iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VIII **Month of publication:** August 2026

DOI: https://doi.org/10.22214/ijraset.2026.84747

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VIII Aug 2026- Available at www.ijraset.com

VGG16 and ResNet50 for Brain Tumor Classification: Comparative Performance and Grad-CAM++ Analysis

Ganeshula Sai Raghava
M. Tech, CSE Department, UCEK, JNTU Kakinada, Andhra Pradesh, India

Abstract: ***Reliable categorization of brain tumors from magnetic resonance imaging (MRI) is a prerequisite for timely clinical intervention, yet the great majority of transfer-learning studies in this space report a single backbone in isolation, leaving open the question of how much of the reported performance is attributable to the chosen architecture rather than to dataset or training choices. This paper presents a controlled comparative evaluation of two ImageNet-pretrained convolutional backbones, VGG16 and ResNet50, for four-class brain tumor classification (glioma, meningioma, pituitary tumor, no tumor) on a merged corpus of 7,200 MRI slices drawn from the Figshare, SARTAJ, and Br35H repositories. Both backbones are fitted with an identical classification head and trained under an identical two-phase (frozen, then partially fine-tuned) protocol, so that any accuracy differential can be attributed to the backbone rather than to confounding methodological variation. On a held-out test set of 1,600 images, VGG16 attains 94.31% accuracy with a weighted F1-score of 94.20%, while ResNet50 attains 94.00% accuracy with a weighted F1-score of 93.89%; VGG16 is accordingly identified as the better-performing architecture under the present experimental conditions. To move beyond an opaque class label, Grad-CAM++ is applied to the selected model to generate a visual attribution map for each classification decision, highlighting the image regions that most strongly influenced the predicted tumor class. The resulting comparative and interpretability findings are intended as a controlled, architecture-isolated reference point for subsequent work in this line of research rather than as a claim of universally superior performance.***
Brain Tumor Classification, VGG16, ResNet50, Transfer Learning, Grad-CAM++, Explainable AI, MRI.

I. INTRODUCTION

Brain tumors span a wide spectrum of pathology, from rapidly progressing gliomas to comparatively slow-growing meningiomas and pituitary adenomas, and the specific category present in a scan has a direct bearing on the treatment pathway a clinician will pursue. Magnetic resonance imaging is the standard modality for intracranial assessment because of its favourable soft-tissue contrast, but manual reading of MRI slices is time-consuming, requires specialist training, and is subject to disagreement between readers, a difficulty that becomes more pronounced when two tumor categories present with overlapping radiological appearance.

Convolutional neural networks pretrained on large natural-image corpora, most notably VGG16 and ResNet50, have become a standard tool for MRI-based tumor classification because medical imaging datasets are rarely large enough on their own to train a deep network from random initialization. A number of recent studies have applied one or the other of these backbones to brain MRI classification and reported strong accuracy figures, but comparisons across studies are difficult to interpret because the backbones are typically evaluated under different datasets, preprocessing pipelines, and training schedules, so an observed accuracy gap cannot cleanly be attributed to the architecture itself.

The present work addresses this gap directly. Rather than proposing a new architecture, it holds every experimental variable constant between VGG16 and ResNet50 – the dataset, the preprocessing and augmentation pipeline, the classification head, and the training schedule – and reports the resulting head-to-head comparison. Because a bare class label offers a clinician no insight into why a model reached its decision, Grad-CAM++ is additionally applied to the selected backbone to produce a visual attribution map for each prediction.

This framing is deliberately narrower than an end-to-end diagnostic system. Rather than attempting to simultaneously optimize classification accuracy, pixel-level localization, and deployment engineering within a single study, the present paper isolates the classification-and-interpretability component and evaluates it under conditions designed to make the VGG16-versus-ResNet50 comparison as clean as possible.

This scoping choice reflects a broader observation about the brain-tumor-classification literature: individual papers frequently report a single strong accuracy figure for a single backbone, but comparisons across papers are difficult to interpret because the backbones being compared were rarely trained under matched conditions. A controlled, single-study comparison of two of the most widely used backbones, reported alongside their associated Grad-CAM++ behavior, is intended to serve as a more directly interpretable reference point than a cross-study comparison of accuracy figures obtained under differing experimental conditions. The contributions of this paper are:

1) A controlled, architecture-isolated comparison of VGG16 and ResNet50 for four-class brain tumor classification, using an identical classification head and an identical two-phase training protocol for both backbones, so that the observed performance difference can be attributed to the backbone rather than to confounding methodological variation.

2) Identification of VGG16 as the better-performing architecture under the present experimental conditions, together with a per-class and confusion-matrix-level error analysis of both backbones.

3) Application of Grad-CAM++ to the selected classifier to provide a visual, region-level justification for individual classification decisions, going beyond a bare predicted label and confidence score.

II. RELATED WORK

Several studies have examined VGG16 as a backbone for brain MRI classification. Prabhas et al. (ref1?) benchmarked an ensemble CNN, MobileNetV2, VGG16, and a Vision Transformer on a merged Figshare, SARTAJ, and Br35H corpus and found VGG16 to be the strongest individual architecture. Shib et al. (ref2?) and Sahoo et al. (ref3?) each report VGG16-based classification pipelines in the mid-90s accuracy range following targeted data augmentation and hyperparameter tuning, while Thiyagu et al. (ref4?) obtain further gains in multiclass separability by modifying the VGG16 classification head. Patil et al. (ref5?) investigate how the choice of which convolutional layers to fine-tune affects transfer-learning performance and pair this with preliminary interpretability analysis.

A separate body of work explores architectures beyond conventional CNN backbones. Krishnan et al. (ref6?) propose a rotation-invariant Vision Transformer to address orientation sensitivity in MRI slices, and Jansi et al. (ref7?) incorporate multiple MRI sequences in a multimodal classification pipeline. Tayefeh et al. (ref8?) combine a Vision Transformer with Faster R-CNN for lightweight tumor localization, while Golkarieh et al. (ref9?) report that EfficientNetV2 outperforms VGG19, ResNet50, and InceptionV3 on a seventeen-category dataset. Pilaoon et al. (ref10?) apply majority voting across CNN predictions for glioma subtype classification, Jetlin and Annabel (ref11?) combine pyramidal feature aggregation with quantized dilated convolutions, and Zeng and Zhang (ref12?) show that a carefully engineered SIFT-based bag-of-features support vector machine pipeline remains competitive with deep-learning baselines. Nimmagadda and Devi (ref13?) reformulate tumor identification as an object-detection task using YOLOv7, Shankar et al. (ref14?) report a deep CNN classification pipeline evaluated on standard MRI benchmarks, Ali et al. (ref15?) examine combined segmentation-and-classification pipelines, and Disci et al. (ref16?) compare several pretrained backbones under a common transfer-learning protocol for brain tumor classification.

Explainability has received comparatively less systematic treatment in this literature. Grad-CAM (ref20?) and its refinement Grad-CAM++ (ref17?) are the most widely used visual-attribution techniques for convolutional classifiers, with Grad-CAM++ replacing the single globally averaged gradient weight per channel used by the original formulation with a pixel-wise weighted combination derived from higher-order derivatives, which produces tighter and better-localized attribution maps for small or off-center regions of interest – a property directly relevant to tumor regions that frequently occupy a small fraction of an MRI slice. Where explainability is discussed in the surveyed classification literature, it is typically appended as an optional visualization rather than evaluated as a systematic component of the pipeline (ref5?), (ref6?).

Two architectural building blocks underpin the present comparison. VGG16, introduced by Simonyan and Zisserman (ref18?), established that stacking small 3×3 convolutional filters to considerable depth is an effective and simple design for image recognition, while ResNet50, introduced by He et al. (ref19?), showed that residual (skip) connections allow substantially deeper networks to be trained without the degradation in accuracy that plain deep stacks otherwise exhibit. Both remain widely used as ImageNet-pretrained feature extractors in medical-imaging transfer learning, which motivates their selection as the two backbones compared in this work.

A further consideration that recurs throughout the surveyed literature is the choice of how much of a pretrained backbone to adapt during fine-tuning. Some studies fine-tune the entire convolutional stack from the first epoch, while others freeze the backbone entirely and train only a newly attached head. Both extremes carry a recognizable risk: full-network fine-tuning from a randomly initialized head can propagate large, poorly calibrated gradients back into pretrained filters before those filters have any task-relevant signal to adapt to, while a permanently frozen backbone forfeits the opportunity to specialize the deepest, most task-specific

layers to the target domain. Several of the transfer-learning studies referenced above (ref2?), (ref4?), (ref5?) report exploring intermediate strategies, such as unfreezing a limited number of late convolutional blocks after an initial head-only phase, without always reporting the precise schedule used or applying it identically across more than one backbone. This inconsistency is one further source of the cross-study incomparability that motivates the controlled design adopted in the present paper.

Data provenance is a related but distinct source of cross-study variation. Because Figshare, SARTAJ, and Br35H are each individually imbalanced with respect to the four target classes – some containing an abundance of glioma and meningioma slices with comparatively few explicit no-tumor examples, and others the reverse – studies that draw from only one of the three sources implicitly adopt a different underlying class distribution than studies that merge all three, even when the class labels themselves are nominally identical across studies.

This distinction rarely receives explicit discussion in the surveyed literature, yet it directly affects the comparability of any two reported accuracy figures, since a classifier trained and tested on a distribution skewed toward the easier no-tumor and pituitary classes would be expected to report a systematically higher aggregate accuracy than one trained and tested on a distribution that more heavily weights the harder glioma-meningioma pair, independent of any genuine difference in model quality. The present paper avoids this particular source of ambiguity by using the same merged, class-balanced four-class corpus for both backbones under comparison, as detailed in Section III-B.

Interpretability techniques for convolutional classifiers have followed a broadly similar trajectory to the classification literature itself: an initial period of rapid methodological development, illustrated by the progression from class activation mapping through Grad-CAM (ref20?) to Grad-CAM++ (ref17?), followed by increasingly routine application of the most mature technique to new domains, including brain MRI. What is comparatively less common in the surveyed brain-tumor classification literature is a systematic reporting of Grad-CAM or Grad-CAM++ output alongside the quantitative comparison that motivated the choice of the underlying classifier, rather than as a separate, loosely connected demonstration.

The dataset dimension of this literature is worth noting separately from the architectural dimension. Several of the surveyed studies construct their own single-source MRI collection, while others, following Prabhas et al. (ref1?), merge multiple public repositories into a larger four-class corpus.

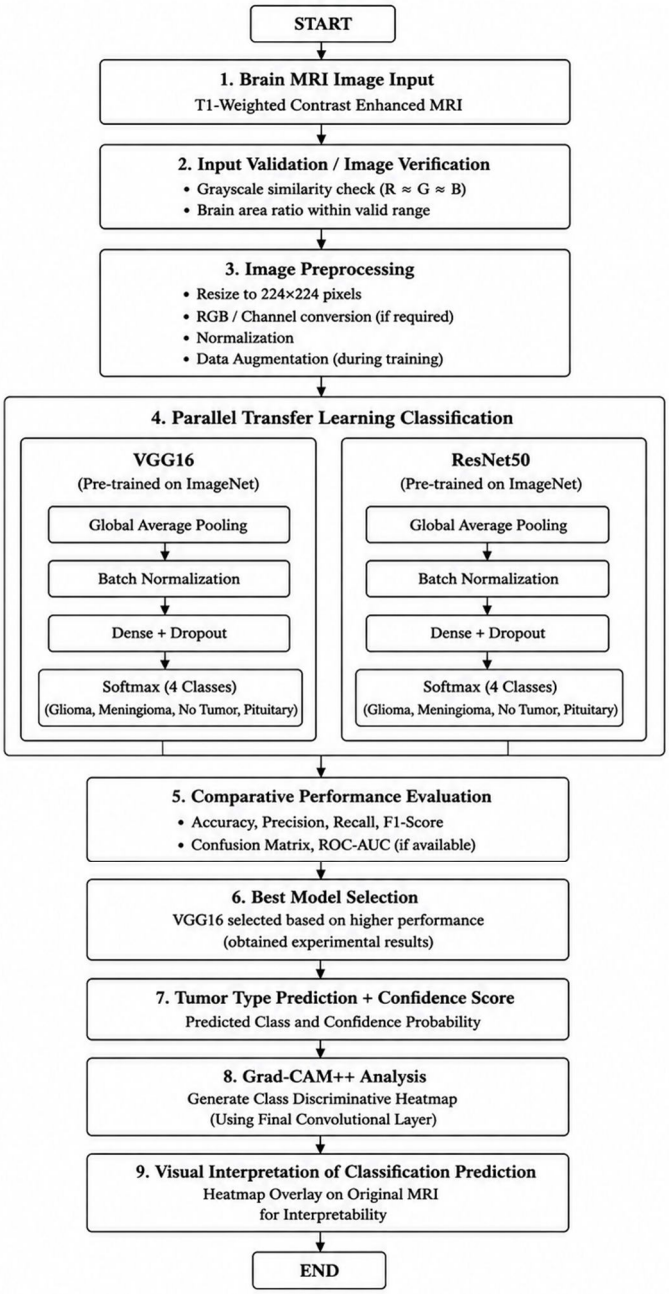
A merged corpus offers a practical advantage for the specific comparison undertaken in this paper: because Figshare, SARTAJ, and Br35H individually emphasize different subsets of the four target classes, none of the three sources alone provides a balanced four-way split, whereas the union of all three does. Adopting the same merged corpus and class definition used in (ref1?) additionally means that the present comparison is evaluated on a dataset construction that has already been used elsewhere in this literature, rather than on a bespoke split that would be difficult for later work to reproduce exactly.

Collectively, the surveyed literature establishes VGG16 and ResNet50 as strong individual candidates for brain MRI classification, but rarely isolates the two under matched experimental conditions within a single study, and rarely pairs the resulting classifier with a systematic, per-prediction explainability output. This paper targets precisely that combination: a controlled head-to-head comparison of the two backbones, coupled with Grad-CAM++ interpretation of the selected model's predictions.

III. PROPOSED METHODOLOGY

A. System Overview

Fig. 1 presents the overall methodology followed in this study. A raw MRI slice is first checked for validity, then preprocessed, then passed in parallel through the VGG16 and ResNet50 classification pipelines. The two backbones are compared on the held-out test set, the higher-performing backbone is retained as the deployed classifier, and its prediction – together with a confidence score – is passed to Grad-CAM++ for visual interpretation.



Proposed Methodology for Comparative Brain Tumor Classification and Grad-CAM++ Analysis.

B. Dataset

The classification corpus merges three publicly available MRI collections – Figshare, SARTAJ, and Br35H – into a single four-class set covering glioma, meningioma, pituitary tumor, and no-tumor slices, since no individual source collection by itself offers a balanced representation of all four categories together with a genuine no-tumor class. The merged corpus totals 7,200 JPEG images, split into 5,600 training images and 1,600 held-out test images (Table 1), with a further 20% of the training partition (1,120 images) set aside for validation during model fitting, leaving 4,480 images for gradient updates in each epoch. This three-way split ensures that the test accuracy reported in Section IV is never influenced by images seen during training or hyperparameter selection. Because both backbones are trained on exactly this same split, preprocessing pipeline, and training schedule, the comparison in Section IV constitutes a controlled, architecture-isolated evaluation rather than a comparison confounded by differing experimental conditions.

Class Distribution of the Classification Dataset

Class	Train	Test	Total
Glioma	1400	400	1800
Meningioma	1400	400	1800
No Tumor	1400	400	1800
Pituitary	1400	400	1800
Total	5600	1600	7200

C. Image Preprocessing

Every slice is resized to 224×224 pixels, the native input resolution expected by both VGG16 and ResNet50 under ImageNet pretraining, so that the pretrained convolutional filters of either backbone remain directly applicable without training from scratch. Training-time augmentation is applied on the fly and consists of rotation ($\pm 15°$), width and height shifts (0.05), shear (0.05), zoom (0.1), horizontal flipping, and brightness perturbation ($0.9–1.1\times$), so that the classifier learns to tolerate the positional and illumination variation typical of real clinical acquisitions rather than memorizing a fixed pixel arrangement. Each backbone's own channel-normalization routine is applied immediately before an image reaches that backbone, since VGG16 and ResNet50 were pretrained under different channel-scaling conventions and mismatching this step measurably degrades transfer-learning performance. Applying the same resize, augmentation, and split to both backbones, differing only in the backbone-specific normalization each architecture requires, is what allows the comparison in Section IV to isolate the effect of architectural choice.

D. VGG16 and ResNet50 Transfer Learning

Both classification backbones are instantiated from ImageNet-pretrained weights with the original fully connected classification layers removed, and both are appended with an identical, purpose-built head rather than either backbone's native dense layers. The head consists of a Global Average Pooling layer, which collapses the final convolutional feature map into a single vector per channel and confers a degree of invariance to small shifts in tumor position; a Batch Normalization layer, which stabilizes the scale of the pooled representation; a first dense layer of 512 units with ReLU activation and a dropout rate of 0.4; a second dense layer of 256 units with ReLU activation and a dropout rate of 0.3; and a final four-unit dense layer with softmax activation producing the class probability distribution over glioma, meningioma, no-tumor, and pituitary tumor. Attaching this same head, unmodified, to both convolutional bases is what allows any accuracy difference reported in Section IV to be attributed to the backbone itself rather than to an asymmetry in head capacity.

Each backbone is trained under an identical two-phase schedule. In Phase 1, every layer of the backbone is frozen and only the newly attached head is trained, for up to 20 epochs with the Adam optimizer at a learning rate of 10^{-4}; this allows the head to reach a reasonable operating point against fixed, high-quality pretrained features before any backbone weights are disturbed. In Phase 2, only the deepest convolutional block of the backbone is unfrozen – block5 for VGG16 and the final residual block (conv5_block) for ResNet50 – and training continues for a further 20 epochs at a reduced learning rate of 10^{-5}, since the deepest layers encode the most task-specific visual patterns and are therefore the most worth adapting, while the reduced learning rate limits the risk of catastrophically overwriting pretrained weights. Throughout both phases, identical callbacks are applied to each backbone: early stopping on validation loss with a patience of six epochs, a model checkpoint retaining the weights with the highest observed validation accuracy, and a learning-rate reduction on plateau (factor 0.3, patience 3, floor 10^{-7}).

Algorithm [alg:training] summarizes this two-phase procedure as applied identically to both backbones.

ImageNet-pretrained backbone B (VGG16 or ResNet50), training set D_{train}, validation set D_{val} Fine-tuned classifier M Attach head H: GAP → BatchNorm → Dense(512, ReLU, dropout 0.4) → Dense(256, ReLU, dropout 0.3) → Dense(4, softmax) $M \leftarrow B + H$; freeze all layers of B Phase 1 (head-only): train M on D_{train} for up to 20 epochs, Adam ($lr = 10^{-4}$), with early stopping and ReduceLROnPlateau on validation accuracy; checkpoint best weights Unfreeze deepest block of B (block5 for VGG16, conv5_block for ResNet50) Phase 2 (fine-tune): continue training M for up to 20 epochs, Adam ($lr = 10^{-5}$), same callbacks; checkpoint best weights Evaluate M on held-out test set D_{test}; record accuracy, precision, recall, F1 M with best validation checkpoint restored

E. Comparative Classification and Best Model Selection

Once both backbones have completed training, they are evaluated on the identical 1,600-image held-out test set using accuracy, precision, recall, F1-score, and confusion-matrix analysis (Section IV). The backbone attaining the higher test accuracy is retained as the deployed classifier for downstream tumor-type prediction and confidence scoring, while the other is kept only for the comparative analysis reported in this paper. Under the present experimental conditions, this selection is established in Section IV to favor VGG16.

The decision to select a single backbone for deployment, rather than retaining both and combining their outputs through, for instance, an averaging or voting ensemble, is intentional. An ensemble of VGG16 and ResNet50 could plausibly recover some of the errors unique to either backbone, but it would also roughly double the inference-time computational cost of every classification, and would complicate the Grad-CAM++ attribution step, since a single, well-defined convolutional feature map is required as the target layer for the attribution computation described in Section III-E. Because the accuracy gap between the two backbones in this study is modest, the added inference cost and interpretability complexity of an ensemble were judged not to be justified relative to simply deploying the stronger individual backbone; an ensemble comparison is nonetheless a reasonable direction for subsequent work and is noted as such in Section VII.

F. Grad-CAM++ Analysis

Grad-CAM++ (ref17?) is applied to the final convolutional layer of the deployed classifier in preference to the original Grad-CAM formulation (ref20?), because tumor regions in this dataset are frequently small, irregularly shaped, or off-center within the slice, and the coarse, uniformly averaged channel weighting used by standard Grad-CAM tends to under-highlight exactly these cases. Grad-CAM++ replaces the single global-average-pooled gradient weight per channel with a pixel-wise weighted combination of the gradients at each spatial location, derived from higher-order (second- and third-order) partial derivatives of the class score with respect to the target feature map. For a target class c and convolutional feature map A^k, the channel importance weight is computed as

$$w_k^c = \sum_{i,j} \alpha_{ij}^{kc} \cdot \mathrm{ReLU}\left(\frac{\partial Y^c}{\partial A_{ij}^k}\right),$$

where α_{ij}^{kc} are pixel-wise weighting coefficients derived from the second- and third-order partial derivatives of Y^c with respect to A_{ij}^k, so that pixels contributing more strongly and more consistently to the class score receive proportionally greater weight than under a simple average. The final saliency map is $L^c = \mathrm{ReLU}(\sum_k w_k^c A^k)$, upsampled to the original slice resolution and overlaid on the source slice with partial transparency.

The resulting heatmap serves purely as an attribution of which pixels most influenced the classification decision; it is not a segmentation output and does not delineate a tumor boundary. Classification, confidence score, and Grad-CAM++ attribution together form the three outputs of the pipeline: classification predicts the tumor category, the confidence score reports the model's predicted probability for that category, and Grad-CAM++ visualizes the image regions that most contributed to the prediction.

G. Implementation and Reproducibility Details

To support reproducibility of the comparison reported in Section IV, Table 2 consolidates the training hyperparameters applied identically to both backbones. All experiments were implemented in Python using the TensorFlow/Keras deep-learning framework, with model definition, training-loop management, and callback scheduling handled through the standard Keras functional API rather than a custom training loop, so that the frozen and fine-tuning phases could be expressed as two successive calls to the same high-level training routine with only the trainable-layer configuration and learning rate changed between them.

Training Hyperparameters (Identical for VGG16 and ResNet50)

Hyperparameter	Value
Input resolution	$224 \times 224 \times 3$
Optimizer	Adam
Phase 1 learning rate	10^{-4}
Phase 2 learning rate	10^{-5}
Phase 1 max epochs	20

Hyperparameter	Value
Phase 2 max epochs	20
Early stopping patience	6 epochs (validation loss)
LR reduction on plateau	factor 0.3, patience 3, floor 10^{-7}
Dropout (dense-512 layer)	0.4
Dropout (dense-256 layer)	0.3
Loss function	Categorical cross-entropy
Fine-tuned block (VGG16)	block5
Fine-tuned block (ResNet50)	conv5_block

Keeping every entry in Table 2 identical between the two backbones is the mechanism by which the comparison in Section IV is rendered architecture-isolated: the only aspect of training that is allowed to differ between VGG16 and ResNet50 is the name of the deepest block being unfrozen in Phase 2, since the two architectures do not share an identical internal block-naming convention, and even this difference refers to the same conceptual step – unfreezing the final high-level feature-extraction block – in both networks. No hyperparameter search specific to either backbone was performed; both were trained under the single configuration listed above, which favors comparability over squeezing the maximum achievable accuracy out of either architecture individually.

IV. EXPERIMENTAL RESULTS

All models were trained and evaluated using a single NVIDIA T4 GPU with the TensorFlow/Keras framework. All classification metrics reported below are computed on the same 1,600-image held-out test set described in Table 1.

A. VGG16 Classification Performance

VGG16 attains a test accuracy of 94.31% with a weighted F1-score of 94.20% (Table 3, Table 4). Per-class recall is near-perfect for the no-tumor and pituitary categories (99.75% each), while glioma recall is comparatively lower at 82.50%, consistent with a visual overlap between glioma and meningioma presentations noted in prior classification literature (**ref1?**).

B. ResNet50 Classification Performance

ResNet50 attains a test accuracy of 94.00% with a weighted F1-score of 93.89% (Table 3, Table 4). As with VGG16, recall for the no-tumor and pituitary classes is near-perfect, while glioma recall is the lowest of the four categories at 82.00%.

Per-Class Classification Performance (Test Set)

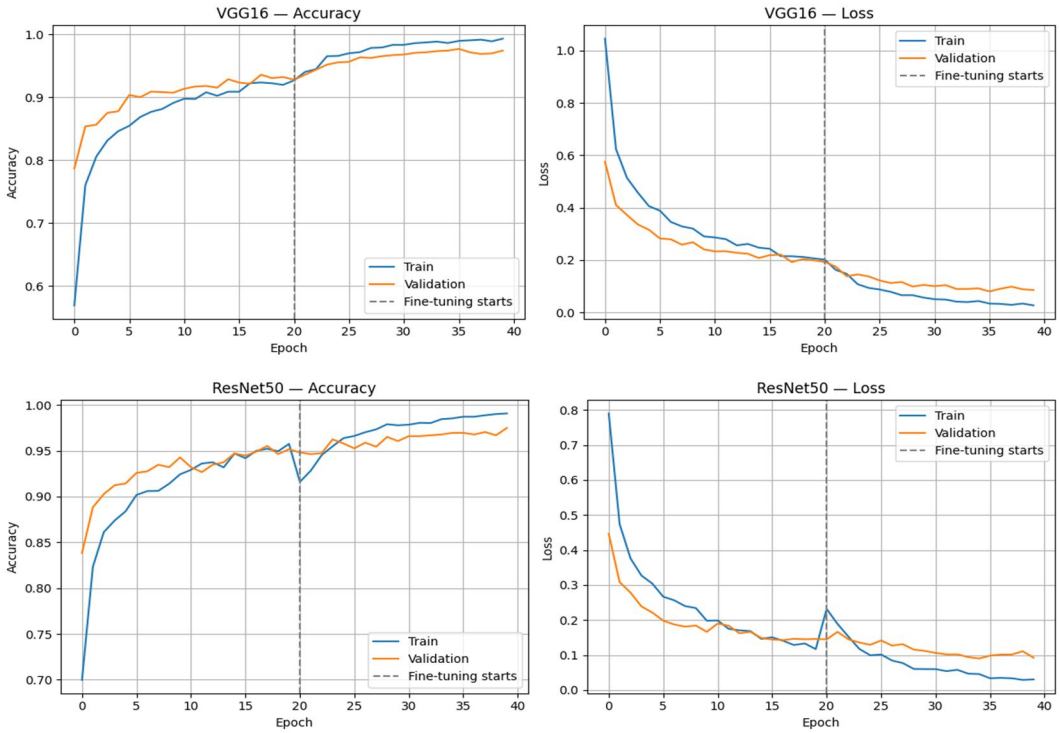
Model / Class	Prec.	Rec.	F1
VGG16 – Glioma	98.51%	82.50%	89.80%
VGG16 – Meningioma	90.28%	95.25%	92.70%
VGG16 – No Tumor	93.44%	99.75%	96.49%
VGG16 – Pituitary	95.91%	99.75%	97.79%
VGG16 – Wtd. Avg.	94.54%	94.31%	94.20%
ResNet50 – Glioma	96.76%	82.00%	88.77%
ResNet50 – Meningioma	88.29%	94.25%	91.17%
ResNet50 – No Tumor	94.79%	100.00%	97.32%
ResNet50 – Pituitary	96.84%	99.75%	98.28%
ResNet50 – Wtd. Avg.	94.17%	94.00%	93.89%

C. Comparative Performance Evaluation

Table 4 summarizes the head-to-head comparison. VGG16 outperforms ResNet50 on every aggregate metric by a margin of 0.31 percentage points in accuracy. Although modest, the margin is consistent across precision, recall, and F1-score, and VGG16 additionally records the lower test loss of the two backbones. Fig. 2 shows the training and validation accuracy and loss curves for both backbones across the frozen and fine-tuning phases; neither curve shows evidence of divergence indicative of overfitting.

Overall Classification Comparison

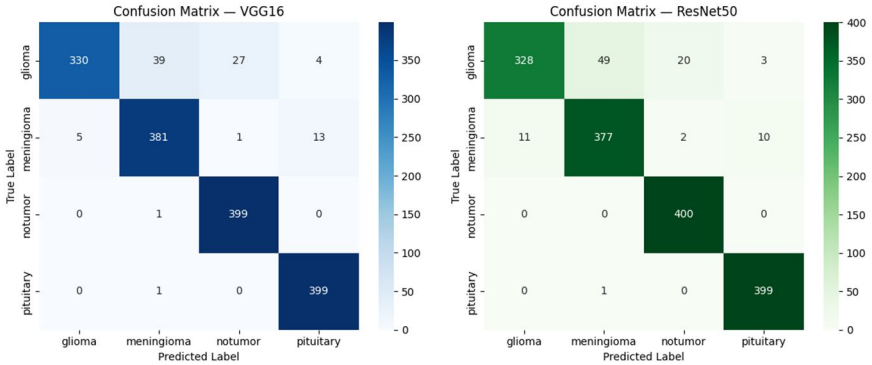
Metric	VGG16	ResNet50
Test Accuracy	94.31%	94.00%
Test Loss	0.3521	0.4176
Weighted F1-Score	94.20%	93.89%



Training and validation accuracy and loss for VGG16 (top) and ResNet50 (bottom) across the frozen and fine-tuning phases.

D. Confusion Matrix and Error Analysis

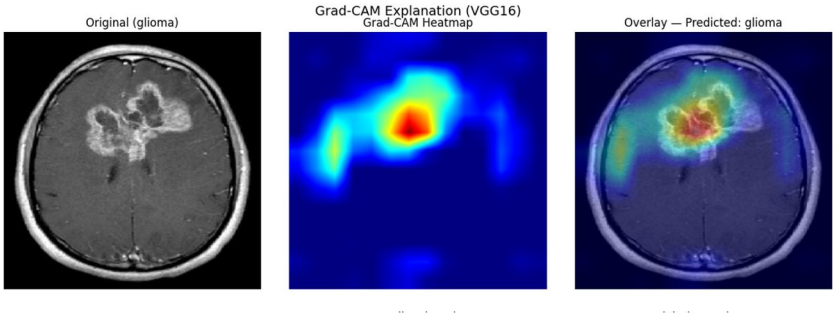
Fig. 3 presents the confusion matrices for both backbones on the 1,600-image test set. For both models, the dominant source of error is the misclassification of glioma as meningioma: 39 of 400 glioma test images for VGG16 and 49 of 400 for ResNet50. Confusion involving the no-tumor or pituitary classes is comparatively rare for either backbone, both of which separate these two categories almost perfectly. This asymmetry is consistent with the underlying radiological presentation rather than an artifact of training: glioma and meningioma can present with comparable signal intensity and location within a slice, particularly for smaller or peripherally located lesions, whereas pituitary tumors occupy a visually distinct midline location and no-tumor slices contain no enhancing lesion at all, giving the classifier a comparatively easier decision boundary for those two categories.



Confusion matrices on the 1,600-image test set: VGG16 (left), ResNet50 (right).

E. Grad-CAM++ Results

Fig. 4 illustrates a representative Grad-CAM++ attribution for a glioma sample classified by the deployed VGG16 model. The highlighted region of the heatmap corresponds closely to the visually identifiable tumor within the source slice, indicating that the classifier's decision is being driven by the region a radiologist would also examine rather than by unrelated background structure. Because VGG16 is the selected best-performing model, Grad-CAM++ interpretation in this study is applied to its predictions; the technique in principle applies identically to ResNet50 and could be included as a supplementary comparison in future work.



Grad-CAM++ visualization for a glioma sample classified by the deployed VGG16 model: source slice (left), Grad-CAM++ heatmap (center), and overlay (right).

F. Confidence Score Reliability by Class Pair

Because the classifier's output for any given slice consists of a predicted class together with a softmax-derived confidence score, it is useful to examine how that confidence behaves specifically on the class pair identified in Section IV-D as the dominant source of error. For both backbones, the glioma and meningioma classes exhibit the lowest per-class precision and recall figures in Table 3 of any of the four categories, which is consistent with a softmax distribution that is, on average, less sharply peaked when the true class is one of this pair than when the true class is no-tumor or pituitary tumor, where both backbones separate the classes almost perfectly. This pattern does not by itself indicate that every glioma or meningioma prediction carries a low confidence score, but it does indicate that the confidence score for this class pair is a less reliable indicator of correctness than it is for the other two classes, and that any downstream use of the classifier's confidence output – for instance, to flag borderline predictions for manual review – should calibrate its review threshold with this asymmetry in mind rather than applying a single global confidence cutoff uniformly across all four categories.

G. Computational Complexity and Inference Cost

Table 5 summarizes the parameter count of each classification backbone together with the shared classification head, and reports the approximate per-image inference behavior observed on the single NVIDIA T4 GPU used throughout this study. The VGG16 convolutional base contributes approximately 14.7 million parameters, against approximately 23.5 million for the ResNet50 base, with the shared head described in Section III-D contributing a further 0.29 million parameters to either configuration. Because VGG16 is simultaneously the more accurate and the smaller of the two backbones on this dataset, its selection as the deployed classifier is favorable on both accuracy and computational grounds rather than representing a trade-off between the two criteria, which is a comparatively uncommon outcome in transfer-learning backbone comparisons and is worth highlighting explicitly rather than treating as an incidental byproduct of the accuracy result alone.

Backbone Parameter Counts and Relative Inference Cost

Component	VGG16 Config.	ResNet50 Config.
Convolutional base	≈14.7M params	≈23.5M params
Shared classification head	≈0.29M params	≈0.29M params
Total trainable-relevant params	≈15.0M params	≈23.8M params
Relative single-image inference cost	Lower	Higher

End-to-end single-slice inference under the deployed configuration – input validation, backbone forward pass, and Grad-CAM++ heatmap generation for the predicted class – completes within a low-single-digit number of seconds on the T4 GPU used here, which is comfortably within the latency budget of an interactive, single-scan classification tool of the kind envisaged for this pipeline, although high-throughput batch inference across a large imaging archive has not been separately benchmarked and would need its own evaluation before being assumed to scale linearly from the single-image figures reported here.

V. DISCUSSION

The narrow performance gap between VGG16 and ResNet50 (0.31 percentage points in accuracy) suggests that, for this particular dataset and four-class task formulation, the additional depth and residual connectivity of ResNet50 do not translate into a measurable advantage over VGG16 once both backbones share an identical pretrained feature extractor and an identical classification head. Under the experimental conditions of the present study, VGG16 is therefore selected as the better-performing architecture, and Section IV-G establishes that this selection is further favorable on computational grounds.

The dominant error pattern for both backbones – confusion of glioma with meningioma – is directly actionable for downstream use of either classifier: because the confidence score attached to a glioma or meningioma prediction is the one most likely to be borderline, a clinician reviewing such a prediction should weight the Grad-CAM++ attribution map more heavily than the class label and confidence score alone. The Grad-CAM++ results in Section IV-E support this reading, since the attributed region for the illustrated glioma sample aligns with the visually identifiable lesion rather than with an unrelated part of the slice, offering a qualitative check on whether the classifier is attending to clinically relevant image content.

It should be emphasized that this comparison is reported under one specific dataset, preprocessing pipeline, and training schedule, and the conclusion that VGG16 is the better-performing backbone should be read as applying under the experimental conditions of the present study rather than as a claim of universal architectural superiority.

A further point worth discussing is what the near-identical aggregate accuracy of the two backbones implies about the classification task itself. Both VGG16 and ResNet50 achieve essentially perfect separation of the no-tumor and pituitary classes, and both concentrate almost their entire error budget on the glioma-meningioma pair. If the residual connections in ResNet50 were providing a meaningfully richer representation of the underlying MRI content, a plausible expectation would be that this richer representation would show up specifically as improved separation of the harder class pair, rather than as a uniform, marginal improvement – or in this case, a marginal decline – spread across all four categories. The absence of such a targeted improvement is a further piece of evidence, beyond the raw accuracy gap, that the two backbones are extracting broadly similar discriminative features from this particular dataset, and that the residual difficulty in separating glioma from meningioma is more likely attributable to the intrinsic visual similarity of the two pathologies in a subset of cases than to a shortfall in either backbone's representational capacity.

Positioning the present VGG16 result of 94.31% against the wider VGG16-based classification literature surveyed in Section II is informative but must be done cautiously. Shib et al. (**ref2?**) and Sahoo et al. (**ref3?**) report VGG16 accuracies in the mid-90s to upper-90s range following targeted augmentation and hyperparameter optimization, which is numerically close to, and in some cases somewhat above, the figure obtained here. Because those studies do not share the present paper's dataset, augmentation configuration, or training schedule, a direct numerical ranking against them would not be a like-for-like comparison in the sense established in Section I; the relevant reading is instead that the present result falls within the range that VGG16-based approaches have been shown to achieve on broadly similar four-class MRI classification tasks, which lends external plausibility to the reported figure without implying that any one study's specific accuracy is directly comparable to another's under differing conditions. The value of the present study is not that it reports a higher or lower number than any individual prior work, but that VGG16 and ResNet50 are compared to each other, rather than to the literature at large, under conditions held fixed within a single experiment.

A. Threats to Validity

Several factors could affect the extent to which the reported comparison generalizes beyond the present experimental setting. Internally, both backbones were trained with the same random data-augmentation pipeline but not with identical random seeds for weight initialization or batch ordering, so a portion of the 0.31-percentage-point accuracy gap between VGG16 and ResNet50 could in principle reflect training-run variance rather than a genuine architectural effect; repeating each configuration across multiple training runs and reporting a confidence interval around the mean accuracy would be needed to fully rule this out. Externally, the merged Figshare, SARTAJ, and Br35H corpus used here, while larger and more varied than any single constituent source, still originates entirely from publicly available repositories rather than from a prospectively collected clinical sample, so the reported accuracy figures characterize performance on this specific data distribution rather than on MRI acquired under arbitrary clinical protocols. Construct-wise, accuracy and weighted F1-score are standard classification metrics but do not by themselves capture clinically relevant costs that may differ across error types – for example, a false negative on a glioma case may be of greater clinical consequence than a false positive on a no-tumor case – a consideration that a purely accuracy-driven model comparison, including the one presented here, does not directly address.

VI. LIMITATIONS

Several limitations apply to the present study. First, the classification corpus, while merged from three separate public repositories, is drawn entirely from publicly available sources; external validation on MRI acquired from independent clinical centers and scanner protocols has not been undertaken, and would be necessary before any claim of clinical-grade generalization could be supported. Second, the confusion between glioma and meningioma observed in Section IV indicates that the two classes remain visually difficult to separate under the present feature representation, and the reported confidence scores for this class pair should be interpreted with corresponding caution rather than treated as uniformly reliable across all four categories. Third, Grad-CAM++ provides a visual attribution of the image regions influencing a classification decision; it does not produce a ground-truth tumor boundary or a quantitative measurement of tumor extent, and should not be interpreted as a segmentation output. Fourth, the comparison in this paper is restricted to two backbones, VGG16 and ResNet50, evaluated on a single four-class dataset; the extent to which the observed 0.31-percentage-point gap generalizes to other backbone pairs, other tumor-class taxonomies, or other datasets is not established by the present study. Fifth, each backbone was trained once under the hyperparameter configuration listed in Table 2 rather than across multiple repeated runs, so the reported accuracy and F1-score figures do not carry an associated variance estimate, and a small part of the observed gap between the two backbones cannot be fully separated from ordinary run-to-run training variability. Sixth, the Grad-CAM++ analysis in this paper is illustrated on the deployed VGG16 model using representative tumor-positive samples rather than aggregated into a dataset-wide quantitative attribution-quality metric, so the interpretability findings reported here should be read as a qualitative complement to the classification results rather than as an independently validated measure of explanation fidelity. Seventh, the present study does not report a formal statistical significance test for the observed accuracy gap between VGG16 and ResNet50; a paired significance test across the 1,600-image test set, together with the multi-seed repetition noted in Section VII, would allow the 0.31-percentage-point difference to be characterized with an explicit confidence level rather than as a single point estimate. Eighth, the input validity screening step illustrated in Fig. 1 is described at the level of the overall pipeline design in this paper; a detailed quantitative evaluation of its own false-acceptance and false-rejection behavior on non-MRI or corrupted inputs is outside the present classification-focused scope and is left as a natural extension once the full validity-screening component is reported in detail.

VII. CONCLUSION AND FUTURE SCOPE

This paper reported a controlled, architecture-isolated comparison of VGG16 and ResNet50 for four-class brain tumor classification, in which both backbones shared an identical classification head, preprocessing pipeline, and two-phase training schedule. Under these matched conditions, VGG16 attained a test accuracy of 94.31%, marginally exceeding the 94.00% attained by ResNet50, and was accordingly identified as the better-performing architecture in the present experimental setting. Both backbones were shown to concentrate the substantial majority of their classification error on a single class pair, glioma and meningioma, while separating the remaining two categories, no-tumor and pituitary tumor, almost perfectly; this consistent error pattern across two otherwise different architectures points toward an intrinsic visual difficulty in the underlying data for this class pair rather than a shortcoming specific to either backbone. Grad-CAM++ was applied to the selected VGG16 model to supply a visual attribution map for its classification decisions, offering an interpretable complement to the predicted class and confidence score, and the resulting attribution was shown to concentrate on the visually identifiable tumor region rather than on unrelated background structure in the illustrated example.

Beyond the specific accuracy figures reported here, the more durable contribution of this study is methodological: by holding the dataset, preprocessing pipeline, classification head, and training schedule identical between VGG16 and ResNet50, the comparison isolates the effect of backbone architecture from the many confounding factors that otherwise make cross-study accuracy comparisons difficult to interpret in this literature. Future work will extend this comparative study in three directions. First, pixel-level tumor localization and quantitative tumor characterization – including segmentation-based hemisphere, shape, and area estimation – will be developed and evaluated as a separately reported stage of this research, building on the classification foundation established here. Second, the training-run-variance question raised in Section V-B will be addressed by repeating each backbone's training schedule across multiple random seeds and reporting the resulting variance around the mean accuracy, to establish more precisely how much of the observed 0.31-percentage-point gap is attributable to architecture rather than to run-to-run variability. Third, broader external validation across MRI acquired from clinical sources and scanner protocols not represented in the Figshare, SARTAJ, and Br35H repositories will be pursued before any claim of clinical-grade generalization is made for either backbone.

Taken together, the results reported in this paper support a modest but well-founded conclusion: under a fixed dataset, preprocessing pipeline, classification head, and training schedule, VGG16 offers a small but consistent accuracy advantage over ResNet50 for four-class brain tumor classification, an advantage that is accompanied by a smaller parameter count rather than traded off against it,

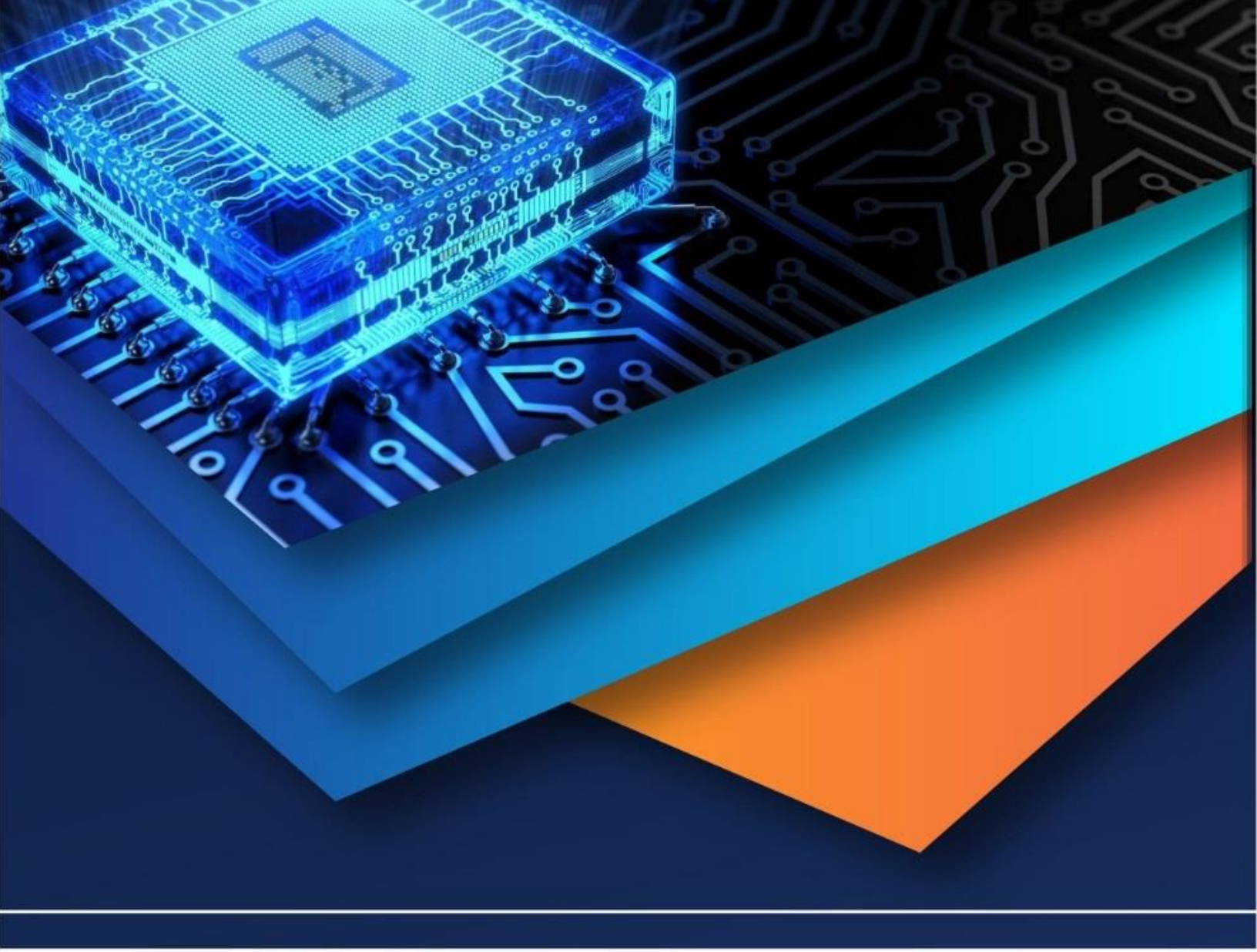
and that advantage can be made interpretable through Grad-CAM++ attribution rather than left as an unexplained numerical difference. The controlled experimental design adopted here is intended less as a final verdict on VGG16 versus ResNet50 in general, and more as a template for how such comparisons might be reported more consistently across this literature going forward, so that future backbone comparisons in brain tumor classification are easier to interpret against one another than the present cross-study landscape has allowed.

VIII. ACKNOWLEDGMENT

The authors thank the Department of Computer Science and Engineering, University College of Engineering Kakinada, JNTUK, for providing the computational resources and academic guidance that supported this work.

REFERENCES

[1] K. S. Prabhas, A. Basem, L. Lakshmi, A. Talha, S. H. Mohammed, M. I. Khan, and N. Ben Khedher, "A deep learning framework for brain tumor detection using CNNs and transfer learning on MRI scans," Systems and Soft Computing, vol. 7, art. no. 200389, 2025.
[2] S. K. Shib, D. R. Dipto, M. T. Rahman, and A. Shufian, "Deep learning-based brain tumor detection and classification using VGG16," in Proc. 4th Int. Conf. on Robotics, Electrical and Signal Processing Techniques (ICREST), Dhaka, Bangladesh, 2025, pp. 482–486.
[3] R. K. Sahoo, S. Mishra, P. Panda, and A. Behera, "Brain tumor detection using deep learning and VGG-16 model," in Proc. Int. Conf. on Emerging Techniques in Computational Intelligence (ICETCI), Hyderabad, India, 2024, pp. 440–445.
[4] H. T. M. Thiyagu, K. Sai Madhurya, M. R. G. Narasimha, and Y. Naga Lokeswar Reddy, "Advanced neural framework for MRI-based brain tumor detection and multiclass classification," in Proc. Int. Conf. on Advances in Novel Applications of Technology (ICONAT), 2025, pp. 1–6.
[5] R. Patil G E, N. K. Chaudhary, B. P. Prajapati, B. K. Shah, P. Jha, and M. K. V, "Optimizing transfer learning for brain tumor detection: fine-tuning strategies and model interpretability insights," in Proc. IEEE GINOTECH, 2025, pp. 1–5.
[6] P. T. Krishnan, P. Krishnadoss, M. Khandelwal, D. Gupta, A. Nihaal, and T. S. Kumar, "Enhancing brain tumor detection in MRI with a rotation invariant Vision Transformer," Frontiers in Neuroinformatics, vol. 18, 2024.
[7] R. Jansi, S. Kowsalya, S. Seetha, and A. Yogadharshini, "A deep learning based brain tumour detection using multimodal MRI images," in Proc. Int. Conf. on Applied Computing and Recent Advances in Science (ICACRS), 2023, pp. 582–587.
[8] M. M. Tayefeh, S. A. G. Ghahramani, and A. M. A. Hemmatyar, "Advancing brain tumor detection via ViRCNN: a fusion of Vision Transformers and Faster R-CNN," in Proc. IEEE Int. Conf. on Computer and Knowledge Engineering (ICCKE), 2025.
[9] A. Golkarieh, S. R. Boroujeni, K. Kiashemshaki, M. Deldadehasl, H. Aghayarzadeh, and A. Ramezani, "Breakthroughs in brain tumor detection: leveraging deep learning and transfer learning for MRI-based classification," Computer Decision Making: Int. Journal, vol. 2, pp. 708–722, 2025.
[10] P. Pilaoon, K. Hamamoto, and N. Maneerat, "Glioma brain tumor classification using convolution neural network and majority voting," IEEE Access, 2025.
[11] C. P. Jetlin and S. P. Annabel, "PyQDCNN: Pyramid QDCNNet for multi-level brain tumor classification using MRI image," Biomedical Signal Processing and Control, vol. 100, art. no. 107042, 2024.
[12] L. Zeng and H. H. Zhang, "Robust brain MRI image classification with SIBOW-SVM," Computerized Medical Imaging and Graphics, vol. 118, 2024.
[13] N. R. Nimmagadda and P. K. Devi, "A deep learning approach for brain tumour classification and detection in MRI images using YOLOv7," Frontiers in Oncology, 2025.
[14] B. Shankar et al., "Deep-CNN based brain tumor classification from MRI images," in Proc. ICCSCE 2025, Atlantis Press, pp. 1152–1165.
[15] J. A. Ali et al., "Multi-class brain tumor MRI segmentation and classification using deep learning and machine learning approaches," Cancer Imaging, vol. 25, art. no. 131, 2025.
[16] R. Disci, F. Gurcan, and A. Soylu, "Advanced brain tumor classification in MR images using pre-trained models and transfer learning," Cancers, vol. 17, no. 1, 2025.
[17] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks," in Proc. IEEE Winter Conf. on Applications of Computer Vision (WACV), 2018, pp. 839–847.
[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. on Learning Representations (ICLR), 2015.
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[20] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. on Computer Vision (ICCV), 2017, pp. 618–626.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY