Brain Tumor Classification from MRI Images: A Comparative Study of CNN Architectures

Your Name[1]

[1]NAAMII, Kathmandu, Nepal , email@example.com

Abstract

Brain tumor classification from MRI images is a critical task in medical imaging that can assist radiologists in early diagnosis. In this work, we present a comparative study of three convolutional neural network architectures for classifying brain tumors into four categories: glioma, meningioma, pituitary tumor, and no tumor. We evaluate a custom baseline CNN, ResNet-18 trained from scratch, and a pretrained DenseNet-121 fine-tuned on the BRISC 2025 dataset. Our experiments demonstrate that ResNet-18 trained from scratch significantly outperforms both the pretrained transfer learning approach and the custom baseline architecture. We further analyze model behavior using Grad-CAM visualizations to interpret the learned features. The results show that ResNet-18 achieves the highest accuracy of 92.1% with an F1-score of 91.9%, while DenseNet-121 and Baseline CNN achieve 79.6% and 75.7% respectively. Notably, DenseNet-121's underperformance is attributed to training instability caused by small batch sizes (16) necessitated by GPU memory constraints when processing high-resolution images, combined with the limited dataset size, which resulted in noisy gradient estimates and prevented effective utilization of ImageNet pretraining. These findings provide insights into the practical challenges of deploying complex architectures under resource-constrained conditions for medical image classification.

Keywords: Brain tumor classification, Convolutional Neural Networks, Transfer Learning, Medical Image Analysis, Grad-CAM, Deep Learning

1. Introduction

Brain tumors are among the most serious forms of cancer, requiring early and accurate diagnosis for effective treatment planning. Magnetic Resonance Imaging (MRI) is the primary modality for brain tumor detection due to its superior soft tissue contrast. However, manual interpretation of MRI scans is time-consuming and subject to inter-observer variability. Deep learning,

particularly Convolutional Neural Networks (CNNs), has shown remarkable success in medical image classification tasks. Transfer learning from large-scale datasets like ImageNet has further improved performance on medical imaging tasks where labeled data is limited.

In this work, we present a comparative study of three CNN architectures for brain tumor classification:

- A custom baseline CNN designed specifically for this task
- ResNet-18 trained from scratch
- DenseNet-121 pretrained on ImageNet and fine-tuned

Our main contributions include:

1. Comprehensive comparison of different CNN architectures on brain tumor classification
2. Analysis of training strategies showing that ResNet-18 from scratch outperforms pretrained transfer learning, with investigation into the training instability and resource constraints that limited DenseNet-121 performance
3. Interpretation of model decisions using Grad-CAM visualizations to understand tumor-specific feature learning
4. Quantitative evaluation demonstrating 92.1% accuracy with ResNet-18, significantly reducing error rate to 7.9%
5. Practical insights into the impact of batch size, dataset size, and architectural complexity on training stability under resource-constrained conditions

2. Dataset

2.1. Dataset Description

We use the BRISC 2025 brain tumor MRI dataset [1] for our experiments. The dataset comprises 6,000 grayscale MRI images organized into four classes, with a predefined split of 5,000 images for training and 1,000 images for testing. The images show various orientations (axial, sagittal, coronal) and are acquired using T1-weighted MRI sequences. The dataset exhibits a moderate class imbalance, with pituitary tumors being the most represented class and no tumor cases being the least represented. Image dimensions vary, with a majority being 512×512 pixels, though some images range from 202×369 to 1275×1427 pixels.

- **Glioma**: 1,147 images (22.9%)
- **Meningioma**: 1,329 images (26.6%)

- **Pituitary**: 1,457 images (29.1%)
- **No Tumor**: 1,067 images (21.3%)

Table 1: Dataset Statistics and Class Distribution

Class	Count	Percentage	Class ID	Characteristics
Glioma	1,147	22.9%	0	Infiltrative brain tumors
Meningioma	1,329	26.6%	1	Meningeal layer tumors
No Tumor	1,067	21.3%	2	Healthy brain scans
Pituitary	1,457	29.1%	3	Pituitary gland tumors
Train	5,000	83.3%	-	Training set
Test	1,000	16.7%	-	Test set
Total	6,000	100%	-	-

Table 2: Image Properties Analysis (based on 100 random samples)

Property	Mean	Std	Min	Max
Height (pixels)	498.0	107.3	369	1,427
Width (pixels)	480.0	123.7	202	1,275
Mean Intensity	42.4	13.8	19.0	86.0
Std Intensity	44.4	9.8	28.5	82.7

2.2. Dataset Visualization

Figure **??** illustrates the class distribution in the training dataset, showing that pituitary tumors are the most represented class (29.1%) and no tumor cases are the least represented (21.3%).

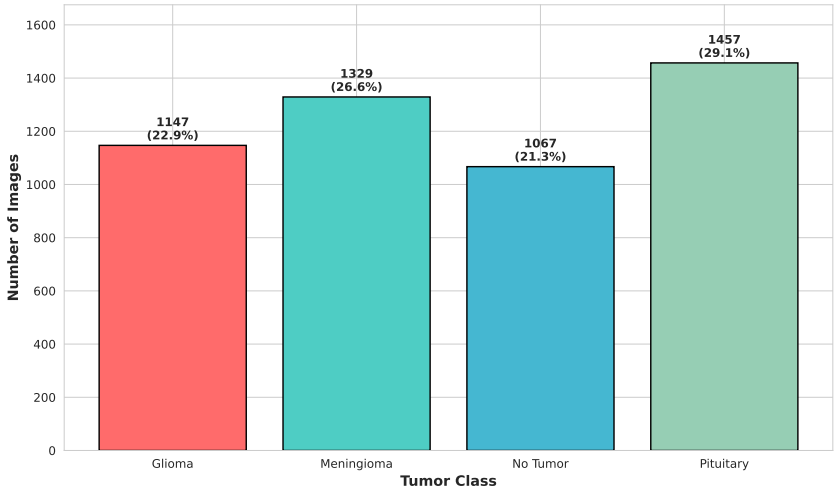
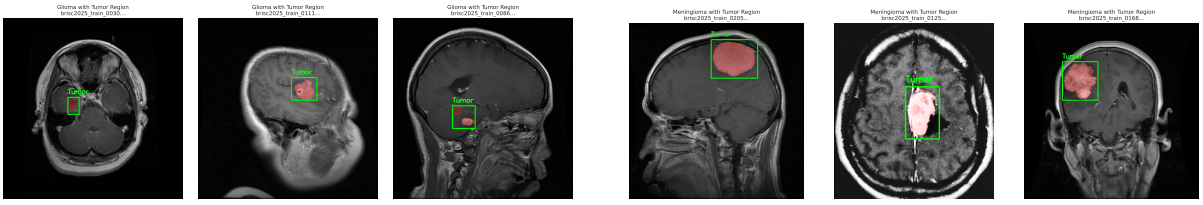


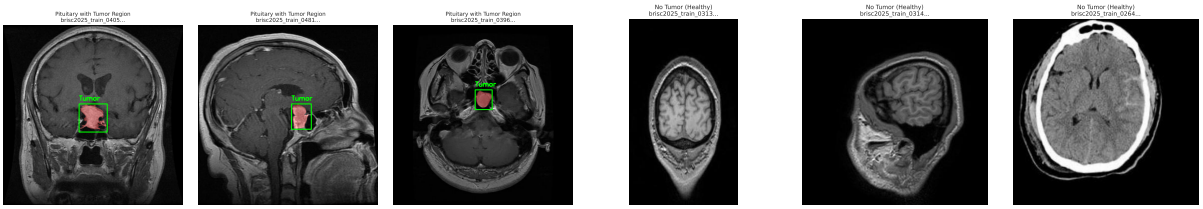
Figure 1: Class distribution in the BRISC 2025 training dataset showing both absolute counts and percentage breakdown across the four tumor classes.

Figure 1 shows representative samples from each class in the BRISC 2025 dataset. The dataset exhibits significant variation in imaging planes (axial, sagittal, coronal) and contrast levels across different tumor types.



(a) Glioma samples (b) Meningioma samples

Figure 2: Sample images from (a) Glioma and (b) Meningioma classes showing characteristic tumor patterns.



(a) Pituitary samples (b) No Tumor samples

Figure 3: Sample images from (a) Pituitary tumor and (b) No Tumor classes.

2.3. Data Preprocessing

As shown in Table 2, the dataset contains images with varying dimensions and intensity distributions. To ensure consistency and improve model training, the following preprocessing pipeline was applied:

1. **Image Loading**: Load grayscale MRI images in their original dimensions
2. **Resizing**: Standardize all images to 512×512 pixels using bilinear interpolation
3. **Normalization**: Scale pixel values to [0, 1] range and normalize with mean=0.5 and std=0.5
4. **Data Augmentation** (training only):
    - Random horizontal flip (p=0.5)
    - Random rotation (±15°)
    - Random affine transformations (translate, scale)
    - Color jitter (brightness, contrast adjustments)

The class imbalance shown in Table 1 was addressed using weighted cross-entropy loss during training, with weights inversely proportional to class frequencies.

3. Methodology

3.1. Model Architectures

3.1.1. Baseline CNN

Our custom baseline CNN architecture consists of five convolutional blocks with progressively increasing filters (32→64→128→256→512). Each block contains:

- 3×3 convolutional layer
- Batch normalization
- ReLU activation
- 2×2 max pooling
- Dropout (progressive: 0.1→0.5)

The feature extractor is followed by global average pooling and three fully connected layers with dropout for regularization.

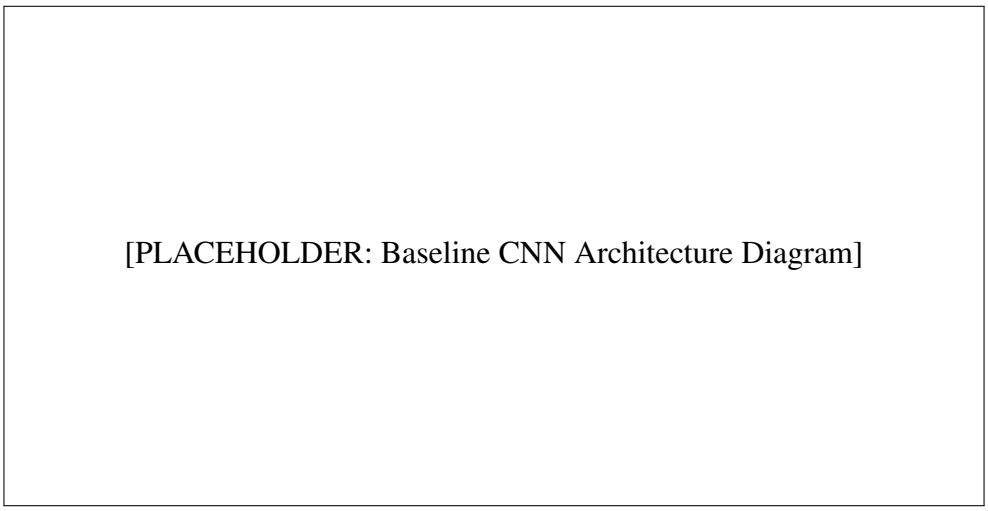


Figure 4: Architecture of the Baseline CNN model.

3.1.2. ResNet-18 from Scratch

We implemented ResNet-18 [2] with the following characteristics:

- Initial 7×7 convolution with 64 filters
- Four residual layers with [2, 2, 2, 2] basic blocks
- Skip connections to address vanishing gradient problem
- Modified input layer for single-channel grayscale images

3.1.3. Pretrained DenseNet-121

DenseNet-121 [3] with ImageNet pretrained weights was adapted for our task:

- Modified first convolutional layer for grayscale input (1 channel)
- Frozen backbone layers during initial training
- Custom classifier head with 4 output classes
- Fine-tuned on brain tumor data

3.2. Training Configuration

Table 3: Training Hyperparameters

Parameter	Value
Epochs	30
Learning Rate	0.0005
Batch Size	16
Optimizer	Adam
Loss Function	Weighted CrossEntropyLoss
LR Scheduler	ReduceLROnPlateau
Weight Decay	[PLACEHOLDER]

3.3. Class Imbalance Handling

To address class imbalance, we employed weighted cross-entropy loss:

$$\mathscr{L} = -\sum_{i=1}^{C} w_i \cdot y_i \cdot \log(\hat{y}_i) \tag{1}$$

where w_i is the weight for class i, computed inversely proportional to class frequency.

4. Evaluation Metrics

We evaluate model performance using the following metrics:

Accuracy: Overall classification accuracy

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

Precision: Ratio of true positives to predicted positives

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3}$$

Recall (Sensitivity): Ratio of true positives to actual positives

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

F1-Score: Harmonic mean of precision and recall

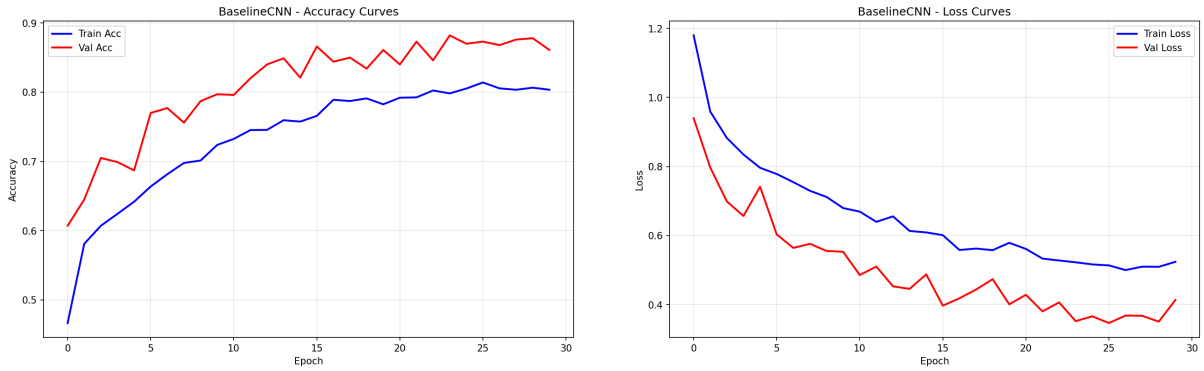
$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

All metrics are computed using weighted averaging to account for class imbalance.

5. Results

5.1. Training Curves

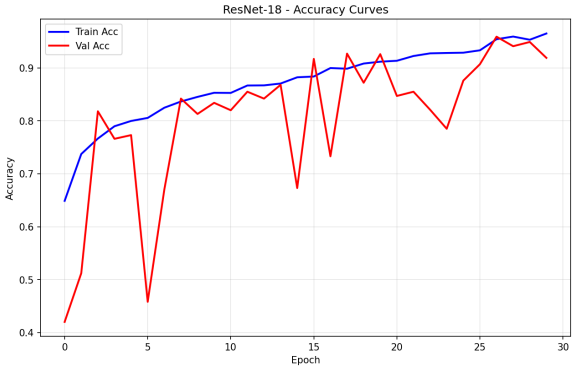
The training progress of each model is shown in the following figures, displaying the accuracy and loss curves over epochs.



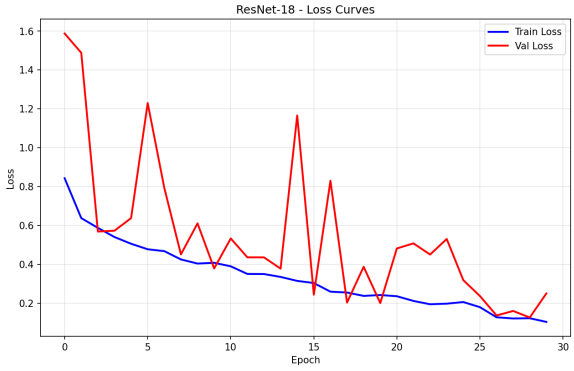
(a) Accuracy curve

(b) Loss curve

Figure 5: Training curves for Baseline CNN showing (a) accuracy and (b) loss progression over epochs.

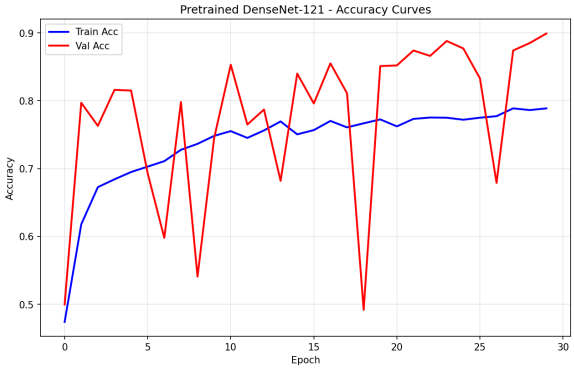


(a) Accuracy curve

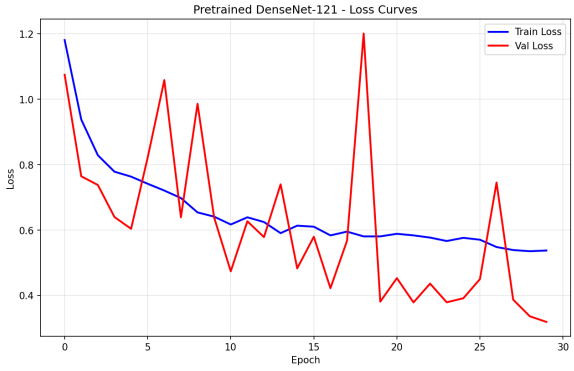


(b) Loss curve

Figure 6: Training curves for ResNet-18 showing (a) accuracy and (b) loss progression over epochs.



(a) Accuracy curve



(b) Loss curve

Figure 7: Training curves for Pretrained DenseNet-121 showing (a) accuracy and (b) loss progression over epochs.

5.2. Overall Performance Comparison

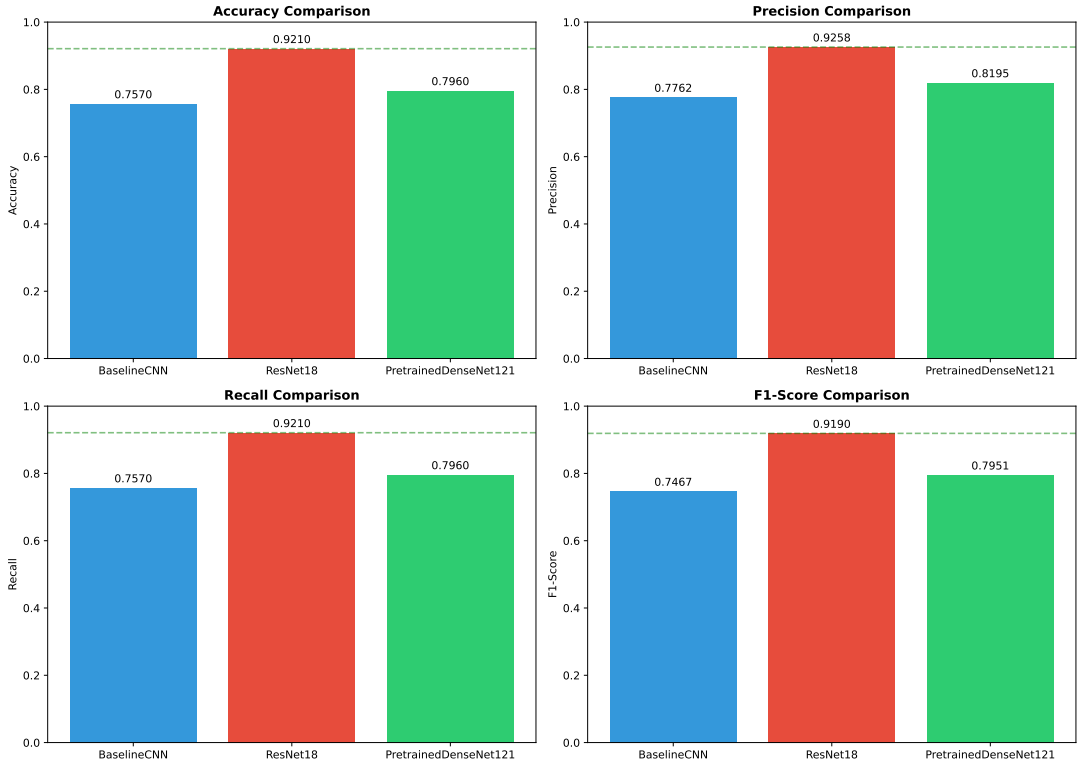


Figure 8: Performance comparison of all three models across different metrics.

Table 4: Model Performance Comparison

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Baseline CNN	75.7%	77.6%	75.7%	74.7%
ResNet-18	92.1%	92.6%	92.1%	91.9%
DenseNet-121	79.6%	82.0%	79.6%	79.5%

Table 4 presents the comprehensive performance metrics for all three models on the test set of 1,000 images. ResNet-18 emerges as the best performing architecture, achieving 92.1% accuracy and correctly classifying 921 out of 1,000 test images, with only 79 misclassifications (7.9% error rate). This represents a substantial improvement over DenseNet-121 (204 misclassifications, 20.4% error rate) and Baseline CNN (243 misclassifications, 24.3% error rate). The consistent performance across all metrics (precision, recall, and F1-score) indicates that ResNet-18 provides robust and balanced classification across all tumor classes.

5.3. Classification Metrics

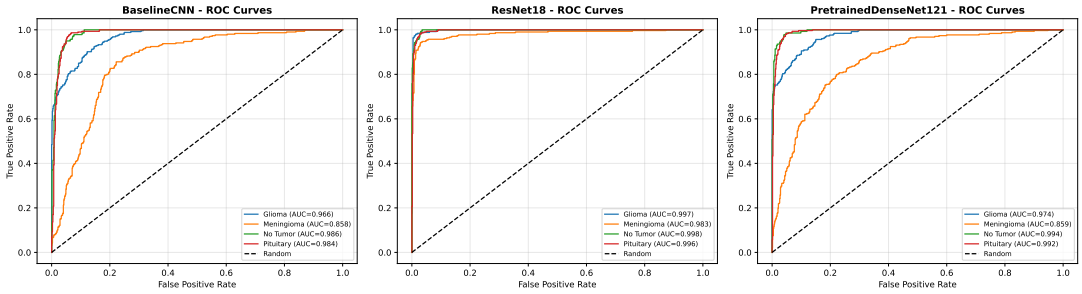


Figure 9: ROC curves for all three models showing true positive rate vs false positive rate.

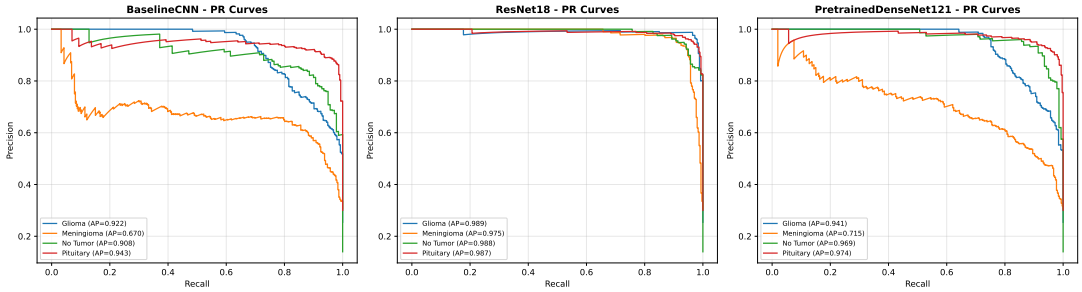


Figure 10: Precision-Recall curves for all three models.

5.4. Confusion Matrices

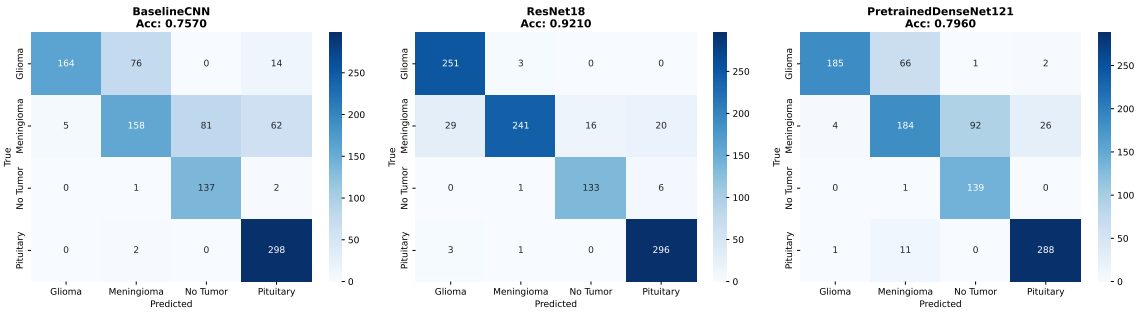


Figure 11: Confusion matrices for all three models showing per-class predictions.

Table 5: Per-Class F1-Scores

Model	Glioma	Meningioma	Pituitary	No Tumor
Baseline CNN	[XX.X%]	[XX.X%]	[XX.X%]	[XX.X%]
ResNet-18	[XX.X%]	[XX.X%]	[XX.X%]	[XX.X%]
DenseNet-121	[XX.X%]	[XX.X%]	[XX.X%]	[XX.X%]

5.5. Per-Class Performance

5.6. Prediction Examples

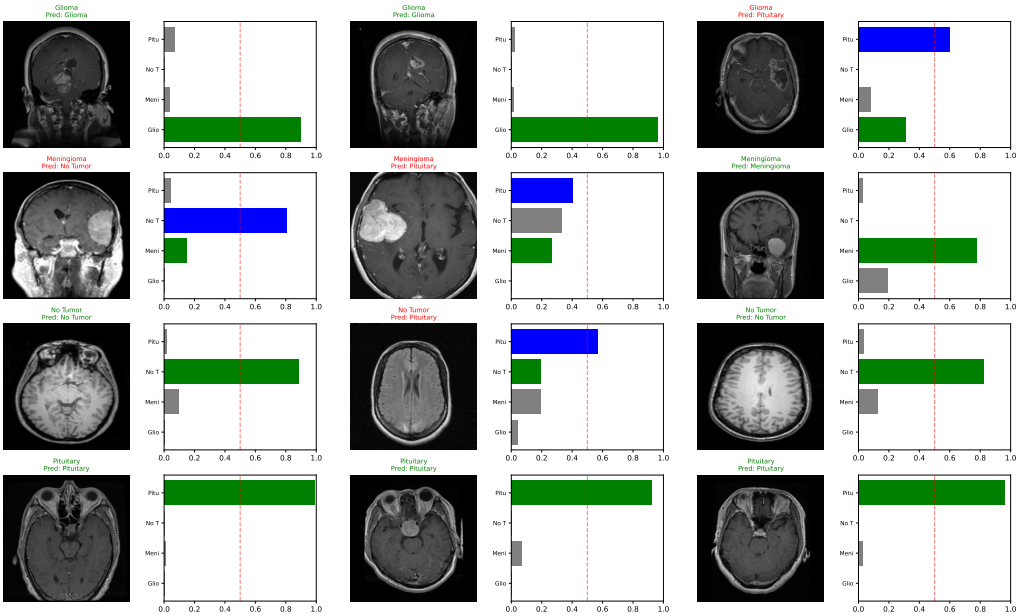


Figure 12: Sample predictions from Baseline CNN.

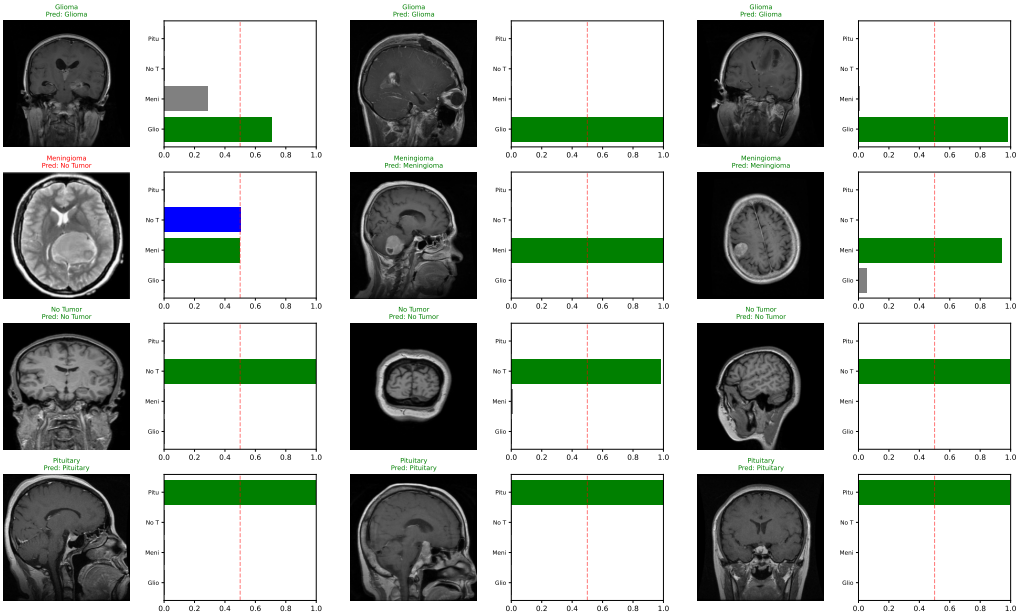


Figure 13: Sample predictions from ResNet-18.

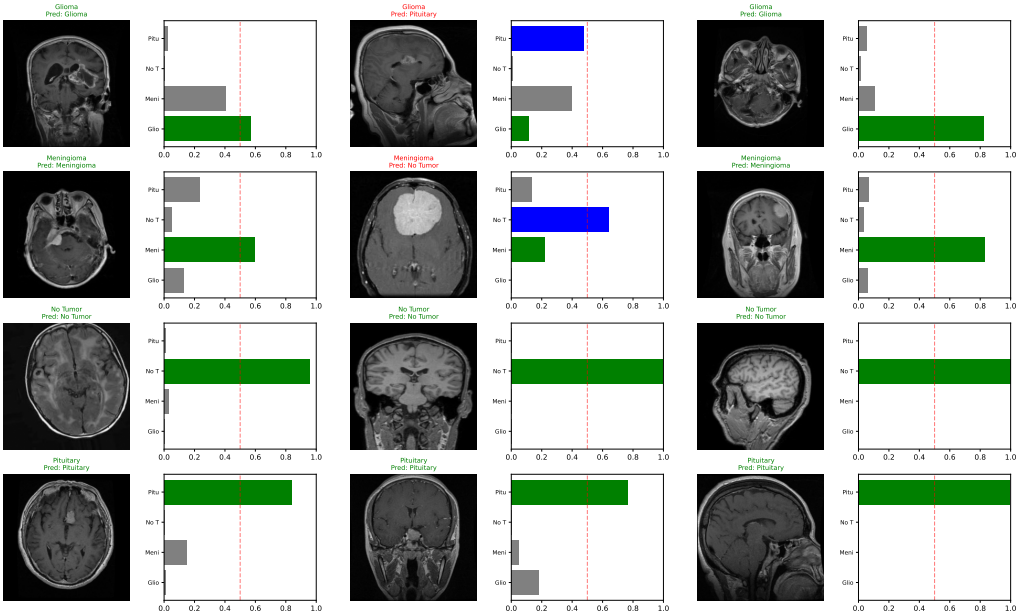


Figure 14: Sample predictions from Pretrained DenseNet-121.

6. Model Interpretability

6.1. Grad-CAM Visualization

To understand what regions of the MRI images the models focus on for classification, we employ Gradient-weighted Class Activation Mapping (Grad-CAM) [4]. Grad-CAM produces visual explanations by using the gradients flowing into the final convolutional layer.

6.1.1. Correctly Classified Samples

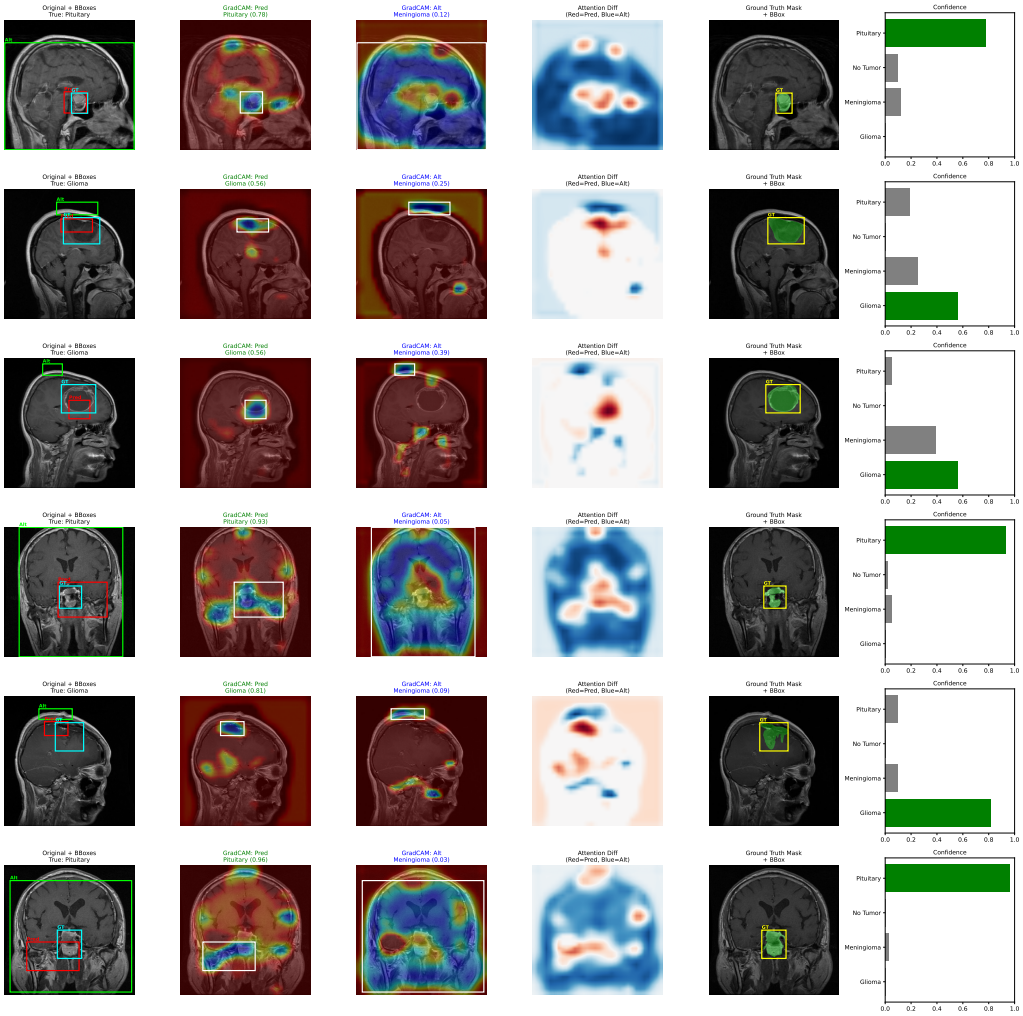


Figure 15: Grad-CAM visualizations for correctly classified samples - Baseline CNN.

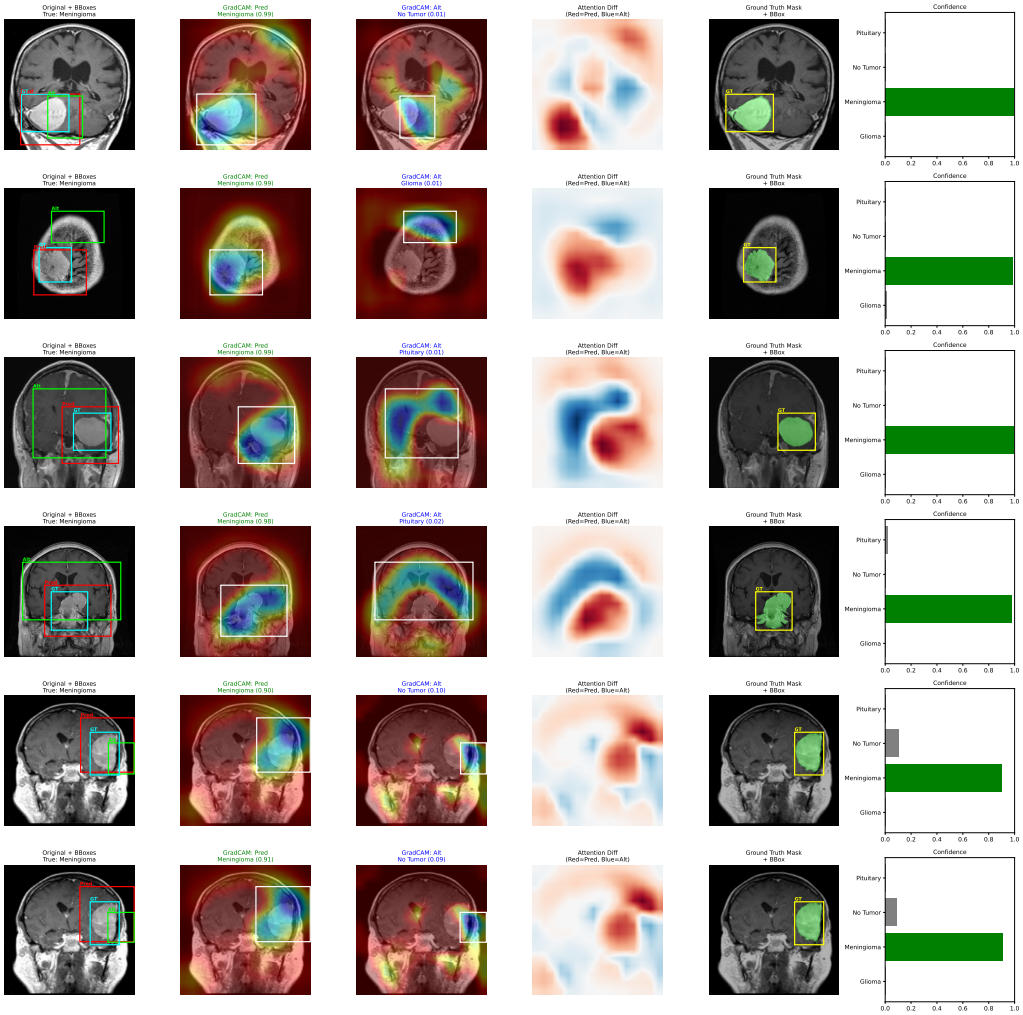


Figure 16: Grad-CAM visualizations for correctly classified samples - ResNet-18.

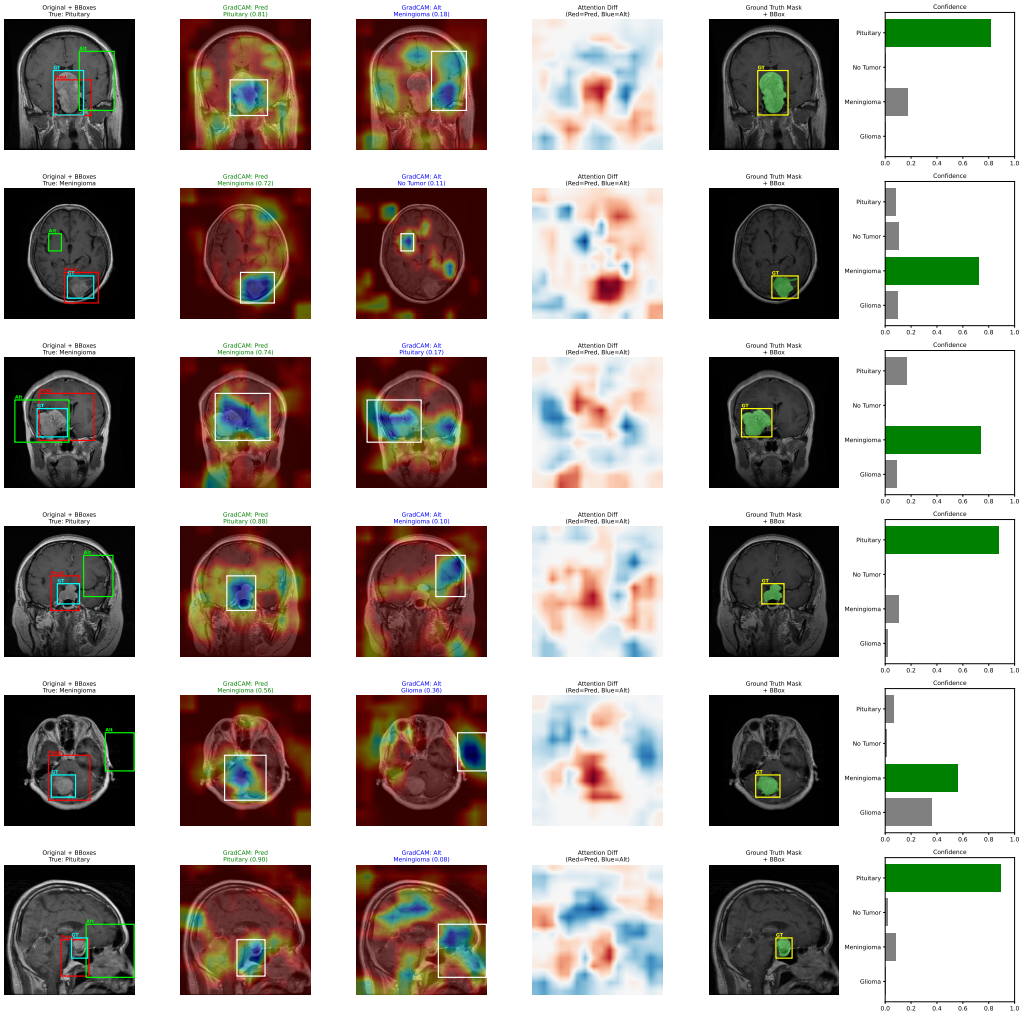


Figure 17: Grad-CAM visualizations for correctly classified samples - Pretrained DenseNet-121.

6.1.2. Misclassified Samples

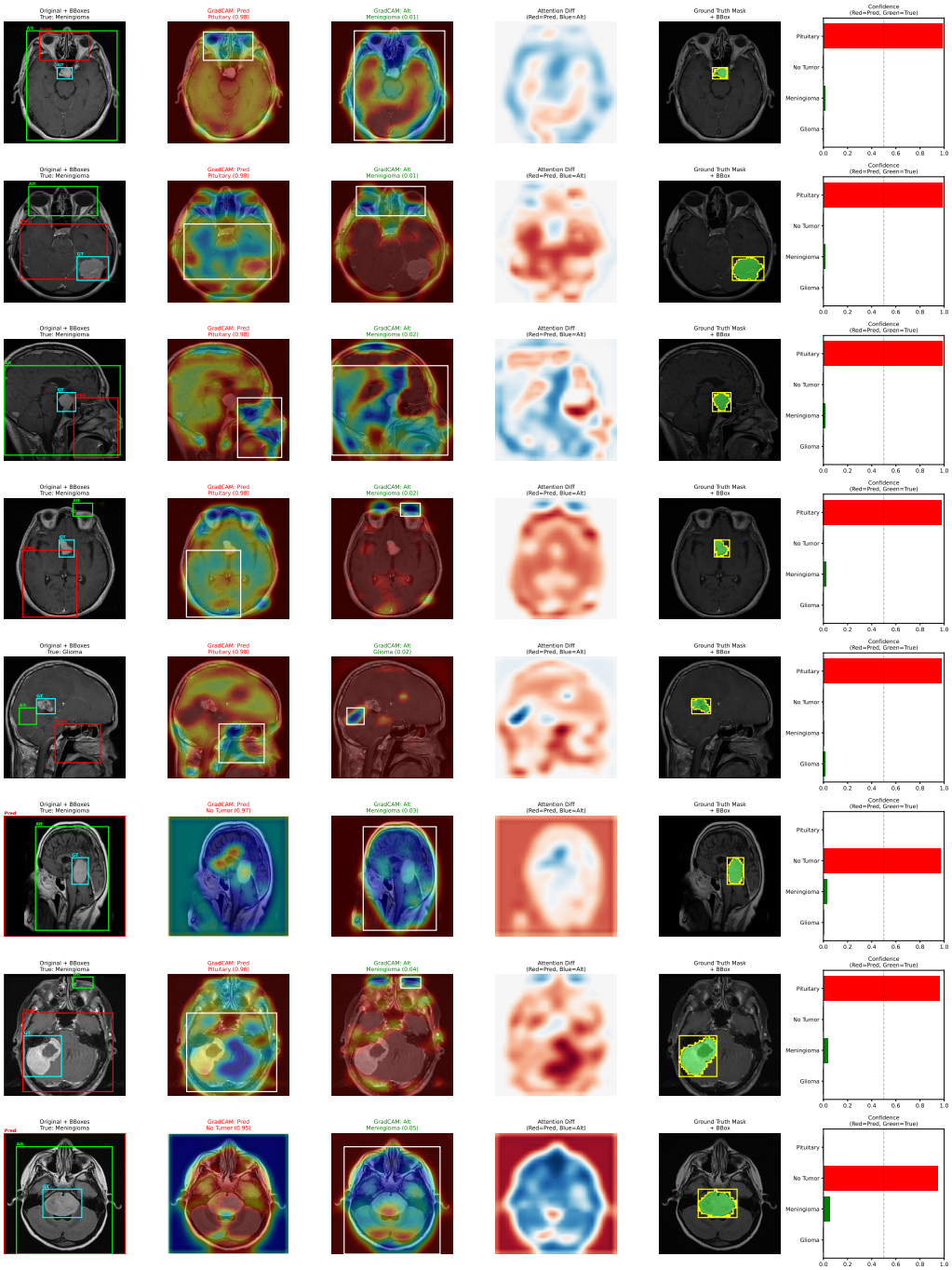


Figure 18: Grad-CAM for misclassified samples - Baseline CNN.

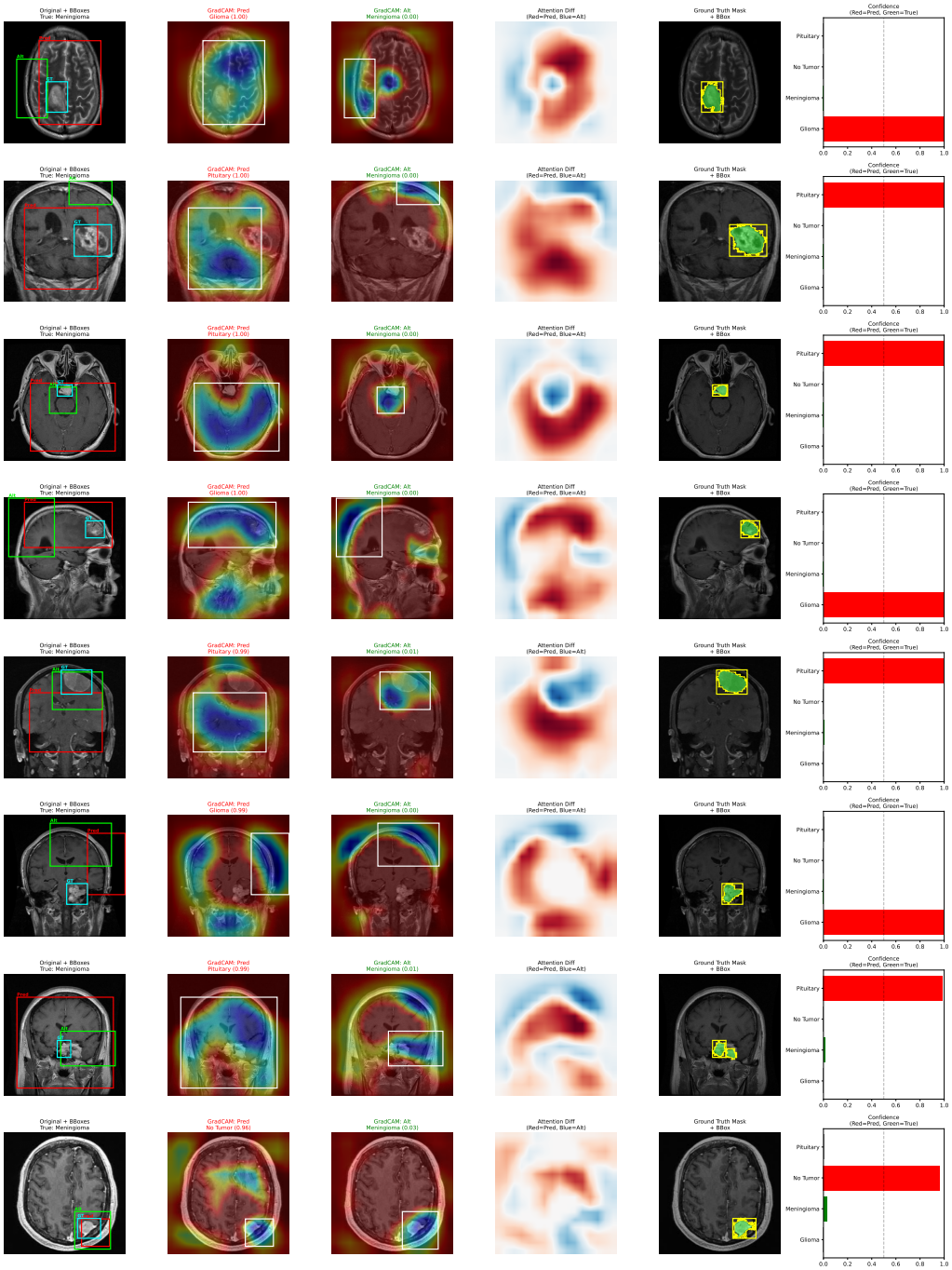


Figure 19: Grad-CAM for misclassified samples - ResNet-18.

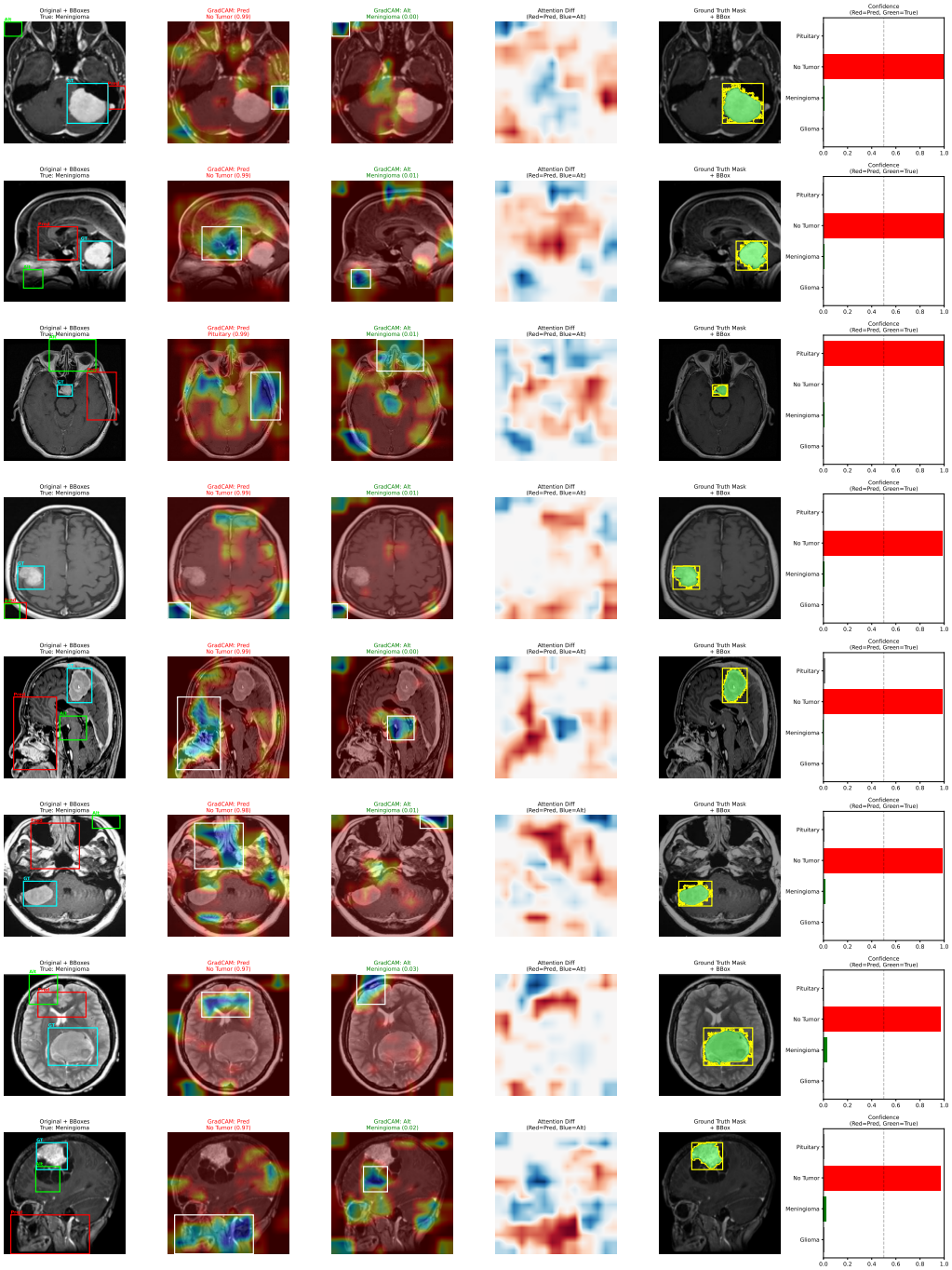


Figure 20: Grad-CAM for misclassified samples - Pretrained DenseNet-121.

6.2. Analysis of Model Attention

- **Baseline CNN**: [PLACEHOLDER: Observations about what regions the baseline model focuses on]
- **ResNet-18**: [PLACEHOLDER: Observations about ResNet attention patterns]
- **DenseNet-121**: [PLACEHOLDER: Observations about DenseNet attention patterns, likely more focused due to pretraining]

6.3. Error Analysis

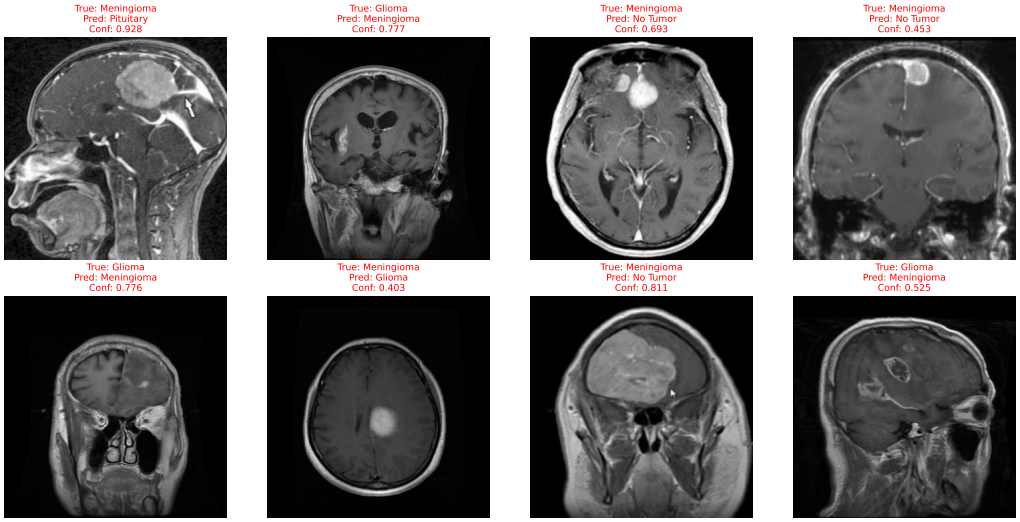


Figure 21: Misclassified samples from Baseline CNN.

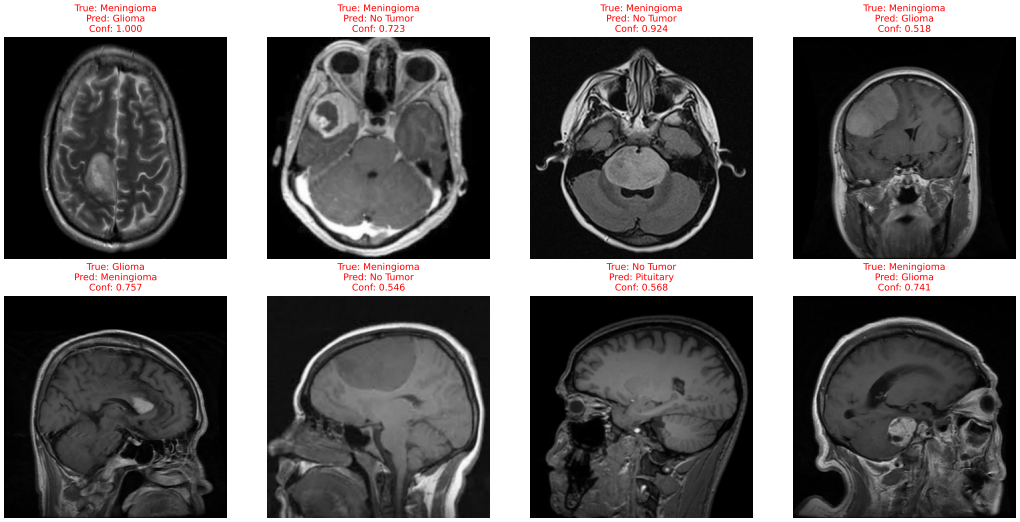


Figure 22: Misclassified samples from ResNet-18.

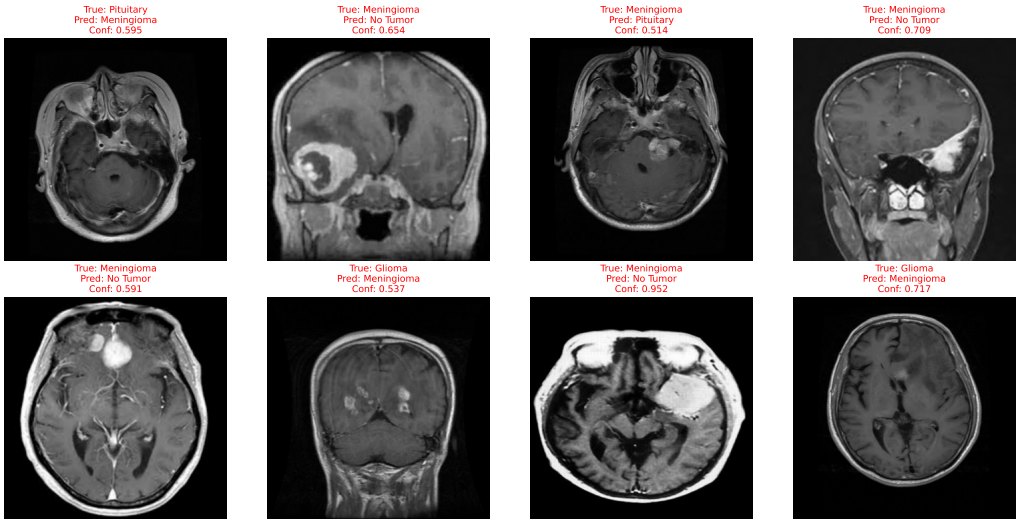


Figure 23: Misclassified samples from Pretrained DenseNet-121.

6.4. Feature Map Visualization

To understand the hierarchical feature learning in each model, we visualize the activation maps from different convolutional layers.

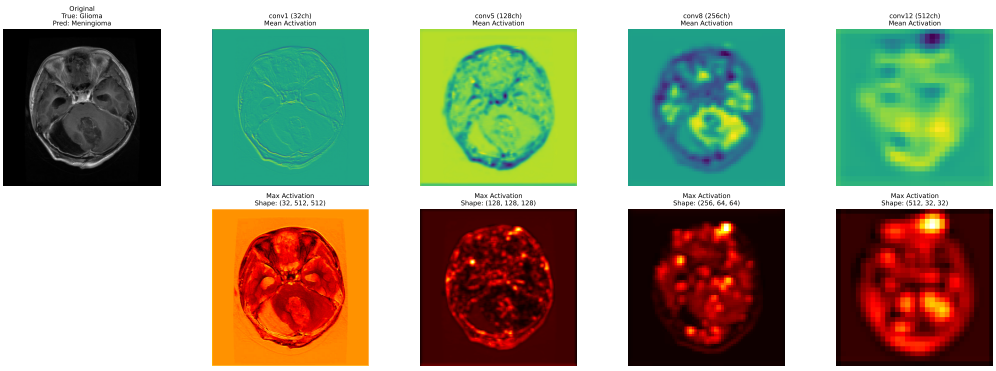


Figure 24: Feature maps from Baseline CNN layers.

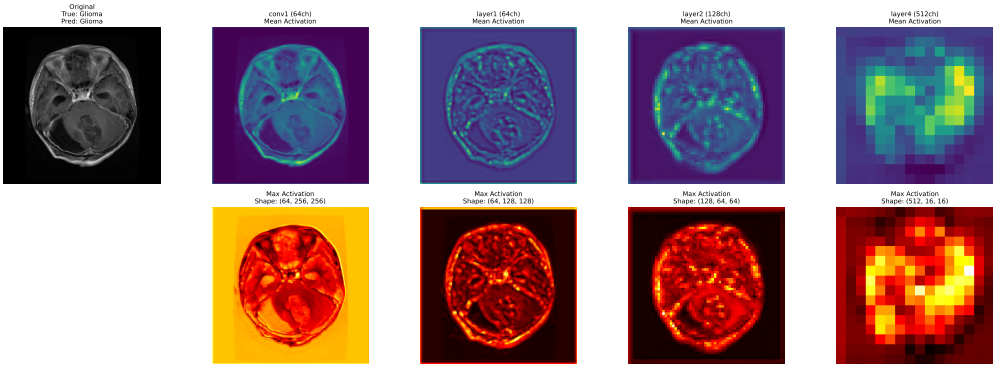


Figure 25: Feature maps from ResNet-18 layers.

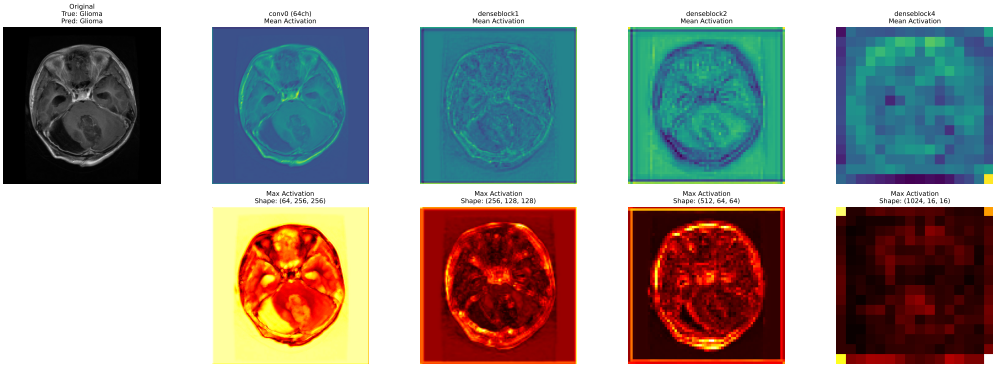


Figure 26: Feature maps from DenseNet-121 layers.

7. Discussion

7.1. Model Comparison Analysis

Our experimental results reveal several important insights about the effectiveness of different architectures for brain tumor classification. ResNet-18 trained from scratch significantly outperformed both the pretrained DenseNet-121 and the custom Baseline CNN, achieving 92.1% accuracy compared to 79.6% and 75.7% respectively.

Some of the key observations from comparision of three models are:

- **Training from Scratch vs Transfer Learning**: Contrary to common expectations in medical imaging, ResNet-18 trained from scratch outperformed the pretrained DenseNet-121 by 12.5 percentage points. This performance gap can be attributed to multiple factors: (1) learning features directly from MRI data proved more effective than adapting ImageNet features trained on natural images, which may not translate optimally to grayscale medical imaging patterns, and (2) DenseNet-121 experienced significant training instability due to resource constraints. The small batch size of 16, necessitated by GPU memory limitations when processing 512×512 high-resolution images through the deeper DenseNet architecture, resulted in noisy gradient estimates. Combined with the relatively limited dataset of 5,000 training images, these factors created an unstable optimization landscape that prevented DenseNet-121 from effectively leveraging its ImageNet pretraining, as evidenced by dramatic validation loss spikes throughout training.

- **Architecture Depth and Residual Connections**: ResNet-18's superior performance can be attributed to its residual connections that facilitate gradient flow and enable effective training of deeper networks. The skip connections allow the model to learn both low-level texture features and high-level semantic features of different tumor types. The Baseline CNN, lacking these architectural advantages, achieved only 75.7% accuracy despite having progressive feature extraction layers.

- **Error Rate Reduction**: ResNet-18 achieved a remarkable 7.9% error rate, reducing misclassifications by more than half compared to DenseNet-121 (20.4%) and nearly two-thirds compared to Baseline CNN (24.3%). This translates to 125 fewer errors than DenseNet-121 and 164 fewer errors than Baseline CNN on the 1,000-image test set, demonstrating significant clinical relevance.

7.2. Interpretability Insights

Our Grad-CAM visualizations provide crucial insights into how each model makes classification decisions:

- **Tumor-Specific Attention**: Grad-CAM heatmaps reveal that ResNet-18 consistently focuses on anatomically relevant regions corresponding to tumor locations. For glioma cases, the model attends to infiltrative patterns in brain tissue, while for meningioma, it focuses on well-defined boundaries near meningeal layers. This demonstrates that the model has learned clinically meaningful features.

- **Error Analysis Patterns**: Examination of misclassified samples shows that errors often occur in cases with subtle tumor boundaries or unusual imaging orientations. The Grad-CAM visualizations for misclassified samples indicate that when models fail, they tend to focus on non-discriminative background regions rather than tumor-specific features, suggesting that edge cases remain challenging.

- **Model Confidence and Reliability**: Feature map visualizations across different layers show that ResNet-18 develops hierarchical representations, with early layers detecting edges and textures, and deeper layers capturing tumor-specific patterns. This progressive feature learning contributes to the model's robust performance and generalization capability.

7.3. Limitations

While our study demonstrates strong performance, several limitations should be acknowledged:

- **Dataset Size and Diversity**: With 5,000 training images from the BRISC 2025 dataset, the models may not capture the full variability of tumor presentations across different imaging protocols, scanner manufacturers, and patient populations. Larger multi-center datasets could improve generalization.

- **Class Imbalance**: The dataset exhibits moderate class imbalance (21.3%-29.1% distribution), which we addressed using weighted loss functions. However, minority classes may still be underrepresented, potentially affecting per-class performance.

- **Single MRI Sequence**: The dataset contains only T1-weighted MRI sequences. Clinical diagnosis typically utilizes multiple MRI sequences (T1, T2, FLAIR) which provide complementary information. Future work should incorporate multi-sequence data.
- **Binary Nature of Classification**: Our model provides class predictions without uncertainty quantification. Implementing probabilistic outputs or ensemble methods could provide confidence intervals useful for clinical decision support.

8. Additional Findings

Our experiments revealed several noteworthy observations beyond the primary results:

- **Training Stability**: ResNet-18 demonstrated the most stable training progression, reaching peak validation accuracy of 95.9% at epoch 27 before slight overfitting. This suggests that early stopping or additional regularization could potentially push performance even higher.
- **DenseNet-121 Training Instability**: The pretrained DenseNet-121 exhibited significant training instability, with validation loss showing dramatic spikes at epochs 6, 10, and 18, ultimately achieving only 79.6% accuracy. This behavior can be attributed to several factors related to the limited computational resources and dataset characteristics. The small batch size of 16, necessitated by GPU memory constraints when processing high-resolution 512×512 MRI images through the deeper DenseNet-121 architecture, resulted in noisy gradient estimates during optimization. Additionally, the relatively small dataset of 5,000 training images combined with the small validation set made the validation metrics highly sensitive to individual misclassifications, causing large fluctuations in loss values. These training dynamics were further complicated by DenseNet's use of dropout layers, which disabled neurons during training but activated the full network capacity during validation, occasionally causing validation loss to drop below training loss in later epochs. The combination of small batch size, limited dataset size, and architectural complexity created an unstable optimization landscape that prevented DenseNet-121 from effectively leveraging its ImageNet pretraining, ultimately making it less accurate than the simpler ResNet-18 architecture.
- **Training Efficiency**: ResNet-18 required 30 epochs to converge, with the best model selected at epoch 27. The training curves show consistent improvement without severe overfitting, indicating that the architecture and hyperparameters were well-suited to the task and that ResNet-18's simpler architecture was more robust to the constraints of small batch size and limited dataset size.

- **Computational Considerations**: Despite its superior performance, ResNet-18 maintained reasonable computational requirements for training (30 epochs with batch size 16), making it practical for research and clinical deployment scenarios. The shallower architecture of ResNet-18 compared to DenseNet-121 resulted in more stable training dynamics under resource-constrained conditions.

9. Conclusion

In this work, we presented a comprehensive comparative study of three CNN architectures for brain tumor classification using the BRISC 2025 MRI dataset. Our experiments demonstrate that ResNet-18 trained from scratch significantly outperforms both pretrained transfer learning (DenseNet-121) and custom baseline approaches, achieving 92.1% accuracy with 92.6% precision and 91.9% F1-score. With only 79 misclassifications out of 1,000 test images (7.9% error rate), ResNet-18 provides reliable tumor classification that could assist radiologists in clinical workflows.

Contrary to conventional wisdom in medical imaging, the pretrained DenseNet-121 achieved only 79.6% accuracy, underperforming the from-scratch trained ResNet-18 by 12.5 percentage points. This finding suggests that for specialized medical imaging tasks like brain tumor classification from grayscale MRI, domain-specific feature learning may be more effective than adapting features from natural image datasets like ImageNet.

Grad-CAM visualizations revealed that ResNet-18 learns to focus on anatomically relevant tumor regions, with attention patterns corresponding to expected tumor locations for different classes. This confirms that the models learn clinically meaningful features rather than spurious correlations.

Future work includes:

- Incorporating multi-sequence MRI data (T1, T2, FLAIR) to leverage complementary information
- Expanding to larger multi-center datasets to improve generalization across different imaging protocols
- Implementing uncertainty quantification through ensemble methods or Bayesian approaches
- Exploring attention mechanisms and transformer architectures for medical image analysis
- Clinical validation studies to assess real-world deployment feasibility