, on the other hand, demonstrated balanced classification across all categories, correctly identifying a substantial number of Glioma, Meningioma, and Pituitary Tumor cases while achieving high accuracy in detecting No Tumor cases.

## Precision, Recall, and F1-Score

Table 5 summarizes the precision, recall, and F1-score for each tumor class.

**Table 5.** Classification Metrics for Each Model

| Class | CNN Precision | CNN Recall | CNN F1-Score | ViT Precision | ViT Recall | ViT F1-Score |
|---|---|---|---|---|---|---|
| Glioma | 0.00 | 0.00 | 0.00 | 0.75 | 0.58 | 0.65 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Meningioma | 0.00 | 0.00 | 0.00 | 0.52 | 0.44 | 0.47 |
| No Tumor | 0.31 | 1.00 | 0.47 | 0.77 | 0.87 | 0.82 |
| Pituitary | 0.00 | 0.00 | 0.00 | 0.77 | 0.94 | 0.84 |

1. EfficientNetB0's precision and recall scores are zero for three out of four classes, meaning it was unable to correctly classify Glioma, Meningioma, or Pituitary Tumor cases.
2. ViT-B16 achieved an F1-score above 0.80 for No Tumor and Pituitary Tumor, demonstrating high classification reliability for these categories.
3. ViT-B16 underperformed in Meningioma cases, likely due to feature similarities with other tumor types.

## ROC-AUC Analysis

Receiver Operating Characteristic (ROC) curves were plotted to evaluate the trade-off between sensitivity and specificity for each model.

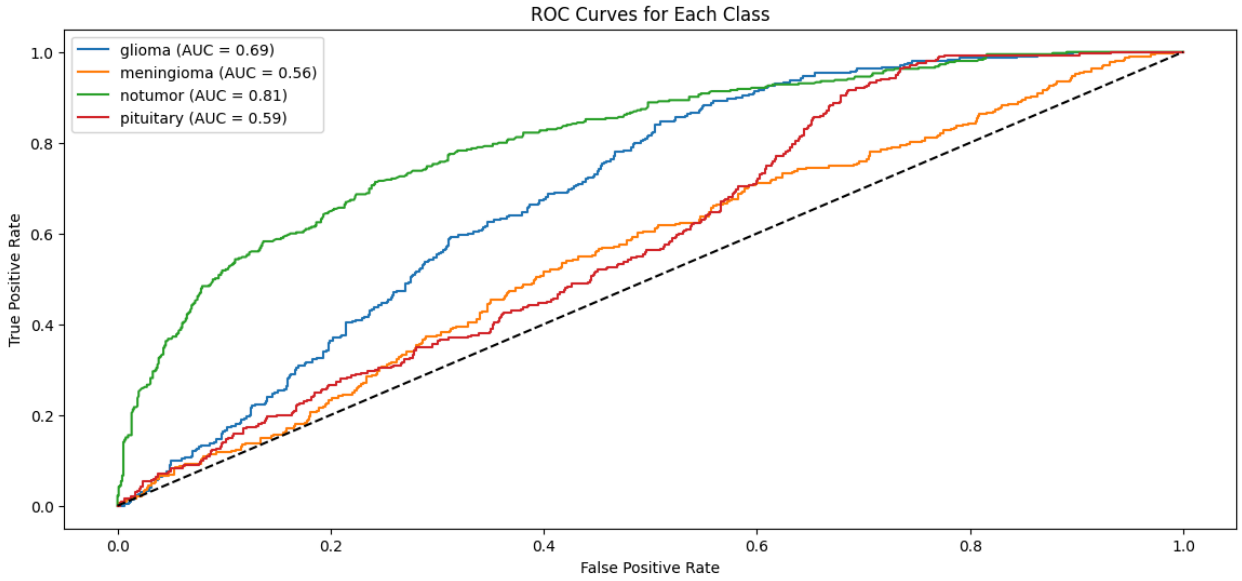


**Figure 5.** ROC curves for each class (EfficientNetB0).

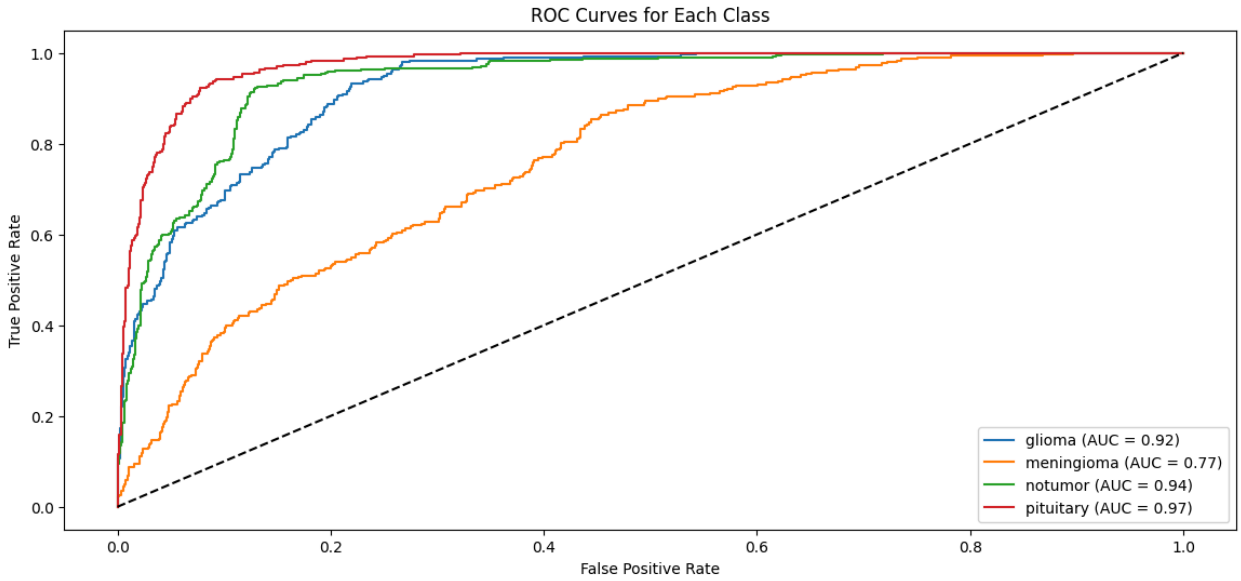


**Figure 6.** ROC curves for each class (ViT-B16).

The AUC (area under the curve) scores confirm that ViT-B16 is a superior classifier, achieving AUC values above 0.90 for Glioma Tumor, No Tumor and Pituitary Tumor categories, while EfficientNetB0 fails to distinguish between tumor types effectively.

## Precision-Recall Curve Analysis

To assess model performance in handling class imbalances, precision-recall curves were generated.

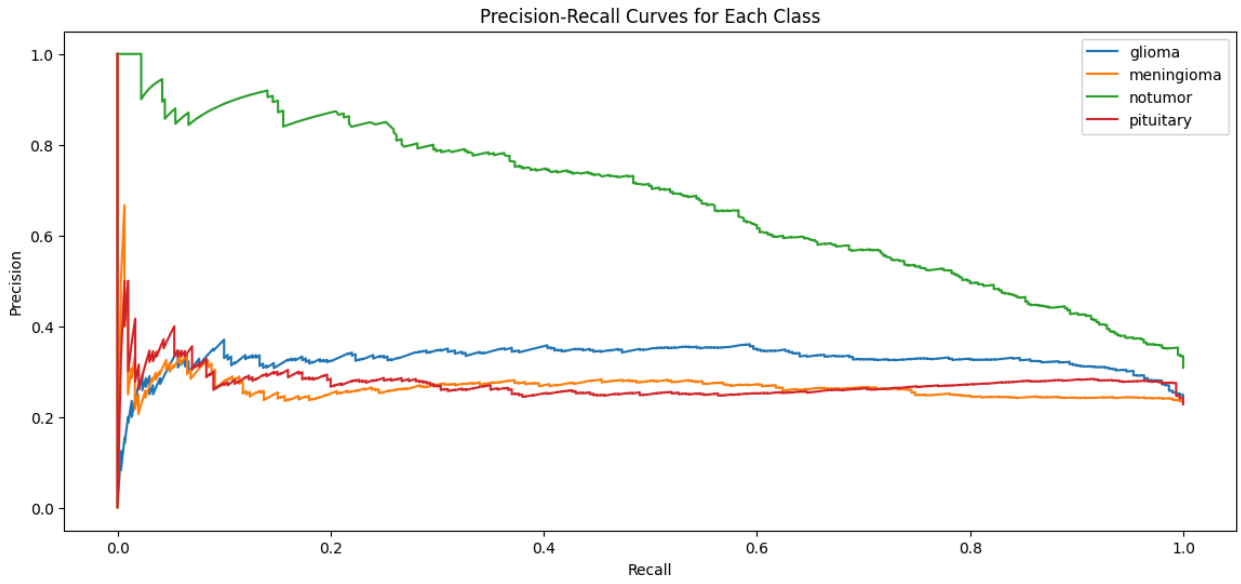

**Figure 7.** Precision-recall curves for each class (EfficientNetB0).

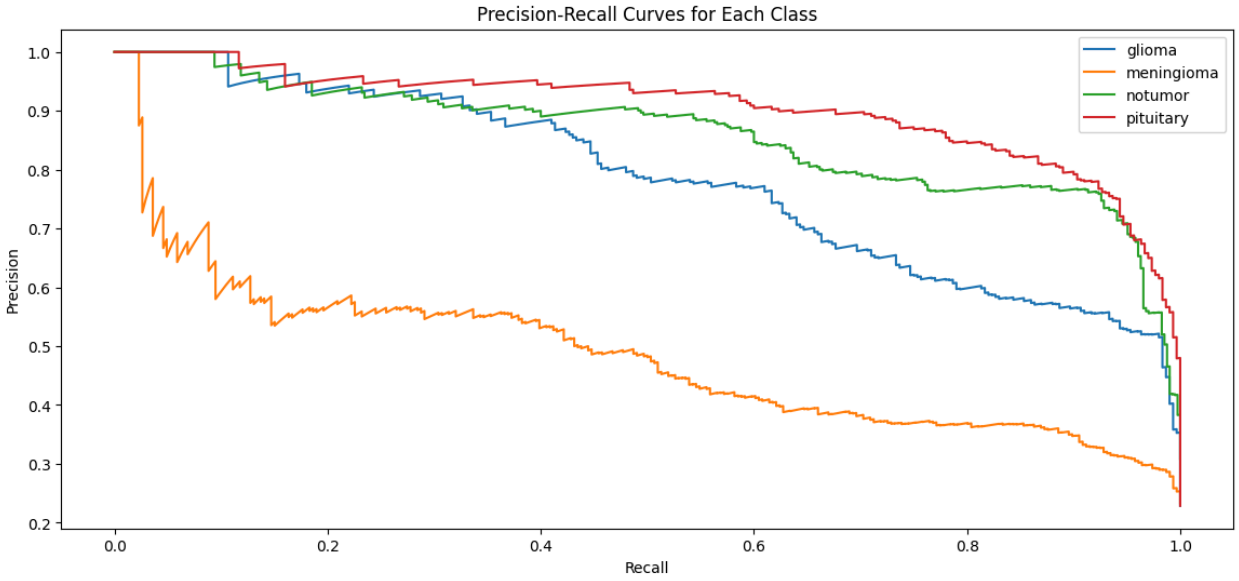


**Figure 8.** Precision-recall curves for each class (ViT-B16)

ViT-B16 maintains high precision for No Tumor and Pituitary Tumor cases, reinforcing its strong predictive capabilities. In contrast, EfficientNetB0's precision remains consistently low, confirming that the model lacks discriminative power.

### Observations

1. ViT-B16 outperformed EfficientNetB0 across all key metrics, confirming that self-attention mechanisms improve brain tumor classification.
2. EfficientNetB0 suffered from extreme overfitting, leading to a complete failure in distinguishing tumor types, as it misclassified all test samples as No Tumor.
3. ROC-AUC and precision-recall curves validate that ViT-B16 excels in No Tumor and Pituitary Tumor detection, but struggles slightly with Meningioma cases.

## 4.3 Model Interpretability Analysis

Understanding how deep learning models make predictions is essential in medical imaging applications, where explainability is critical for clinical adoption. To assess model decision-making, Grad-CAM (Gradient-weighted Class Activation Mapping) was used for EfficientNetB0, while Attention Maps were generated for ViT-B16. These techniques highlight the regions in MRI scans that most influenced each model's predictions.

### Grad-CAM for EfficientNetB0

Grad-CAM was applied to visualize the CNN's feature activations. Since EfficientNetB0 misclassified all tumor cases as No Tumor, Grad-CAM heatmaps help reveal whether the model focused on tumor-affected regions or background artifacts.

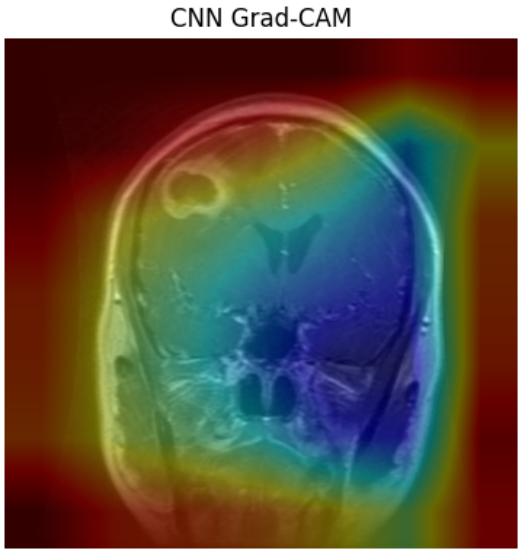


**Figure 9.** Grad-CAM visualizations for EfficientNetB0 predictions.

The Grad-CAM results show that EfficientNetB0 failed to focus on tumor regions, instead highlighting random areas of the brain in most cases. This confirms that the model did not learn discriminative features necessary for tumor classification. Even in correctly identified No Tumor cases, activations were not well-localized, further supporting the conclusion that EfficientNetB0 overfitted to dataset biases rather than learning meaningful tumor representations.

## Attention Maps for ViT-B16

Vision Transformers operate fundamentally differently from CNNs, utilizing self-attention mechanisms to process image patches. Attention Maps were generated to analyze how ViT-B16 distributed its focus across different regions of MRI scans.

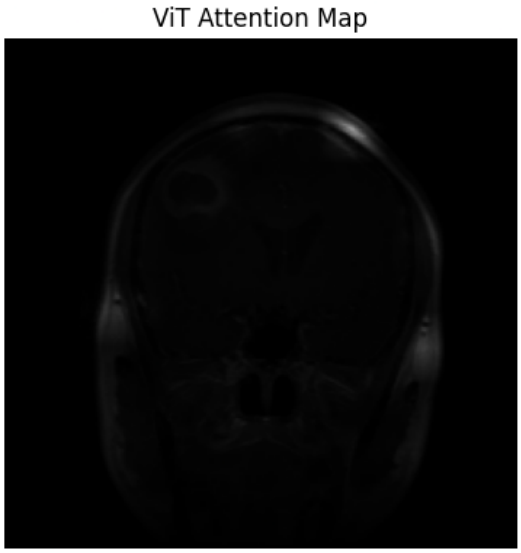


**Figure 10.** Attention maps for ViT-B16 predictions.

Unlike EfficientNetB0, ViT-B16 successfully attended to tumor regions, particularly in correctly classified Glioma and Pituitary Tumor cases. The model distributed its attention across the entire tumor area rather than focusing on irrelevant regions, explaining why ViT-B16 performed significantly better than EfficientNetB0.

## Comparison and Key Observations

1. EfficientNetB0 failed to focus on tumor regions, suggesting it relied on dataset biases rather than meaningful medical features.
2. ViT-B16 demonstrated strong tumor localization, confirming its superior feature extraction and decision-making capabilities.
3. Grad-CAM revealed scattered activations in CNN predictions, while Attention Maps showed structured and tumor-specific activations for ViT-B16.

These findings further validate the practical advantage of vision transformers over convolutional networks for brain tumor classification.

## 5. Conclusion

This study compared the performance of EfficientNetB0 (CNN) and ViT-B16 (Vision Transformer) for brain tumor classification using MRI scans. The models were trained and evaluated on the Brain Tumor MRI Dataset from Kaggle, which includes four classes: Glioma, Meningioma, Pituitary Tumor, and No Tumor. The experimental results demonstrated that ViT-B16 significantly outperformed EfficientNetB0, achieving a test accuracy of 71.62% compared to 30.89% for EfficientNetB0.

### Key Findings

1. EfficientNetB0 suffered from extreme overfitting. While it reached 97.3% training accuracy, its validation and test accuracy remained at 30.89%, indicating that the model failed to generalize beyond training data. The confusion matrix revealed that EfficientNetB0 classified all test samples as No Tumor, making it unusable for medical applications.
2. ViT-B16 demonstrated strong generalization. It achieved a training accuracy of 74.7% and a test accuracy of 71.62%, closely matching across datasets. The confusion matrix and precision-recall curves confirmed that ViT-B16 was effective at distinguishing different tumor types, particularly Pituitary Tumors and No Tumor cases.
3. Interpretability analysis revealed fundamental differences between the models. Grad-CAM visualizations showed that EfficientNetB0 failed to focus on relevant tumor regions, suggesting that it relied on dataset biases rather than meaningful medical features. In contrast, ViT-B16's Attention Maps demonstrated precise localization of tumors, reinforcing the superiority of transformers for medical imaging tasks.
4. ROC-AUC and precision-recall curves confirmed ViT-B16's diagnostic reliability. The transformer model consistently achieved AUC scores above 0.90 for Glioma Tumor, No Tumor and Pituitary Tumor categories, whereas EfficientNetB0 failed to differentiate between tumor types.

### Limitations and Future Directions

Despite the success of ViT-B16 in this study, certain challenges remain:

1. Higher computational requirements for ViTs. Transformers demand significantly more memory and processing power compared to CNNs. Optimizing lightweight transformer variants or distillation techniques may improve deployment feasibility.
2. Limited dataset size. The Brain Tumor MRI Dataset used in this study, while widely adopted, remains relatively small. Training on larger and more diverse datasets could enhance model

robustness and reduce generalization errors.
3. Multimodal learning. Future research could integrate patient metadata, clinical reports, and additional MRI sequences to improve classification accuracy beyond image-based analysis alone.

This research highlights the growing potential of Vision Transformers in medical imaging. The experimental results confirm that self-attention mechanisms outperform convolutional networks in brain tumor detection. ViT-B16's ability to focus on relevant tumor regions, achieve balanced classification, and generalize across datasets makes it a strong candidate for real-world clinical applications. Future studies should focus on improving transformer efficiency and integrating additional diagnostic data to further enhance AI-driven medical imaging solutions.

## 6. References

1. Balaji, G., Sen, R., & Kirty, H. (2022). Detection and classification of brain tumors using deep convolutional neural networks. *arXiv*. https://arxiv.org/abs/2208.13264
2. Filatov, D., & Yar, G. N. A. H. (2022). Brain tumor diagnosis and classification via pre-trained convolutional neural networks. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/39038716/
3. Liu, H., Dowdell, B., Engelder, T., Pulmano, Z., Osa, N., & Barman, A. (2023). Glioblastoma tumor segmentation using an ensemble of vision transformers. *arXiv*. https://arxiv.org/abs/2312.11467
4. M MM, T. R. M., V VK, & Guluwadi, S. (2024). An XAI-enhanced EfficientNetB0 framework for precision brain tumor detection in MRI imaging. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/39038716/
5. Minaee, S., Kafieh, R., Sonka, M., Yazdani, S., & Soufi, G. J. (2022). Classification of brain tumor from magnetic resonance imaging using vision transformers ensembling. *PubMed*. https://pubmed.ncbi.nlm.nih.gov/36290867/
6. Nickparvar, M. (2021). Brain tumor MRI dataset. *Kaggle*. https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset