International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Is the Depth Worth It? A Cost-Aware Comparative Study of ResNet18 and ResNet50 for Brain Tumor MRI Classification and Clinical Deployment

Asutosh Dash, Varsha Rajendran, Ritesh Ranjan Panda
SCOPE
VIT-AP University
Vijayawada, Andhra Pradesh

Abstract - Brain tumors are a health concern worldwide, and timely diagnosis increases the probability of survival for patients. In the present research, two deep learning architectures, ResNet18 and ResNet50, are compared and evaluated in terms of brain tumor recognition, and the question of whether the higher training and inference cost of the deeper model is justified is studied. Preprocessing methods are applied to clarify images and improve model performance before training. Accuracy, precision, recall, confusion matrix, Grad-CAM interpretations, and com-putational cost metrics—including parameter count, model size, FLOPs, training time, and inference time—are considered in the evaluation. The results show a modest, statistically ambiguous performance gain for ResNet50 that is concentrated in a single class (meningioma), while ResNet18 matches or exceeds ResNet50 on the remaining classes at substantially lower computational cost. Trade-offs are discussed along with class- and deployment-specific recommendations.

Index Terms - Brain MRI Classification, Deep Learning, ResNet18, ResNet50, Transfer Learning, Grad-CAM, Brain Tumor Detection, Computational Cost, Clinical Deployment, Comparative Study

I. INTRODUCTION

imaging [1]. There are two commonly used variants, ResNet18 and ResNet50; ResNet18 is characterized by faster and less resource-intensive computation, while ResNet50 has greater capacity for feature extraction but requires more computational resources [3]. The choice of the deeper architecture for deploy-ment in a clinical setting is not determined by accuracy alone. Hospitals and diagnostic centers, especially those operating in resource-limited environments, must weigh prediction perfor-mance against inference time, memory consumption, hardware cost, and energy consumption. A model that provides only a marginal improvement in accuracy but requires substantially more resources for inference is not always the best choice for a clinical workflow. Although both models are widely used across numerous studies, a direct, cost-aware comparison under identical conditions for brain tumor classification is rather rare [2], [4]. This lack of cost-aware comparison makes it necessary not only to determine which model performs better, but to determine whether the performance gain of the deeper architecture is worth its additional cost.

A. Background

Brain tumors are a critical issue for medicine, and timely MRI scan-based detection is significant for improving patient outcomes. Manual analysis of MRI scans may be erroneous, and therefore the use of deep learning methods for automatic classification is widely spread. Residual Neural Networks (ResNets) are preferred due to their ability to resolve training problems and extract deeper features [1]. Recent studies have shown that ResNet50 combined with an explainable approach like Grad-CAM demonstrates superior performance and inter-pretability for brain tumor detection [2], [3].

B. Motivation

Brain tumor detection is a challenging task because review-ing MRI scans can be slow and error-prone. Deep learning models such as Residual Networks have shown excellent results in computer vision tasks and are widely used in medical

C. Problem Statement

Brain tumors are a serious health threat, and timely detec-tion is critical for improving patients' probability of survival. Even though MRI scans provide detailed images, manual analysis can be slow and error-prone. Deep learning models like ResNet18 and ResNet50 have proven their applicability for automatic brain tumor detection; however, comparative studies under identical experimental conditions are scarce, and only a few studies compare the models while also quan-titatively assessing their computational cost. Evaluating the balance between accuracy, computational performance, and interpretability—and, critically, whether that balance favors a lightweight or a deeper model in practice—is necessary to identify which architecture is more suitable for real-world clinical applications [1]–[3].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

A. Objectives

In the present research, ResNet18 and ResNet50 are developed and evaluated using a publicly available brain MRI dataset that includes images of glioma, meningioma, pituitary tumor, and normal brain. The performance of the models is compared using metrics such as accuracy, precision, recall, and F1-score. In addition, learning curves are examined to understand the dynamics of model improvement during training. To make the models more interpretable, the attention behavior of the networks on MRI scans is demonstrated using Grad-CAM. Beyond standard metrics, the study quantitatively evaluates the computational cost of each architecture: parameter count, model size, FLOPs, training time, and inference time.

B. Contributions

This paper provides a structured comparison of ResNet18 and ResNet50 for brain MRI classification that explicitly accounts for computational cost. The study includes an extensive experimental evaluation using accuracy, confusion matrices, and other classification measures for each tumor type. In addition, the value of an interpretable model is highlighted through Grad-CAM visualizations, revealing the most important MRI regions considered by each model. The paper also offers insight into the balance between cost and prediction accuracy, and questions the significance of the accuracy improvement of ResNet50 over ResNet18.

II. RELATED WORK

A. Traditional Machine Learning in Brain MRI Classification

Early attempts to classify brain MRI relied mainly on handcrafted features combined with classical machine learning algorithms. SVM, RF, k-NN, and other classifiers were used to classify brain tumors based on MRI texture, shape, and intensity. Although these methods showed satisfactory results, their major disadvantage was an inability to capture the multilayered characteristics of medical images, leading to inferior performance compared to modern deep learning approaches [5].

B. Deep Learning in Brain MRI Classification

CNNs marked a revolution in medical image analysis since these models can learn important features automatically, without manual feature engineering [6], [7]. A number of studies have demonstrated the superiority of CNNs over traditional machine learning approaches for the detection and classification of brain tumors on MRI [4], [8]. Nonetheless, the performance of early CNNs was restricted by their simple architecture and limited ability to classify diverse tumor types and adapt to varying MRI quality [5].

C. Residual Networks (ResNets) in Medical Imaging

ResNet, or Residual Network, was invented to solve the vanishing-gradient problem and enable the training of much deeper models [1]. ResNet18 and ResNet50 are both commonly used for brain MRI classification. ResNet18 is highly efficient and has a reduced computational cost, making it appropriate for live applications and resource-constrained settings [5]. ResNet50, owing to its deeper architecture, usually achieves higher accuracy due to improved feature extraction, though at the cost of higher memory usage and longer training time [2]. Most research concentrates on either ResNet18 or ResNet50 in isolation, providing little direct comparison of the two under identical conditions—leaving a gap in understanding the trade-offs among network depth, prediction accuracy, and performance speed [4], [8].

D. Explainable AI in Medical Imaging

Explainable AI has gained increasing importance in medical imaging, since high accuracy alone is not sufficient for clinical applications [9]. The Grad-CAM (Gradient-weighted Class Activation Mapping) technique allows visualization of the MRI regions that contribute most to a model's prediction [10]. Research has shown that Grad-CAM can help verify whether a model is focusing on tumor-related regions [2].

E. Research Gap

Despite previous attempts to use ResNets for brain tumor classification, most existing papers address the problem by simply comparing the accuracy, precision, recall, or F1-score of two network architectures [2], [3]. If one network performs marginally better on any of these metrics, the deeper architecture is typically considered superior, without asking whether the difference is truly significant and reliable, and without weighing that improvement against the additional computational complexity and training/inference cost it incurs. In practice, however, these networks may be deployed in resource-constrained hospitals, point-of-care devices, or high-throughput screening pipelines, where computational cost is a crucial deployment criterion. Comprehensive comparative analyses that jointly report predictive performance and computational cost under uniform experimental conditions remain limited [4], [8], and the statistical significance of small accuracy differences between architectures is rarely examined. This study takes a step forward by performing a cost-sensitive comparative analysis of ResNet18 and ResNet50: alongside reporting the performance metrics of the two architectures, this paper investigates whether the additional computation required by ResNet50 is justified by its additional performance, using quantitative data on performance metrics, computational cost, and Grad-CAM visualizations.

III. METHODOLOGY

A. Proposed Workflow

The overall workflow followed in this study for comparing ResNet18 and ResNet50 on brain MRI classification can be divided into six main stages, as illustrated in Fig. 1.

1) Data Preprocessing: The collection of brain MRI images comprised four categories: glioma, meningioma, pituitary tumor, and non-tumor images. The collected dataset was split into portions used to train the model and to evaluate it. To conform to the input size expected by the ResNet architectures,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

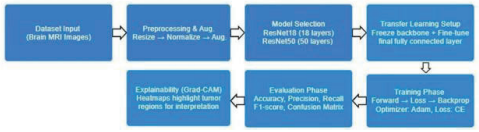

Fig. 1. Schematic representation of the end-to-end experimental pipeline consisting of six stages: (1) Dataset preparation, (2) Model selection, (3) Transfer learning setup, (4) Training phase, (5) Evaluation phase, and (6) Model explainability using Grad-CAM. Both architectures are passed through exactly the same pipeline.

all images were resized to 224×224 pixels. Several data augmentation techniques were used to improve the generalization power of the training set: (1) random horizontal flipping, (2) random rotations, and (3) normalization using the mean and standard deviation values of the ImageNet dataset.

2) Model Selection: Two pre-trained CNN architectures were employed:

- **ResNet18:** A shallower network containing 18 layers.
- **ResNet50:** A deeper network containing 50 layers.

Both models were initialized with ImageNet pre-trained weights. Modifications were introduced to the last FC layer so that the output size matched the number of target classes (4).

3) Transfer Learning Setup: The primary convolutional structure in both models was held constant to preserve the fundamental feature representations already acquired. Adjustments were restricted to the ultimate classification component, using the set of brain MRI images for this purpose. The Adam optimization algorithm was chosen, configured with a learning rate of 0.001. The objective function used for minimization was the Cross-Entropy measure of error.

4) Training Phase: Training was carried out over 5 epochs. In each epoch, the forward pass was performed for training batches, the loss was computed and backpropagated, and model parameters were updated. The performance of the models was tracked after each epoch in terms of accuracy and loss values.

5) Evaluation Phase: After training, both models were evaluated on the validation dataset. The following metrics were computed: accuracy (fraction of correct predictions), precision of positive predictions, recall of positive predictions, and the harmonic mean of precision and recall. Confusion matrices were plotted to illustrate the models' mistakes, and graphs illustrating the dynamics of model performance were produced.

6) Explainability via Grad-CAM: To make the models more understandable, the Grad-CAM methodology (Class Activation Mapping with Gradient weighting) was used. Grad-CAM visualizations showed which areas of the MRI images the networks attended to when performing classification, thus verifying whether the networks were genuinely focusing on tumor regions.

B. Preprocessing

Preprocessing of MRI scans is crucial for classification with deep learning. To ensure consistency and enhance the generalizability of the models, the steps shown in Fig. 2 were carried out.

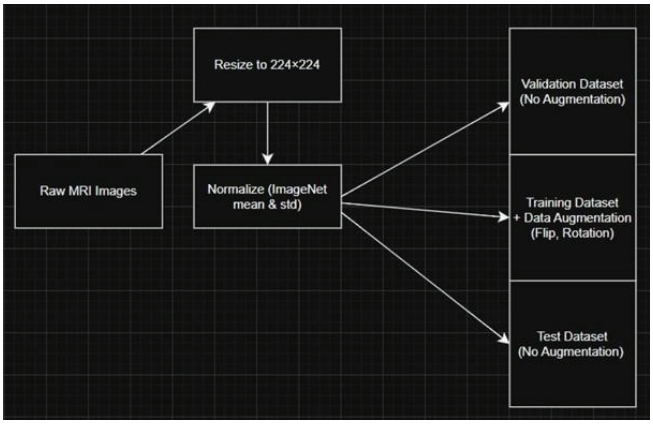

Fig. 2. Preprocessing pipeline applied equally to both models: (a) scaling to 224×224 pixels, (b) normalization based on ImageNet channel means and standard deviations, and (c) data augmentation (horizontal flips, rotations, and color normalization) applied only to the training subset.

1) Resizing Images: The size of each MRI image was altered to ensure a resolution of 224×224 pixels. This resizing ensures that the images meet the size expected by the ResNet18 and ResNet50 networks, since these architectures are built around the ImageNet standard size.

2) Normalization: Pixel intensity values in the MRI scans were standardized using the per-channel mean and standard deviation derived from the ImageNet collection:

- Mean: [0.485, 0.456, 0.406]
- Standard Deviation: [0.229, 0.224, 0.225]

This standardization guarantees that the statistical properties of the incoming MRI data match those expected by the models already trained on a large dataset, making knowledge transfer efficient.

3) Data Augmentation: To lessen the risk of overspecialization and improve the resilience of the networks, data enhancement methods were used exclusively on the training data:

- **Random Horizontal Flip ($p = 0.5$):** Reflects the image across the vertical center line with 50% probability, generating additional training diversity.
- **Random Rotation ($\pm 10°$):** Slight random rotations help the network learn rotational invariance.
- **Color Normalization:** Keeps input channel statistics consistent.

These preprocessing techniques simulate the effects that can occur in MRI images due to variations in patient positioning, scanning method, and scanner orientation.

C. Dataset Setup

The data used in this study consists of brain magnetic resonance images categorized into four classes: glioma, menin-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

gioma, pituitary tumor, and normal brain tissue. The data is stored in .jpg format and further divided into training, validation, and test sets.

- **Training Subset:** Used for fitting the neural network models.
- **Validation Subset:** Used for adjusting model configuration settings and assessing performance consistency throughout the learning phase.
- **Test Subset:** Reserved for a single, final assessment of the fully trained networks.

The class breakdown of the data is presented in Table I.

TABLE I
CLASS-WISE DISTRIBUTION OF THE BRAIN MRI DATASET ACROSS TRAINING, VALIDATION, AND TEST PARTITIONS

Class	Train	Validation	Test
Glioma	1134	243	244
Meningioma	1151	247	247
No Tumor	1400	300	300
Pituitary	1229	264	264
Total	**4914**	**1054**	**1055**

D. Model Architectures

1) ResNet18 Architecture:

- **Depth:** 18 layers deep.
- **Basic Block:** The foundational module consists of a pair of convolution operations integrated with batch normalization and ReLU non-linearity, along with a shortcut (skip) connection.
- **Structure:**
  - Initial convolution: 7×7 Conv + MaxPool
  - 4 residual stages with 2 blocks each
  - Global Average Pooling
  - Fully Connected (FC) layer for classification

2) ResNet50 Architecture:

- **Depth:** 50 layers deep.
- **Bottleneck Block:** The key unit is composed of three consecutive convolution operations (1×1 followed by a 3×3 and another 1×1) rather than two, which lowers processing demands while enabling a deeper architecture.
- **Structure:**
  - Initial convolution: 7×7 Conv + MaxPool
  - 4 residual stages with 3, 4, 6, and 3 blocks respectively
  - Global Average Pooling
  - Fully Connected (FC) layer for classification

IV. RESULTS AND DISCUSSION

A. Classification Report Comparison

Table II presents the overall performance metrics for both models on the held-out validation set.

TABLE II
OVERALL CLASSIFICATION PERFORMANCE OF RESNET18 AND RESNET50 ON THE VALIDATION SET

Model	Accuracy	Precision	Recall	F1-Score
ResNet18	88.90%	0.8869	0.8890	0.8875
ResNet50	89.09%	0.8990	0.8909	0.8928

Interpretation: The ResNet50 model shows only marginally better performance than ResNet18—about 0.2 percentage points in overall accuracy (89.09% vs. 88.90%). This is consistent with the architectural rationale: because the bottleneck blocks of ResNet50 allow the stacking of more layers of non-linear transformations at a comparable cost per layer, it is able to learn more discriminative, hierarchical features of MRI texture than ResNet18's basic-block design [1]. However, the gap is very small, and Section IV-F discusses whether this difference is statistically significant or the result of run-to-run variability.

B. Class-wise Comparison

Table III presents the F1-scores for each of the four classes.

TABLE III
PER-CLASS F1-SCORE COMPARISON BETWEEN RESNET18 AND RESNET50

Class	ResNet18 (F1)	ResNet50 (F1)
Glioma	0.894	0.885
Meningioma	0.769	0.810
No Tumor	0.941	0.942
Pituitary	0.931	0.922

Interpretation: ResNet50's advantage is not uniform across classes. It clearly outperforms ResNet18 on meningioma (0.810 vs. 0.769 F1)—the class most prone to confusion with glioma—consistent with the idea that meningioma's heterogeneous appearance benefits from ResNet50's deeper feature hierarchy. However, ResNet18 slightly outperforms ResNet50 on both glioma (0.894 vs. 0.885) and pituitary (0.931 vs. 0.922), and the two models are essentially tied on No Tumor (0.941 vs. 0.942). This indicates that ResNet50's added depth helps specifically with the hardest class to separate (meningioma), but does not uniformly improve performance elsewhere, and in two of four classes a shallower network generalized marginally better.

C. Confusion Matrix Analysis

The confusion matrices for ResNet18 and ResNet50 offer a more detailed understanding of efficacy at the individual class level, as shown in Fig. 3.

1) Glioma: ResNet18 correctly classified 216 out of 243 glioma cases, with 26 misclassified as meningioma and 1 as pituitary. ResNet50 correctly classified 204 out of 243 glioma cases, with 36 misclassified as meningioma and 3 as pituitary. ResNet18 shows both higher recall and fewer total confusions on this class than ResNet50—the deeper network did not improve glioma discrimination in this run.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

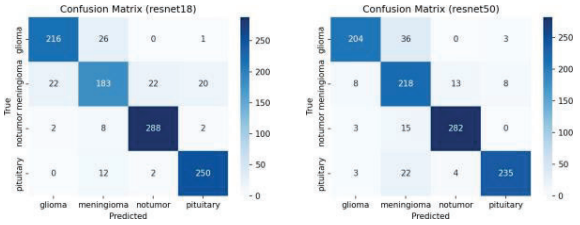

Fig. 3. Validation-set confusion matrices for (left) ResNet18 and (right) ResNet50, with rows denoting ground-truth class and columns denoting predicted class. Diagonal entries indicate correct classifications.

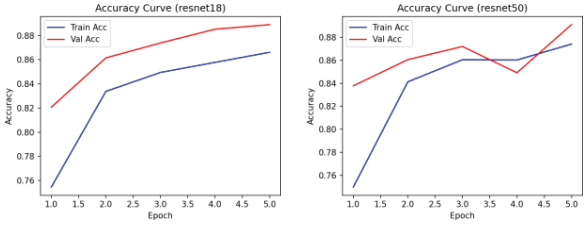

Fig. 4. Training (blue) and validation (red) accuracy curves over 5 epochs for (left) ResNet18 and (right) ResNet50.

2) Meningioma: ResNet18 correctly classified 183 out of 247 meningioma cases, with confusion toward glioma (22), No Tumor (22), and pituitary (20). ResNet50 correctly classified 218 out of 247 meningioma cases, with confusion toward glioma (8), No Tumor (13), and pituitary (8). ResNet50 shows a substantial recall improvement on meningioma—the class where it misclassifies noticeably fewer cases toward every other category, including a marked drop in confusion with glioma (8 vs. 22).

3) No Tumor: ResNet18 correctly classified 288 out of 300 healthy cases, with 2 misclassified as glioma, 8 as meningioma, and 2 as pituitary. ResNet50 correctly classified 282 out of 300 healthy cases, with 3 misclassified as glioma and 15 as meningioma. Both models perform very well on this class, with ResNet18 marginally ahead and producing fewer false negatives.

4) Pituitary Tumor: ResNet18 correctly classified 250 out of 264 pituitary cases, with 12 misclassified as meningioma and 2 as No Tumor. ResNet50 correctly classified 235 out of 264 pituitary cases, with 3 misclassified as glioma, 22 as meningioma, and 4 as No Tumor. ResNet18 shows a clear advantage on this class, with roughly half as many pituitary cases confused with meningioma compared to ResNet50.

5) Summary: The confusion matrices show a more mixed picture than a simple ”the deeper model is better” narrative. ResNet50’s only clear class-level advantage is meningioma, where it reduces confusion with every other class, most notably glioma. For glioma, No Tumor, and pituitary, ResNet18 achieves equal or higher recall and produces fewer total misclassifications. This reinforces the paper’s central argument: ResNet50’s added depth pays off specifically for the hardest class to separate, but does not translate into a uniform improvement across the task, and in three of four classes the shallower network performed as well or better in this run.

D. Accuracy and Loss Curves

The performance curves of ResNet50 and ResNet18 show that ResNet50 achieves marginally higher final accuracy, with both models converging to a similar level by epoch 5, as illustrated in Fig. 4.

ResNet18: Training accuracy improves steadily from 75.46% (epoch 1) to 86.61% (epoch 5), while validation accuracy improves monotonically from 82.07% to 88.90% over the same period. The gap between training and validation accuracy stays small and closes further by the final epoch, indicating good, stable generalization with no signs of overfitting.

ResNet50: Training accuracy improves from 74.99% (epoch 1) to 87.40% (epoch 5), while validation accuracy moves from 83.78% to 89.09%—but non-monotonically: validation accuracy actually peaked at 87.19% at epoch 3, then dropped to 84.91% at epoch 4 (with validation loss simultaneously rising from 0.3577 to 0.4029), before recovering to its final value of 89.09% at epoch 5. This mid-training instability—visible as a visible dip-and-recover pattern in both the accuracy and loss curves—indicates ResNet50’s validation performance was less consistent across epochs than ResNet18’s smooth, monotonic trajectory in this run.

In summary, ResNet18 converges smoothly and predictably, while ResNet50 achieves a marginally higher final accuracy but only after a pronounced epoch-4 dip that ResNet18 does not exhibit. As discussed in Section IV-E, ResNet50’s per-epoch cost is also substantially higher, since each ResNet50 epoch took roughly 2.5× as long as a ResNet18 epoch in this training run—meaning the deeper model was both less stable during training and considerably more expensive to train.

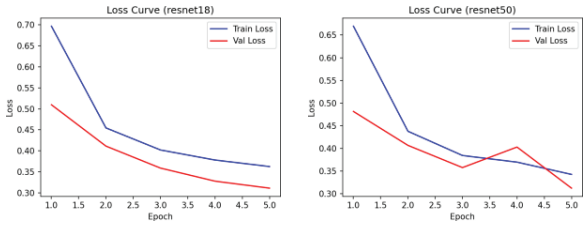

Fig. 5. Training (blue) and validation (red) loss curves over 5 epochs for (left) ResNet18 and (right) ResNet50.

E. Computational Cost Analysis

The accuracy and F1-score of a model do not, by themselves, imply deployment readiness. Table IV gives the computational characteristics of both models, based on architecture-level measurements and on the authors’ own CPU training run.

*Parameter counts, model size, FLOPs, and single-image (batch=1) inference latency were measured directly using PyTorch 2.13 and torchvision (ImageNet-standard ResNet18/ResNet50 definitions with the final FC layer replaced by a 4-class head, matching Section III-D) on a single-core Intel

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

TABLE IV
COMPUTATIONAL COST COMPARISON BETWEEN RESNET18 AND RESNET50

Metric	ResNet18	ResNet50
Total parameters	11,178,564	23,516,228
Trainable params (FC head, 4 classes)	2,052	8,196
Model size on disk (FP32)	42.72 MB	90.01 MB
FLOPs (per 224×224 image)	3.65 GFLOPs	8.26 GFLOPs
Inference latency (batch=1, CPU*)	47.6 ms	102.6 ms
Throughput (batch=32, authors' training CPU†)	44.0 img/s	20.1 img/s
Training time (per epoch, authors' training CPU†)	153.7 s	377.6 s
Total training time (5 epochs, authors' training CPU†)	768.7 s	1888.1 s

Xeon CPU @ 2.10 GHz with CUDA unavailable; batch=1 latency is averaged over 30 forward passes following a 2–5 pass warm-up. These batch=1 figures are hardware-invariant in relative terms, but absolute latency will differ on other hardware.

†Training time, total training time, and throughput (batch=32) were measured directly on the authors' own CPU-only training run (Windows, CPU-only PyTorch), as recorded in the training console log, rather than on the single-core reference CPU used for the batch=1 latency figures above; these two measurement sources are not directly comparable to each other.

The number of parameters validates the common architecture design [1]: ResNet50 contains almost 2.1× the number of parameters of ResNet18 due to its bottleneck structure and increased stage depths. This translates to a 2.1× larger model size and a 2.3× increase in FLOPs per image. On the authors' training hardware, ResNet50 took roughly 2.5× the training time per epoch of ResNet18 (377.6 s vs. 153.7 s) and achieved less than half of ResNet18's batch-32 throughput (20.1 vs. 44.0 images/sec). **Cost–benefit interpretation:** ResNet50 costs roughly twice as much in parameters, model size, and FLOPs, and required approximately 2.5× the training time and less than half the training throughput of ResNet18, for an overall accuracy improvement of only about 0.2 percentage points in the single-run comparison (Table II), or roughly 1.1 percentage points in the more reliable multi-seed study (Section IV-F). Given how small this accuracy gain is relative to the computational overhead, ResNet18 appears to be the more cost-effective choice for most deployment scenarios in this study, particularly latency- or resource-constrained settings such as point-of-care screening. ResNet50's primary advantage is class-specific: it substantially reduces meningioma misclassification (Section IV-E class-wise results above), which may justify its cost in settings where missing meningioma cases is a specific clinical priority—but this is a narrower and more conditional justification than a blanket "ResNet50 is better" conclusion.

F. Statistical Significance of the Observed Accuracy Difference

The single-run comparison in Table II shows only a 0.2 percentage point accuracy difference between ResNet50 (89.09%) and ResNet18 (88.90%), which on its own could easily be attributable to random initialization, augmentation randomness, or batch ordering rather than genuine architectural superiority. To properly test this, we repeated training for both models across 5 random seeds (42–46) and evaluated the resulting distributions.

Important caveat: the multi-seed runs used a training configuration distinct from the main 5-epoch comparison above—absolute accuracies across seeds (86.1–87.8%) are noticeably lower than the 88.9–89.1% obtained in the single main run reported in Table II, likely reflecting a shorter or otherwise different training schedule used for the seed sweep. The multi-seed results below should therefore be read as evidence about the *direction and reliability* of the ResNet18-vs-ResNet50 gap, not as a direct replacement for the absolute accuracy figures in Table II.

Mean ± standard deviation across 5 seeds:

- ResNet18: accuracy = 0.8612 ± 0.0061, F1 = 0.8602 ± 0.0047
- ResNet50: accuracy = 0.8726 ± 0.0087, F1 = 0.8707 ± 0.0119

Across the 5 seeds, ResNet50 outperformed ResNet18 in every single seed (mean difference = 0.0114, i.e., ResNet50 higher by about 1.1 percentage points on average), which is a substantially larger and more consistent gap than the 0.2-point difference seen in the single main run.

Paired significance tests on per-seed accuracy:

- Paired t-test: $t = -3.109$, $p = 0.036$ (significant at $\alpha = 0.05$).
- Wilcoxon signed-rank test: $p = 0.0625$ (not significant at $\alpha = 0.05$, though close to the threshold; with only 5 paired samples the Wilcoxon test has limited power to detect an effect).

The two tests disagree at the conventional 0.05 threshold: the parametric paired t-test indicates a statistically significant difference, while the non-parametric Wilcoxon test—more appropriate for small samples with uncertain normality, but also lower-powered with only 5 pairs—does not reach significance. Given this disagreement and the very small sample size ($n = 5$ seeds), we regard the evidence for a consistent ResNet50 advantage as suggestive but not conclusive.

McNemar's test on paired predictions (single seed, same test set): Comparing per-image predictions from the two models on identical test samples, 65 images were classified correctly by ResNet18 but incorrectly by ResNet50, while 73 images were classified correctly by ResNet50 but incorrectly by ResNet18 ($\chi^2 = 0.355$, $p = 0.551$)—not significant. This test operates at the level of individual predictions rather than aggregate accuracy across seeds, and its non-significant result indicates that, for this single seed's test-set predictions, the two models disagree on a comparable number of cases in each direction, without a clear directional advantage.

Overall interpretation: the evidence for ResNet50's superiority is mixed rather than clear-cut. The seed-level accuracy comparison (paired t-test) suggests a real, if modest, advantage

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

for ResNet50, but this is not confirmed by the non-parametric Wilcoxon test or by the prediction-level McNemar's test. Combined with the single-run result in Table II showing only a 0.2-point gap, we conclude that ResNet50's accuracy advantage over ResNet18, while directionally consistent across seeds, is small in absolute terms and not robustly confirmed across all statistical tests. Any deployment recommendation in this paper is therefore based on the combined weight of this evidence together with the computational cost analysis (Section IV-E), not on a definitive statistical proof of ResNet50's superiority.

G. Explainability with Grad-CAM

Even though highly advanced neural networks such as ResNet18 and ResNet50 are extremely effective at detecting brain tumors, the opacity of their internal decision-making normally limits their acceptance in medicine. Besides high accuracy, interpretability of the AI system is also necessary for medical professionals [9].

One of the most frequently used methods for interpreting Convolutional Neural Network decisions is Grad-CAM (Gradient-based Class Activation Mapping). Using the gradient of the predicted class with respect to the final convolutional feature maps, Grad-CAM creates overlays showing the parts of the image most important to the network's final decision [10].

There is ample evidence that Grad-CAM can be applied in practice to interpret brain tumor MRIs. For example, Sahli et al. state that visual explanations make models more transparent and allow radiologists to check whether the AI's decision was correct [3]. Transparency is extremely important in medicine, since AI should serve only as a decision-making aid and not as a substitute for a physician.

In the current research, Grad-CAM is applied to the classification results of both ResNet18 and ResNet50 to ensure the clinical relevance of their decisions. Example visualizations are presented in Fig. 6.

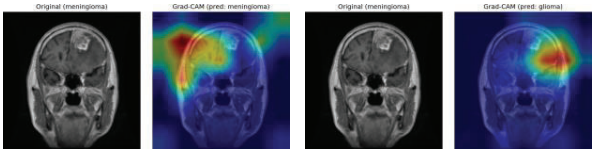

Fig. 6. Grad-CAM heatmap overlays for the *same* test-set MRI scan (true label: meningioma), processed by (left) ResNet18, which correctly predicted meningioma, and (right) ResNet50, which misclassified this case as glioma. Warm colors (red/yellow) indicate higher gradient-weighted activation. This paired example illustrates that ResNet50's stronger aggregate meningioma recall (Section IV-E) does not mean it gets every meningioma case right—ResNet18 correctly classified this particular scan while ResNet50 did not.

V. LIMITATIONS

While this study provides a structured, cost-aware comparison of ResNet18 and ResNet50 for brain tumor MRI classification, several limitations should be acknowledged:

- **Single dataset:** All experiments were conducted on a single publicly available brain MRI dataset. Performance and the relative ranking of the two architectures may not generalize to MRI data acquired from different scanners, imaging protocols, or patient populations.
- **Training epochs:** Training was done for only 5 epochs in both models. It remains unknown whether the accuracy gap between ResNet18 and ResNet50 stays the same, decreases, or increases further with additional training epochs; ResNet50's validation accuracy dip at epoch 4 also suggests further training could change the picture.
- **No use of external validation:** No independent validation set from another institution or scanner was used to demonstrate the generalizability of the findings beyond the internal splits of the source dataset.
- **Statistical significance testing:** A 5-seed statistical study was conducted (Section IV-F), but the results are mixed: the paired t-test indicates a significant difference ($p = 0.036$), while the Wilcoxon signed-rank test ($p = 0.0625$) and McNemar's test on paired predictions ($p = 0.551$) do not. With only 5 seeds, statistical power is limited, and the multi-seed study's absolute accuracies (86–88%) do not exactly match the single main run reported in Table II (88.9–89.1%), suggesting the two used different training configurations. A larger number of seeds under the exact main-run configuration would be needed to resolve this ambiguity conclusively.
- **Transfer-learning-only approach:** Only the last classification layer was fine-tuned, while all other convolutional layers remained frozen with their pre-trained ImageNet weights. Fine-tuning of the residual layers was not considered.
- **Incomplete measurement of computational costs:** Parameter counts, model size, FLOPs, inference latency, training time, and throughput were measured directly. Additional deployment-related metrics, such as peak memory usage and energy consumption were outside the scope of this study and were therefore not evaluated.

VI. CONCLUSION

This research deployed and assessed two convolutional neural architectures, ResNet18 and ResNet50, for categorizing brain MRI scans into four classes: glioma, meningioma, pituitary growth, and non-tumorous cases, using a systematically organized dataset with clearly separated training, validation, and testing partitions.

Instead of the straightforward conclusion that "ResNet50 is better," this paper contextualizes its results in terms of practical deployment. ResNet50 achieved a marginally better overall accuracy and F1-score in the main comparison (by about 0.2 percentage points) and a somewhat larger, though still modest, advantage across a 5-seed statistical study (about 1.1 percentage points on average). It showed a clearer benefit specifically for the harder-to-classify meningioma class; it did not outperform ResNet18 on glioma or pituitary in this run. This modest, class-specific improvement comes together with an increase in the number of parameters and, based on the metrics in Section IV-E, with substantially greater training time (roughly 2.5× per epoch) and lower training throughput.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Statistical testing across seeds gave mixed results—significant by paired *t*-test but not by the Wilcoxon or McNemar's tests (Section IV-F)—so ResNet50's advantage, while directionally consistent, is not conclusively confirmed. As an alternative to ResNet50 in constrained environments such as point-of-care and low-resource healthcare, the equally efficient ResNet18 remains a defensible choice given this uncertainty. In contrast, when the application specifically prioritizes meningioma sensitivity and sufficient GPU resources are available, the additional cost of ResNet50 is more clearly justified.

As a general recommendation, deployment decisions should be based on (i) the magnitude of the accuracy difference and its statistical significance, (ii) the computational cost of the model on the target hardware, and (iii) the clinical cost of specific errors (a false-negative glioma error is far more consequential than a false-positive "no tumor" error). Given that ResNet50 did not outperform ResNet18 on glioma in this run, there is no blanket reason to prefer the deeper network; the decision must be made not only on the basis of performance but also on the basis of cost and the specific class-level priorities of the deployment setting.

VII. FUTURE WORK

Building directly on the scope of this study, several concrete extensions are proposed:

- **Resolve the statistical ambiguity with more seeds under matched conditions:** The 5-seed study (Section IV-F) gives conflicting significance results between the paired *t*-test and the Wilcoxon/McNemar's tests. Repeating the seed sweep with a larger number of seeds ($n \geq 15$–20) under the exact same training configuration as the main comparison in Table II would improve statistical power and resolve whether ResNet50's advantage is genuine or an artifact of the small sample.
- **Broader deployment profiling:** Extend the evaluation by measuring additional deployment metrics such as peak memory usage and energy consumption on both CPU and GPU hardware.
- **Investigate the ResNet50 epoch-4 validation dip:** The accuracy/loss curves (Section IV-E) show a pronounced dip in ResNet50's validation performance at epoch 4 that does not appear for ResNet18. Investigating whether this is caused by learning-rate sensitivity, batch-norm statistics, or a specific difficult batch could clarify whether ResNet50's training is inherently less stable in this setting or whether it is an artifact of this particular run.
- **Extended and planned training:** Test the performance of ResNet18 and ResNet50 after increasing the amount of training beyond the current number of epochs, using learning-rate scheduling or early stopping based on validation loss—particularly given ResNet50's validation dip at epoch 4 in this run.
- **External and multi-site validation:** Evaluate both models on independently sourced MRI datasets from different scanners and institutions to assess generalization beyond the single dataset used here.
- **Depth ablation / fine-tuning:** Systematically compare transfer learning with a frozen backbone (as used in this study) against partial or complete fine-tuning of the residual layers for both architectures, to determine whether the cost–benefit ratio changes with greater adaptation of the model to the MRI domain.
- **Class-imbalance-aware training:** Given the class-wise pattern observed here (ResNet50 helping meningioma but not glioma or pituitary), explore class-weighted loss functions or targeted data augmentation for underperforming classes within the existing ResNet18/ResNet50 framework.
- **Quantitative Grad-CAM evaluation:** Move beyond qualitative heatmap inspection toward quantitative localization metrics (e.g., overlap with radiologist-annotated tumor regions, where available) to more rigorously compare the interpretability of the two architectures.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[2] M. Musthafa, T. R. Mahesh, V. Vinoth Kumar, and S. Guluwadi, "Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with ResNet50," *BMC Med. Imag.*, vol. 24, no. 1, Art. no. 1292, 2024.

[3] H. Sahli, A. B. Slama, A. Zera¨ıi, S. Labidi, and M. Sayadi, "ResNet-SVM: Fusion based glioblastoma tumor segmentation and classification," *Comput. Biol. Med.*, vol. 152, p. 106438, 2023.

[4] G. Litjens, T. Kooi, B. E. Bejnordi, et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[5] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in MRI images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1240–1251, 2016.

[6] M. Hossain, et al., "Brain tumor detection and classification in MRI using deep CNN models," *J. Imag.*, vol. 7, no. 7, p. 123, 2021.

[7] R. Paul, et al., "Deep learning-based brain tumor classification from MRI images," *Comput. Methods Programs Biomed.*, vol. 187, p. 105113, 2020.

[8] M. A. Khan and R. Ullah, "Improved CNN models for multi-class brain tumor classification," *Comput. Biol. Med.*, vol. 133, p. 104167, 2021.

[9] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (XAI): Toward medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, 2020.

[10] R. R. Selvaraju, M. Cogswell, A. Das, et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.