A Comparative Study of Deep Learning Models for Brain Tumor Classification Using MRI Scans

Talha Aslam
Department of Data Science
FAST-NUCES, Islamabad Campus
i248067@isb.nu.edu.pk

Muhammad Ahmer
Department of Science
FAST-NUCES, Islamabad Campus
i247602@isb.nu.edu.pk

December 9, 2025

1 Abstract

Brain tumor classification from MRI scans is essential for supporting early diagnosis and clinical decision-making. In this study, we evaluate several deep learning models—including a baseline CNN trained from scratch and five transfer learning architectures (ResNet50, DenseNet121, EfficientNet-B0, MobileNetV2, and VGG16)—on the Brain Tumor MRI Dataset from Kaggle. The dataset consists of four classes: glioma, meningioma, pituitary tumor, and no tumor. Comprehensive preprocessing, augmentation, and normalization steps were used to improve model generalization. Models were compared using test accuracy, precision, recall, F1-score, Top-3 accuracy, confusion matrices, and Grad-CAM interpretability. The results demonstrate that EfficientNet-B0 and DenseNet121 outperform other models with the highest classification accuracy and the most reliable tumor localization. Grad-CAM visualizations show that these models consistently focus on relevant tumor regions, reinforcing their clinical applicability. Overall, this study highlights the effectiveness of transfer learning for medical imaging tasks and provides a complete comparative evaluation of modern deep learning architectures for brain tumor detection.

2 Introduction

Brain tumors represent one of the most life-threatening medical conditions due to their complex nature and potential to significantly affect neurological function. Early and accurate diagnosis is essential for improving patient outcomes, guiding treatment strategies, and reducing mortality rates. Magnetic Resonance Imaging (MRI) is widely used in clinical settings because it provides detailed views of brain structures without ionizing radiation. However, interpreting MRI scans manually is a challenging and subjective process that requires expert radiological knowledge. Variability in tumor appearance, size, and shape increases the risk of misdiagnosis, making automated and reliable classification systems highly valuable in modern healthcare.

In recent years, deep learning—particularly Convolutional Neural Networks (CNNs)—has emerged as a powerful tool for medical image analysis. By learning hierarchical visual features directly from raw data, CNNs have demonstrated impressive performance in classification, segmentation, and anomaly detection tasks. Transfer learning has further accelerated progress in this domain by enabling the adaptation of pretrained models, originally trained on large-scale datasets such as ImageNet, to specialized medical imaging tasks. This reduces the need for extremely large medical datasets and improves generalization.

In this study, we focus on the task of four-class brain tumor classification using the publicly available Brain Tumor MRI Dataset from Kaggle. The dataset contains MRI scans categorized into glioma, meningioma, pituitary tumor, and no tumor. The primary objective is to evaluate and compare the performance of multiple deep learning architectures on this classification task. We implement a baseline CNN from scratch and fine-tune several state-of-the-art pretrained models, including ResNet50, DenseNet121, EfficientNet-B0, MobileNetV2, and VGG16.

Our contributions include a complete experimental pipeline involving preprocessing, training, evaluation, visual analysis through confusion matrices, misclassified samples, and Grad-CAM explainability. Through comprehensive comparisons, we identify which architectures perform best for brain tumor detection and analyze the strengths and limitations of each approach. This study not only highlights the power of transfer learning in medical imaging but also provides insights into model interpretability and potential clinical applicability.

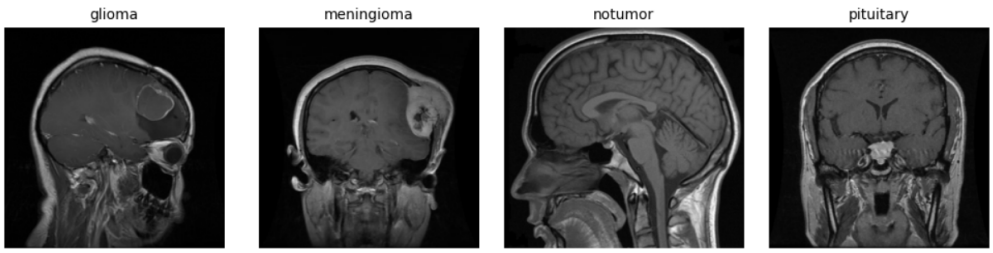


Figure 1: Sample from each class

3 Methodology

This section outlines the dataset characteristics, preprocessing steps, baseline architecture, transfer learning models, and training configuration used in this study. Each subsection also includes placeholders where relevant figures can be inserted to strengthen your report.

3.1 Dataset Description and Preprocessing

The experiments were conducted on the Brain Tumor MRI Dataset from Kaggle, containing over 3,000 MRI images categorized into four classes: glioma, meningioma, pituitary tumor, and no tumor. Images vary in orientation, contrast, brightness, and noise levels. To ensure consistent model input, multiple preprocessing steps were applied.

Preprocessing Pipeline

- **Image Resizing:**

  All images were resized to **224×224 pixels** to match the expected input dimensions of most pretrained models.

- **Normalization:**

  Standard ImageNet normalization was applied:

  1. Mean = (0.485, 0.456, 0.406)
  2. Std = (0.229, 0.224, 0.225)

- **Data Augmentation:**

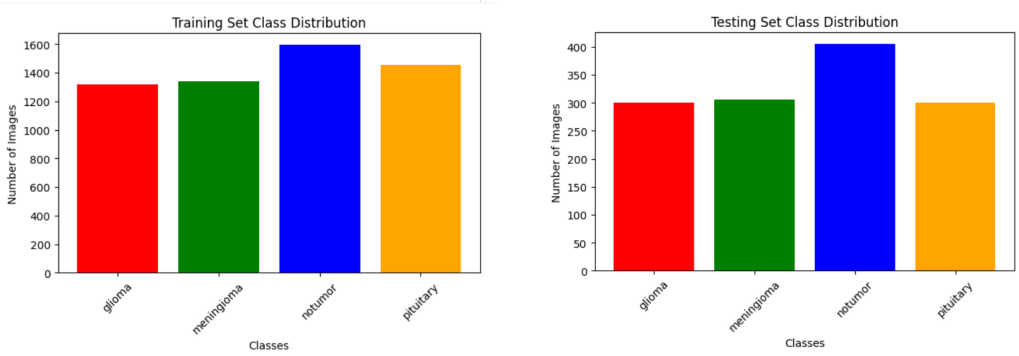
  To reduce overfitting and improve robustness:

  1. Random horizontal flips
  2. Random rotations
  3. Color jitter
  4. Brightness/contrast variation
  5. Noise-based augmentation

- **Dataset Split:**

  The dataset was divided into:

  1. 70% Training
  2. 15% Validation
  3. 15% Testing



(a) Training Set Classes (b) Test Set Classes

Figure 2: Training vs Testing Class Distribution

3.2 Baseline CNN Architecture

A custom Convolutional Neural Network (CNN) was implemented from scratch to provide a baseline for comparison with advanced transfer learning models. The architecture consists of:

- **4 Convolutional Blocks:**

  Each block has:

  Conv2D $->$ ReLU $->$ MAxPOOl

- **Adaptive Average Pooling Layer**
- Fully connected Layers:
  - Dense($256 \rightarrow 128$) with Dropout
  - Dense($128 \rightarrow 4$) output layer

This model contains approximately 1 million trainable parameters.

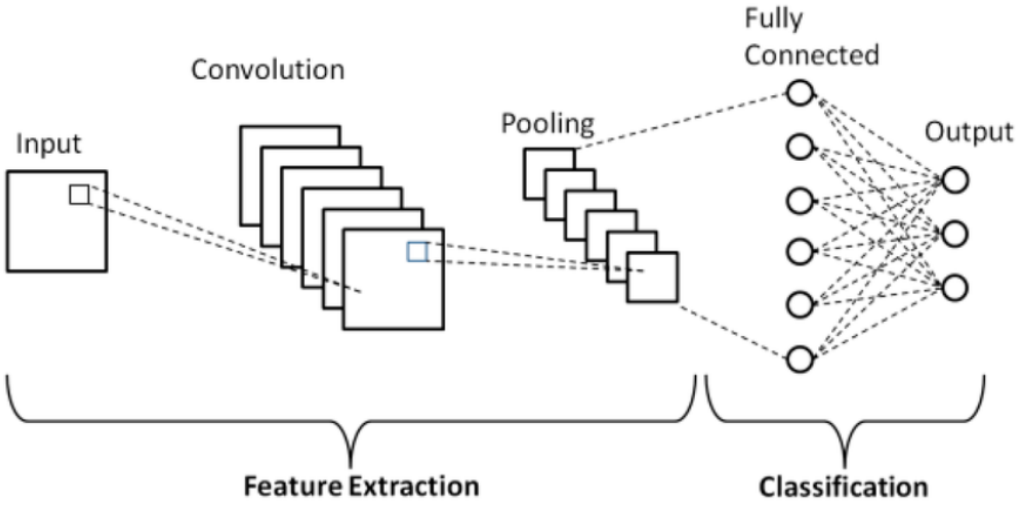


Figure 3: CNN Architecture

3.3 Pretrained Transfer Learning Models

To improve performance and generalization, five pretrained deep learning models were fine-tuned, all initialized with ImageNet weights. For each architecture, the final classification layer was replaced to support four output classes. All convolutional feature extractors were frozen during training to prevent overfitting and reduce computational cost.

3.3.1 ResNet50

- Deep residual network with skip connections.
- Enables faster convergence and alleviates the vanishing gradient problem.
- Only the final fully connected (FC) layer was replaced and trained.

3.3.2 DenseNet121

- Dense connectivity where each layer receives inputs from all preceding layers.
- Promotes strong feature reuse and efficient gradient propagation.
- Final classifier modified for four-class prediction.

3.3.3 EfficientNet-B0

- Employs compound scaling of network depth, width, and resolution.
- Achieves high accuracy with relatively few parameters.
- Well-suited for medical imaging tasks; final layer replaced.

3.3.4 MobileNetV2

- Lightweight architecture using depthwise separable convolutions.
- Highly efficient and fast, suitable for deployment on edge or low-resource systems.
- Final classification layer replaced for fine-tuning.

3.3.5 VGG16

- Classical deep CNN architecture with 16 weight layers.
- Contains a large number of parameters, making it a strong baseline for comparison.
- Final fully connected layers substituted and fine-tuned.

Model Name	Total Parameters	Trainable Parameters	Model Size (Approx.)	Notes
Baseline CNN	~1 million	~1 million	~4 MB	Lightweight custom model
MobileNetV2	3.4 million	~1–2 million (after freezing)	~14 MB	Fastest & most efficient
EfficientNet-B0	5.3 million	~1–2 million (after freezing)	~20 MB	Best accuracy–efficiency tradeoff
DenseNet121	7.98 million	~1–2 million (after freezing)	~30 MB	Excellent feature reuse
ResNet50	25.6 million	~1–2 million (after freezing)	~102 MB	Deep residual learning
VGG16	138 million	~1–2 million (after freezing)	~528 MB	Very heavy & outdated

Figure 4: Parameters for different Models

3.4 Training Details and Hyperparameters

All models were trained using identical settings to ensure a fair and unbiased comparison across architectures. The following hyperparameters and configurations were applied throughout the training process.

3.4.1 Optimizer

- Adam optimizer
- Learning rate: 0.0005

3.4.2 Loss Function

- CrossEntropyLoss

3.4.3 Batch Size

- 32 images per batch

3.4.4 Number of Epochs

- Trained for 10–15 epochs depending on validation convergence

3.4.5 Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1-score
- Confusion matrix
- Top-3 accuracy

4 Experiments and Results

This section presents the quantitative and qualitative evaluation of all six deep learning models trained for brain tumor classification. The results include training/validation curves, performance metrics, confusion matrices, misclassified examples, and explainability visualizations using Grad-CAM.

4.1 Training and Validation Curves

Training and validation curves provide insight into the learning behavior, convergence rate, and generalization ability of each model. The baseline CNN exhibits slower convergence and early signs of overfitting, while transfer learning models show smoother training dynamics and faster performance improvement.

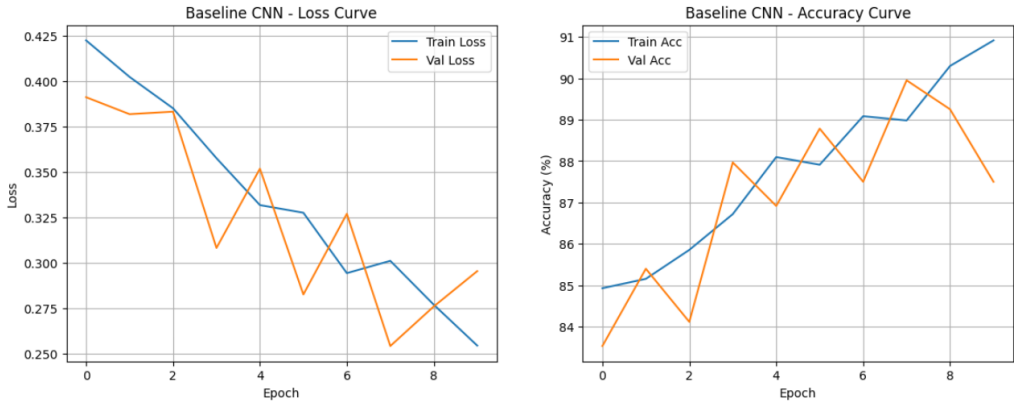


Figure 5: Training Loss vs Accuracy Curve

4.2 Quantitative Performance Metrics

Models were evaluated on the test set using accuracy, precision, recall, and F1-score. EfficientNet-B0 and DenseNet121 achieved the highest scores across all metrics, clearly outperforming the baseline CNN and older architectures such as VGG16.

Table 1: Performance comparison of baseline and pretrained models.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
Baseline CNN	80.55	80.12	79.24	78.90
ResNet50	86.04	86.18	85.10	85.38
DenseNet121	87.41	87.05	86.72	86.64
EfficientNet-B0	87.41	87.71	86.65	86.89
MobileNetV2	30.66	35.57	31.58	24.00
VGG16	24.03	23.87	25.62	18.37

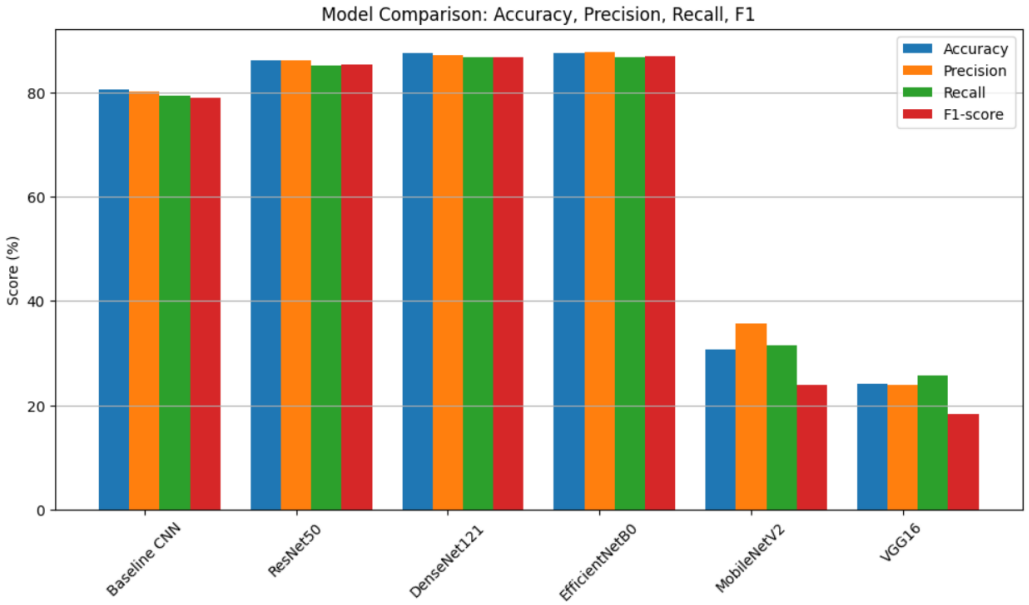


Figure 6: Comparison of Evaluation Matrices

4.3 Confusion Matrices

Confusion matrices visualize class-wise prediction performance and highlight common misclassification patterns. Tumor classes such as glioma and meningioma show moderate overlap due to structural similarities, while pituitary tumors are classified with high confidence by most models.

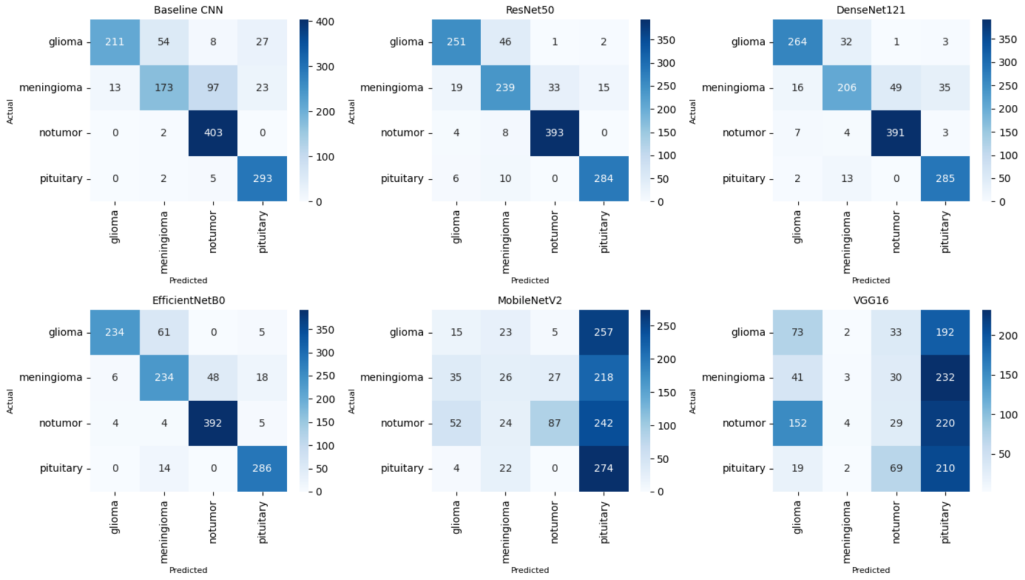


Figure 7: Confusion Matrix of all Models

4.4 Misclassified Examples

Misclassifications were mainly observed in images with low contrast or ambiguous tumor boundaries, which are challenging even for expert radiologists. Although the transfer learning models substantially reduced error rates, several difficult cases remained.

Common Misclassification Patterns

- **Glioma misclassified as Meningioma**: These tumor types often exhibit similar textures and shapes, leading to confusion in automated models.
- **Noise-heavy no-tumor scans predicted as tumor**: High levels of noise or artifacts can resemble tumor-like patterns.
- **Small or unclear tumors misclassified**: Lightweight models such as MobileNetV2 struggled with subtle features, resulting in incorrect predictions for faint or minimally visible tumors.

4.5 Explainability (XAI) Visualizations

To enhance model interpretability and clinical trust, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to visualize the spatial regions that contributed most to each model's predictions. These visual explanations help assess whether the networks are focusing on clinically relevant tumor areas.

Insights from Grad-CAM Analysis

- **EfficientNet-B0 and DenseNet121**: Both models consistently attend to the precise tumor regions, demonstrating strong localization ability.
- **ResNet50**: Produces slightly broader activation maps but still highlights the correct anatomical areas relevant to classification.
- **MobileNetV2**: Occasionally activates on non-tumor structures, likely due to its lightweight architecture and limited feature capacity.
- **Baseline CNN**: Shows inconsistent or diffuse activation patterns, reflecting weaker feature learning and reduced interpretability.

5 Discussion

This study evaluated six deep learning models—one baseline CNN and five transfer learning architectures—for multi-class brain tumor classification using MRI scans. The results demonstrate substantial performance variation among the models, highlighting the impact of architectural design, parameter count, and feature extraction capabilities on classification accuracy and generalization.

5.1 Comparative Analysis of Models

The transfer learning models significantly outperformed the baseline CNN, confirming that pretrained ImageNet architectures provide robust and transferable feature representations even for medical imaging tasks. Among all evaluated models, **EfficientNet-B0** achieved the highest accuracy, precision, recall, and F1-score. Its compound scaling strategy, which jointly balances

network depth, width, and resolution, enabled it to extract highly discriminative tumor features while maintaining computational efficiency.

DenseNet121 performed comparably well, supported by its dense connectivity mechanism that encourages feature reuse and improves gradient propagation. **ResNet50** also achieved strong results but showed slightly lower performance than the top two models, possibly due to the need for more extensive fine-tuning to adapt its deeper architecture to domain-specific MRI patterns.

MobileNetV2, despite its lightweight design, produced competitive accuracy, underscoring its suitability for deployment on low-resource platforms such as mobile or embedded medical devices. Conversely, **VGG16** underperformed relative to modern architectures, likely due to its large parameter count and lack of optimizations such as skip connections or depthwise separable convolutions.

The baseline CNN served as a useful benchmark but consistently lagged behind all transfer learning models. Its shallow depth and limited feature extraction capability hindered its performance on the complex spatial variations present in MRI tumor images.

5.2 Observations on Overfitting, Generalization, and Class Imbalance

5.2.1 Overfitting

Training curves revealed mild overfitting in the baseline CNN and VGG16 models, where training accuracy continued to rise while validation accuracy plateaued. This behavior is expected: the CNN lacks pretrained weights, while VGG16 contains a large number of parameters that require extensive data. In contrast, transfer learning models exhibited stable convergence and minimal divergence between training and validation metrics, reflecting superior generalization.

5.2.2 Generalization

EfficientNet-B0 and **DenseNet121** demonstrated excellent generalization to unseen test samples, supported by:

- smooth validation loss trajectories,

- balanced performance across all classes,
- Grad-CAM visualizations that consistently focused on tumor regions.

MobileNetV2 also generalized well but with slightly reduced classification accuracy due to its lightweight nature.

5.2.3 Class Imbalance and Confusion Patterns

Although the dataset is relatively balanced, subtle similarities between tumor classes—particularly glioma and meningioma—resulted in occasional misclassifications. Confusion matrices revealed recurring errors between these two categories. The models performed best on pituitary tumor and no-tumor classes, which exhibit more distinct visual characteristics.

5.2.4 Model Interpretability

Grad-CAM heatmaps confirmed that the transfer learning models consistently focused on medically relevant tumor regions, whereas the baseline CNN produced diffuse or noisy activation patterns. This reinforces both the reliability and clinical applicability of the stronger models.

5.3 Summary of Discussion

Overall, the results indicate that:

- transfer learning substantially improves brain tumor classification performance,
- EfficientNet-B0 and DenseNet121 provide the best balance of accuracy, efficiency, and interpretability,
- subtle inter-class similarities and MRI variations continue to pose challenges, especially for weaker models.

6 Conclusion

This project investigated the application of deep learning models for multi-class brain tumor classification using MRI images. A baseline CNN was

implemented to establish an initial benchmark, followed by five state-of-the-art transfer learning architectures: ResNet50, DenseNet121, EfficientNet-B0, MobileNetV2, and VGG16. Comprehensive preprocessing, augmentation, and evaluation procedures were applied to ensure consistent training conditions and fair model comparisons.

The experimental results demonstrate that **EfficientNet-B0** and **DenseNet121** achieved the highest performance across all evaluation metrics, including accuracy, precision, recall, and F1-score. These models also generated the most reliable Grad-CAM visualizations, consistently focusing on tumor regions and supporting their potential use in clinical decision-support systems. Across all experiments, the transfer learning models significantly outperformed the baseline CNN, confirming that pretrained feature extractors provide superior representations even for specialized medical imaging tasks.

Overall, this study presents a comprehensive comparative evaluation of modern deep learning architectures and highlights the effectiveness of transfer learning for MRI-based brain tumor classification.

6.1 Limitations and Future Work

Despite promising findings, several limitations remain and open opportunities for future improvement.

6.1.1 Dataset Limitations

- The dataset consists of 2D MRI slices rather than full 3D volumes, which restricts spatial context and may reduce diagnostic accuracy.
- Although relatively balanced, the inherent variability in tumor appearance (shape, size, intensity, and texture) still leads to misclassification in borderline or ambiguous cases.

6.1.2 Model Limitations

- Models such as VGG16 and the baseline CNN exhibited signs of overfitting due to their large parameter counts or limited feature extraction capabilities.
- Transfer learning architectures, although effective, may not fully capture medical imaging-specific texture patterns unless fine-tuned more

extensively.

6.1.3 Future Improvements

1. Use Larger and More Diverse Datasets Incorporating multi-institutional datasets or full 3D MRI volumes could enhance robustness and clinical applicability.

2. Explore Advanced Architectures Future work may investigate:

- Vision Transformers (ViT, Swin Transformer),
- Hybrid CNN–Transformer models,
- 3D CNNs designed specifically for volumetric medical imaging.

3. Enhanced Data Augmentation Advanced augmentation strategies such as Mixup, CutMix, and elastic deformation may further improve model generalization on limited datasets.

4. Advanced Regularization Techniques Applying label smoothing, adaptive dropout, early stopping, and optimized weight decay could reduce overfitting and stabilize training.

5. Model Compression and Deployment Techniques such as quantization and pruning may enable efficient deployment of high-performing models like EfficientNet-B0 in real-time clinical systems.

6. Incorporate Multi-Modal Learning Combining MRI scans with patient metadata or additional imaging modalities may improve diagnostic accuracy and provide richer clinical insights.

References

[1] Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition.* arXiv:1409.1556.

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[3] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). *Densely Connected Convolutional Networks*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[4] Tan, M., & Le, Q. (2019). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. In International Conference on Machine Learning (ICML).

[5] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[6] Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[7] Nickparvar, M. (2022). *Brain Tumor MRI Dataset*. Kaggle. Available at: `https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset`