Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 2327~2335
e-ISSN: 2548-6861

Comparative Analysis of CNN, ResNet50, and Vision Transformer Architectures for Brain Tumor Classification from MRI Images

Matthieu Kayembe [1]*, Franklin Mwamba [2], Fiston Oshasha [3]***, Pierre Kafunda [4]*, John Poma [5]***
* Department of Mathematics, Statistics and Computer Science, University of Kinshasa, Kinshasa, DR. Congo
**Health Sciences Research Institute, Kinshasa, Democratic Republic of the Congo
***General Commissariat for Atomic Energy, Regional Center for Nuclear Studies of Kinshasa, P.O. Box 868, University of Kinshasa
mknministries@gmail.com [1], franklin.mwamba@irss.cd [2], fiston.oshasha.fv@cgea-rdc.org [3], pierre.kafunda@unikin.ac.cd [4], pomaesendo7@gmail.com [5]

Article Info

Article history:

Received 2026-03-27
Revised 2026-05-04
Accepted 2026-05-25

Keyword:

Medical Image Classification,
Brain MRI,
CNN,
ResNet50,
Vision Transformer,
Deep Learning,
Brain tumors.

ABSTRACT

Brain tumor classification from Magnetic Resonance Imaging (MRI) is a critical task in computer-aided medical diagnosis. Deep learning has emerged as a powerful approach for automating this process, yet a rigorous and unified comparison of architectures evaluated on the same dataset under identical conditions remains limited in the literature. This study presents a systematic comparative analysis of three deep learning architectures: a Convolutional Neural Network (CNN) trained from scratch, a transfer learning-based model using ResNet50 with two phase fine tuning, and a pre trained Vision Transformer (ViT), evaluated on a publicly available multi class MRI dataset comprising 3,264 images distributed across four categories: glioma tumor, meningioma tumor, pituitary tumor, and no tumor. The training set contains 2,870 images and the test set contains 394 images. All models are trained and evaluated under identical experimental conditions. Experimental results show that the CNN achieves a test accuracy of 0.24 and a weighted F1-score of 0.20, constrained by its local feature extraction and shallow architecture. ResNet50 reaches a high training accuracy of 0.97 but suffers from severe overfitting (test accuracy: 0.28), largely due to the domain gap between ImageNet pre-training and MRI data. In contrast, the Vision Transformer achieves the best overall performance, with a test accuracy of 0.76, a weighted F1-score of 0.73, and consistent generalization attributed to its multi-head self-attention mechanism capturing global spatial dependencies across the full image. These findings confirm the superiority of attention-based architectures for complex medical image classification tasks. Limitations related to dataset size, model interpretability, and the gap to clinical validation thresholds are critically discussed. Future work will explore hybrid CNN-Transformer architectures, Grad-CAM-based visualization, and extended clinical validation to enhance real-world applicability.

I. INTRODUCTION

Brain tumors represent a serious and potentially life-threatening medical condition requiring accurate and timely diagnosis. Magnetic Resonance Imaging (MRI) is the gold standard imaging modality for detecting and characterizing brain abnormalities, owing to its ability to provide high-contrast, detailed soft-tissue images without ionizing radiation. However, manual interpretation of MRI scans by radiologists is time-consuming, subject to inter-observer variability, and increasingly challenged by growing patient volumes in clinical settings. Automating brain tumor classification is therefore a high-priority research objective with direct clinical implications.

Deep learning has enabled the development of automated systems for medical image analysis with high accuracy and reproducibility. Convolutional Neural Networks (CNNs), notably introduced by Krizhevsky et al. [1] with the AlexNet architecture, have demonstrated remarkable performance in

image classification tasks and have since been applied extensively to brain tumor classification [5], [6]. CNNs learn hierarchical, spatially-local features through stacked convolutional layers and have established themselves as the dominant paradigm in medical image analysis for the past decade.

To overcome CNN limitations in low-data regimes, transfer learning strategies leveraging pre-trained models such as ResNet50 [2] have been widely adopted. These approaches allow models to benefit from feature representations learned on large-scale datasets such as ImageNet, substantially reducing the amount of domain-specific training data required [7], [8]. However, the effectiveness of transfer learning may be constrained by domain shift, which refers to the fundamental difference between the statistics of natural RGB images used in pre training and the grayscale or pseudo colored texture patterns of MRI scans.

More recently, Transformer-based architectures, originally proposed for natural language processing by Vaswani et al. [3], have been successfully adapted to computer vision. The Vision Transformer (ViT) introduced by Dosovitskiy et al. [4] reformulates image classification as a sequence modeling problem by partitioning images into fixed-size patches and applying multi-head self-attention mechanisms. Its capacity to model long-range spatial dependencies across the entire image makes it particularly attractive for complex medical images such as brain MRIs, where global structural relationships between regions carry critical diagnostic information [11], [12].

Despite the growing literature on individual architectures, a rigorous and unified experimental comparison of CNN, ResNet50, and ViT under strictly identical conditions, using the same dataset, preprocessing pipeline, batch size, and evaluation protocol, remains scarce in the brain tumor classification domain. Most existing comparative studies either focus on a subset of architectures, use heterogeneous preprocessing strategies, or lack comprehensive class level performance analysis. This work fills this gap and makes the following four original contributions:

- Unified experimental comparison: Three representative deep learning architectures (custom CNN, ResNet50 with two-phase fine-tuning, and pre-trained ViT) are evaluated on the same standardized MRI dataset under identical experimental conditions.
- Comprehensive multi-metric evaluation: Performance is assessed using accuracy, per-class precision, recall, F1-score, and normalized confusion matrices, providing fine-grained analysis for each of the four tumor categories.
- Generalization and overfitting analysis: The train-test performance gap is quantified and analyzed for each architecture, with learning curve evidence supporting the reported findings.
- Critical clinical discussion: Limitations related to model interpretability, dataset size, generalization to multi-institutional data, and the gap between reported metrics and clinical deployment thresholds are explicitly addressed, providing actionable research directions.

II. RELATED WORK AND POSITIONING

Brain tumor classification from MRI images has witnessed significant advances driven by the evolution of deep learning architectures. Existing approaches can be organized into three main categories, each with distinct strengths and limitations that motivate the present comparative study.

A. CNN-Based approaches

CNNs have been the dominant paradigm in medical image classification for the past decade due to their ability to automatically extract hierarchical, spatially-local features directly from raw pixel data. The AlexNet architecture introduced by Krizhevsky et al. [1] established the viability of deep convolutional models for large-scale image recognition, demonstrating that learned feature representations substantially outperform hand-crafted features such as SIFT or HOG descriptors.

In the medical imaging domain, several works have applied CNNs specifically to brain tumor classification from MRI. Cheng et al. [5] proposed a deep feature-based approach using tumor region augmentation for multi-class classification, achieving strong results on a custom dataset. Similarly, Badža and Barjaktarović [6] developed a lightweight custom CNN achieving high accuracy on MRI scans, demonstrating that compact architectures can be effective when carefully tuned to the target domain. These works confirm the relevance of CNNs as baseline models while also highlighting their sensitivity to training set size and the risk of limited generalization when data is scarce, a fundamental limitation that motivates the use of transfer learning and attention based strategies.

B. Transfer Learning and Pre-Trained Models

To overcome data scarcity, transfer learning has become a standard strategy in medical image analysis. He et al. [2] introduced deep residual networks (ResNet), enabling the training of very deep architectures by addressing the vanishing gradient problem through identity skip connections of the form $H(x) = F(x) + x$. Their ResNet50 variant, pre-trained on ImageNet, provides rich visual feature representations that can subsequently be adapted to new target domains through fine-tuning.

In the brain tumor classification domain, Khan et al. [7] applied ResNet50 fine-tuned on a dedicated MRI dataset and reported competitive results, while Deepak and Ameer [8] demonstrated that transfer learning consistently outperforms CNN models trained from scratch across several evaluation metrics. Despite these advantages, both studies acknowledge that fine-tuning on small medical datasets can lead to severe overfitting, particularly when the source and target domains

differ substantially. The domain gap between natural RGB images (ImageNet) and grayscale or pseudo-colored MRI scans represents a fundamental challenge, as the low-level texture features learned during ImageNet pre-training do not necessarily transfer to the distinct visual statistics of MRI data. This domain gap limitation constitutes one of the central phenomena analyzed in the present study.

C. Transformer-based approaches

The introduction of the Transformer model by Vaswani et al. [3] revolutionized sequence modeling through the self-attention mechanism, which computes pairwise dependencies between all elements of an input sequence simultaneously. Its adaptation to vision tasks, the Vision Transformer (ViT) proposed by Alexey Dosovitskiy et al. [4], reformulates image classification as a sequence modeling problem by partitioning images into fixed size patches (P × P pixels), projecting each patch into an embedding vector, and processing the resulting sequence through standard Transformer encoder blocks equipped with multi head self-attention.

Touvron et al. [9] further improved ViT through data-efficient training strategies and knowledge distillation, extending its applicability to moderately-sized datasets without requiring the massive pre-training datasets originally used by Dosovitskiy et al. In medical imaging specifically, surveys by Azad et al. [12] and Takahashi et al. [13] confirm the growing adoption of Vision Transformers, highlighting their advantages in modeling long-range spatial dependencies that are critical for understanding global tumor morphology and tissue relationships in MRI scans. Hybrid architectures combining CNN feature extractors with Transformer encoders, such as TransUNet [14], have demonstrated state-of-the-art performance in medical segmentation tasks. Similarly, Oshasha et al. [11] demonstrated the effectiveness of cross-scale Transformer-based architectures for robust visual classification under challenging conditions.

D. Positioning of this work

Despite the advances described above, several important gaps persist in the existing literature. First, most comparative studies do not evaluate all three paradigms, namely custom CNN, pre trained CNN, and ViT, under strictly identical experimental conditions using the same dataset, preprocessing, batch size, and evaluation protocol. Second, class-level performance analysis across all tumor categories is rarely reported comprehensively, making it difficult to identify which specific tumor types remain challenging. Third, the generalization behavior of these architectures on limited medical datasets is insufficiently characterized, with few works explicitly quantifying train-test performance gaps. Fourth, the clinical implications of reported accuracy levels are seldom discussed critically, with most works omitting discussion of interpretability and regulatory requirements.

This work directly addresses all four gaps by conducting a fully unified experimental comparison with consistent preprocessing, training protocols, and multi-level evaluation, accompanied by a rigorous discussion of clinical limitations and future research directions.

III. METHODOLOGY

A. Dataset Description

The dataset used in this study is a publicly available MRI brain tumor classification benchmark obtained from Kaggle (Brain Tumor Classification MRI). It comprises a total of 3,264 MRI images stored in JPEG format with variable original resolutions, distributed across four categories: glioma tumor, meningioma tumor, pituitary tumor, and no tumor. Table I details the class distribution across the training and test splits.

TABLE I.
DATASET DISTRIBUTION BY CLASS AND SPLIT

Class	Training Set	Test Set	Total
Glioma tumor	826	100	926
Meningioma tumor	822	115	937
No tumor	395	105	500
Pituitary tumor	827	74	901
Total	**2,870**	**394**	**3,264**

The dataset is partitioned into a training set (2,870 images) used for model parameter optimization, and a test set (394 images) used for epoch-level validation monitoring and final evaluation. No separate validation set was predefined in the original data; a *holdout validation strategy* was therefore adopted throughout. Images are provided in JPEG format, resized to 224 × 224 pixels during preprocessing. It is acknowledged that using the test set for epoch monitoring introduces a degree of optimistic bias; a k-fold cross-validation strategy would provide more reliable estimates but was not feasible given the dataset size constraints. This constitutes a recognized limitation discussed in Section V.

It is further acknowledged that this dataset, while standard in the literature, represents a limited size for training attention-based architectures such as ViT, which typically benefit from large-scale pre-training data. The impact of this constraint on model behavior is analyzed in the Discussion section.

B. Data Preprocessing

Preprocessing was applied consistently across all models to ensure fair comparison. The following steps were applied:

- Resizing: All images resized to 224 × 224 pixels, compatible with ResNet50 and ViT (patch size P = 16 yields $N = (224/16)^2 = 196$ patches).
- Normalization: Pixel values rescaled to [0, 1] by dividing by 255 for CNN and ViT. For ResNet50, Keras's built-in preprocess_input function applies channel-wise mean subtraction and scaling per ImageNet statistics.

- Label encoding: One-hot encoded vectors for CNN and ResNet50 (categorical_crossentropy). Integer class indices for ViT (PyTorch CrossEntropyLoss).
- Data augmentation (training set only): CNN and ResNet50: random rotations (±15°), zoom (10%), horizontal flips via Keras ImageDataGenerator. ViT: random horizontal flips, rotations (±10°), width/height shifts (10%), zoom (15%) via PyTorch transforms. No augmentation applied to the test set.

C. Studied Architectures

Three architectures were compared in this study:

1). CNN Trained from Scratch

The custom CNN follows a standard encoder-classifier design with three convolutional blocks followed by fully connected layers. Each convolutional block applies a Conv2D layer with ReLU activation $f(x) = max(0, x)$ followed by MaxPooling2D (2×2) for spatial dimensionality reduction. The complete architecture is detailed in Table II.

TABLE II.
CNN ARCHITECTURE — LAYER-BY-LAYER DETAIL

Layer	Output Shape	Params.	Activation
Conv2D (32 filters, 3×3)	222 × 222 × 32	896	ReLU
MaxPooling2D (2×2)	111 × 111 × 32	0	—
Conv2D (64 filters, 3×3)	109 × 109 × 64	18,496	ReLU
MaxPooling2D (2×2)	54 × 54 × 64	0	—
Conv2D (128 filters, 3×3)	52 × 52 × 128	73,856	ReLU
MaxPooling2D (2×2)	26 × 26 × 128	0	—
Flatten	86,528	0	—
Dense (128 units)	128	11,075,712	ReLU
Dropout (rate = 0.5)	128	0	—
Dense (4 units)	4	516	Softmax
Total	—	**11,169,476**	—

With 11,169,476 total parameters, the majority of which (11,075,712) are concentrated in the first dense layer, this model serves as the baseline to evaluate performance achievable without transfer learning or global attention mechanisms.

2). ResNet50 (Transfer learning)

ResNet50 [2] is a 50-layer deep residual network pre-trained on ImageNet (1.2M images, 1,000 classes). Its core innovation is the residual skip connection $H(x) = F(x) + x$, which enables training of very deep networks by mitigating the vanishing gradient problem. Adaptation to the brain tumor classification task was performed in two sequential phases:

- Phase 1 — Feature extraction (20 epochs): All ResNet50 base layers frozen (trainable = False). The original classification head was replaced by: GlobalAveragePooling2D → BatchNormalization → Dense(128, ReLU) → Dropout(0.5) → Dense(4, Softmax). Compiled with Adam at default learning rate.
- Phase 2 — Fine-tuning (10 epochs): The last 30 layers of the ResNet50 base unfrozen (trainable = True). Re-compiled with Adam(lr = 1×10^{-5}) to preserve pre-trained representations while adapting to the MRI domain.

3). Vision Transformer (ViT)

The ViT model used is *google/vit-base-patch16-224*, loaded via HuggingFace Transformers [4]. Its configuration is: patch size P = 16×16, number of patches N = 196, embedding dimension d = 768, number of Transformer blocks L = 12, number of attention heads H = 12, MLP hidden dimension = 3,072, total parameters ≈ 86M. The processing pipeline is:

- Patch extraction: The input image (224×224×3) is divided into N = 196 non-overlapping patches of 16×16 pixels each.
- Linear projection: Each patch $x_i \in \mathbb{R}^{P^2 \cdot C}$ is projected to an embedding vector $e_i \in \mathbb{R}^d$ via a learned linear transformation.
- Positional encoding: A learnable positional embedding E_{pos} is added to each patch embedding to preserve spatial information.
- Transformer encoding: The sequence passes through $L = 12$ Transformer blocks, each with multi-head self-attention (MHSA) and feed-forward MLP with LayerNorm:

$$Attention(Q, K, V) = softmax\left(QK^T/\sqrt{d_k}\right).V$$

Classification: The [CLS] token representation is passed to a linear 4-class head with softmax. The classification head was replaced to match the 4-class output, and the full model was fine-tuned end-to-end with $Adam(lr = 1 \times 10^{-4})$.

D. Training Parameters

Table III summarizes the complete training configuration for each model.

TABLE III.
TRAINING HYPERPARAMETERS PER MODEL

Parameter	CNN	ResNet50	ViT
Framework	Keras / TF	Keras / TF	PyTorch + HuggingFace
Optimizer	Adam (default lr)	Adam (default) → Adam (1e−5)	Adam (lr = 1e−4)

Loss function	Categorical CE	Categorical CE	Cross-Entropy
Epochs	10	20 + 10 (fine-tune)	10
Batch size	32	32	32
Input size	224 × 224 × 3	224 × 224 × 3	224 × 224 × 3
Pre-training	None	ImageNet	ImageNet-21k

E. Evaluation Metrics

To compare the models, several metrics are used [8]:

• Accuracy:

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

• Precision:

$$Precision = \frac{TP}{TP + FP}$$

• Recall:

$$Recall = \frac{TP}{TP + FN}$$

• F1-score:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

Additionally, a normalized confusion matrix makes it possible to analyze classification errors by class.

F. Experimental Protocol

To ensure the reproducibility and fairness of the comparative study, all three models were developed and evaluated following a strictly unified experimental protocol. The key elements of this protocol are described below.

1). Development Environment and Hardware

All experiments were conducted in Google Colaboratory (Colab) with GPU acceleration (NVIDIA Tesla T4, 16 GB VRAM). The CNN and ResNet50 models were implemented using TensorFlow 2.x with the Keras high-level API. The ViT model was implemented using PyTorch and the HuggingFace Transformers library (transformers v4.x). Python 3.10 was used throughout. Key dependencies include: NumPy, scikit-learn (metrics computation and confusion matrices), Matplotlib and Seaborn (visualization), and torchvision (image transforms for the ViT pipeline).

2). Reproducibility Measures

To ensure reproducibility, random seeds were fixed at the start of each experiment (NumPy seed = 42, TensorFlow global seed = 42, PyTorch manual_seed = 42). The dataset directory structure (Training/ and Testing/ subfolders named after class labels) was kept identical across all three model pipelines. The same 394-image test set was used for final evaluation in all cases, and no test-set images were used during training or augmentation. The models were trained sequentially in independent notebook sessions to avoid any cross-contamination of GPU memory or state.

3). Training Pipeline and Data Flow

For CNN and ResNet50, images were loaded using Keras's ImageDataGenerator with flow_from_directory(), which applies augmentation on-the-fly during training. The test generator was configured with shuffle=False to guarantee consistent alignment between predicted labels and ground-truth class indices during evaluation. For the ViT, images were loaded using PyTorch's ImageFolder with a torchvision.transforms pipeline, and batched using DataLoader with shuffle=True for training and shuffle=False for evaluation. In all three cases, the training loop iterated over the full training set for the specified number of epochs, computing training loss and accuracy at each step. Epoch-level validation loss and accuracy were computed on the test set at the end of each epoch and recorded in the history object for subsequent learning curve visualization.

4). Evaluation Procedure

After training, each model was evaluated on the full 394-image test set. For CNN and ResNet50, predictions were obtained via model.predict(test_generator), and the predicted class for each image was determined by applying argmax over the softmax output vector: $\hat{y}$ = argmax(softmax(logits)). Ground-truth labels were retrieved from test_generator.classes (Keras). For the ViT, predictions were collected within a torch.no_grad() evaluation loop, with the predicted class identified by torch.max(logits, dim=1). In all cases, scikit-learn's classification_report() and confusion_matrix() functions were used to compute per-class precision, recall, F1-score, and weighted averages. Normalized confusion matrices (row-normalized by true class support) were generated using Seaborn's heatmap function to facilitate visual comparison across models.

5). Model Comparison Framework

Model comparison was performed along two complementary dimensions. First, absolute performance was compared using test-set accuracy and weighted F1-score as primary metrics, chosen because they account for class imbalance through support-weighted averaging. Second, generalization capacity was assessed through the train–test accuracy gap (Δ = Train Accuracy − Test Accuracy), where a large positive gap indicates overfitting. A gap below 0.05 was considered indicative of good generalization. This dual axis framework enables a nuanced comparison that distinguishes between models that perform well in absolute terms and models that generalize reliably to unseen data, a distinction of particular importance in the medical imaging context where overfitting carries direct clinical risk.

IV. EXPERIMENTAL RESULTS

A. CNN — Baseline Model

The CNN trained from scratch achieved a training accuracy of 0.80 after 10 epochs. On the test set, performance dropped substantially to **test accuracy = 0.24**, with a test loss of 3.17, indicating a severe train–test gap of 0.56 percentage points. The learning curve exhibited progressive divergence between training and validation accuracy from epoch 3 onward, confirming poor generalization. Table IV presents the class-level classification report.

TABLE IV.
CNN CLASSIFICATION REPORT (TEST SET, N = 394)

Class	Precision	Recall	F1-score	Support
Glioma tumor	0.16	0.03	0.05	100
Meningioma tumor	0.28	0.25	0.26	115
No tumor	0.24	0.49	0.32	105
Pituitary tumor	0.17	0.14	0.15	74
Weighted average	**0.22**	**0.24**	**0.20**	**394**

Figure 1 shows the learning curve, with a notable gap between training and validation, highlighting the model's poor generalization.

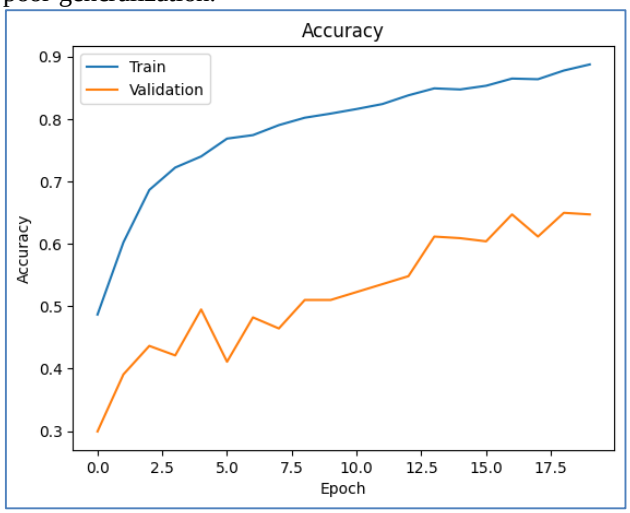

Figure 1: CNN learning curve

The confusion matrix reveals that the model predominantly confuses glioma tumor with other classes, achieving a recall of only 0.03 for this critical category. The no_tumor class is best identified (recall = 0.49), likely due to the absence of a tumor mass providing a visually distinctive signal. The globally poor performance confirms the CNN's inability to learn sufficient discriminative representations for complex intra-class variation from a dataset of only 2,870 training images without pre-training.

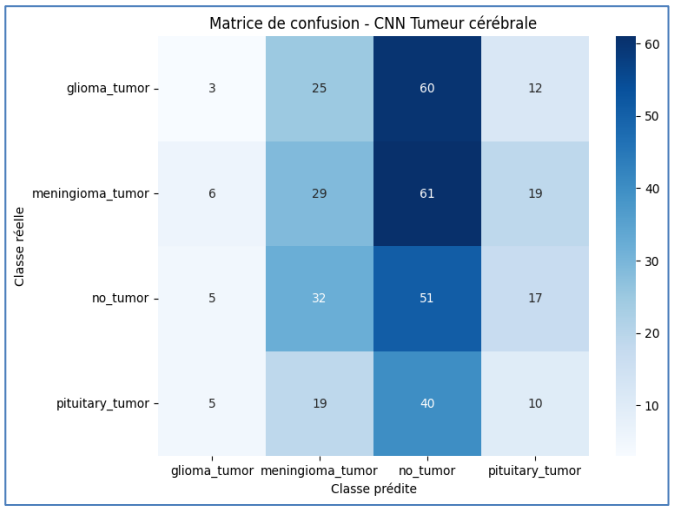

Figure 2:CNN confusion matrix

B. ResNet50 — Transfer Learning with Fine-Tuning

ResNet50 reached a training accuracy of 0.97 and a very low training loss of 0.068 after the combined 30-epoch procedure (20 epochs frozen + 10 epochs fine-tuning). However, the test accuracy dropped drastically to 0.28, representing a train to test gap of 0.69, the largest observed across all three models and a hallmark of severe overfitting. Table V presents the class level results.

TABLE V.
RESNET50 CLASSIFICATION REPORT (TEST SET, N = 394)

Class	Precision	Recall	F1-score	Support
Glioma tumor	0.00	0.00	0.00	100
Meningioma tumor	0.62	0.07	0.12	115
No tumor	0.27	0.96	0.42	105
Pituitary tumor	0.00	0.00	0.00	74
Weighted average	**0.25**	**0.28**	**0.15**	**394**

Figure 3 presents the learning curve. The significant gap between training and validation highlights the model's difficulty in generalizing despite transfer learning.

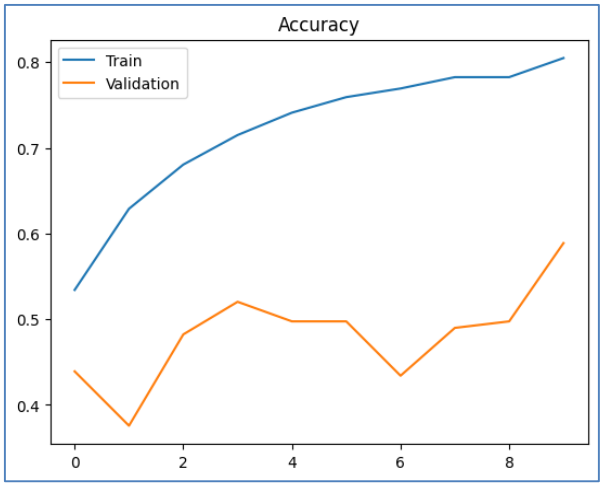

Figure 3:ResNet50 learning curve

The confusion matrix reveals an extreme prediction bias toward the no_tumor class, which is predicted for nearly all test samples (recall = 0.96), while glioma and pituitary tumor achieve zero recall. This behavior indicates a class collapse: rather than learning genuine tumor discriminative features, the fine-tuned model converged to a majority-class prediction strategy. This failure is attributable to two compounding factors: (1) the significant domain gap between ImageNet natural images and MRI scans, which undermines the transferability of pre-learned features; and (2) the limited dataset size, insufficient to properly adapt the 30 unfrozen layers during fine-tuning.

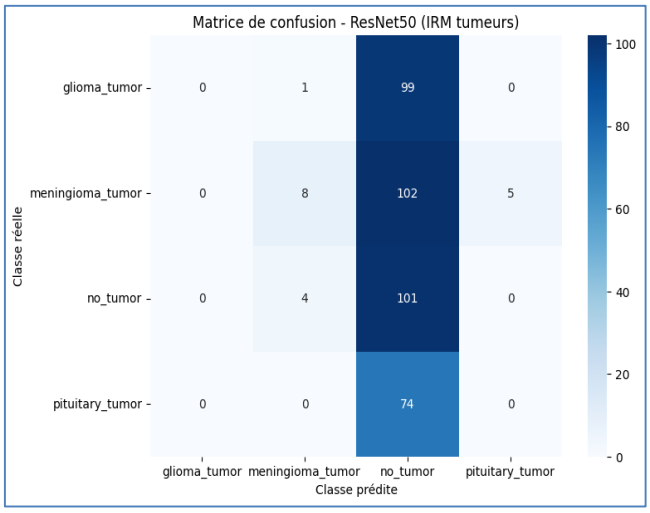

Figure 4: ResNet50 confusion matrix

C. Vision Transformer (ViT) — Best Overall Performance

The ViT model achieved a training accuracy of 0.76 and a test accuracy of 0.76 after 10 epochs, with a train–test gap of less than 0.01, indicating strong generalization and complete absence of overfitting. This is the most notable result of the study. Table VI presents the class-level performance.

TABLE VI.
VISION TRANSFORMER (VIT) CLASSIFICATION REPORT (TEST SET, N = 394)

Class	Precision	Recall	F1-score	Support
Glioma tumor	0.96	0.26	0.41	100
Meningioma tumor	0.70	0.98	0.82	115
No tumor	0.71	1.00	0.83	105
Pituitary tumor	0.97	0.77	0.86	74
Weighted average	**0.82**	**0.76**	**0.73**	**394**

Three of the four classes, namely meningioma tumor (F1 = 0.82), no tumor (F1 = 0.83), and pituitary tumor (F1 = 0.86), are classified with high and balanced precision recall scores. The glioma tumor class remains the most challenging, achieving a recall of only 0.26 despite a precision of 0.96: the model correctly identifies glioma when it predicts it, but misses 74% of actual glioma cases. This asymmetry is analyzed in the Discussion.

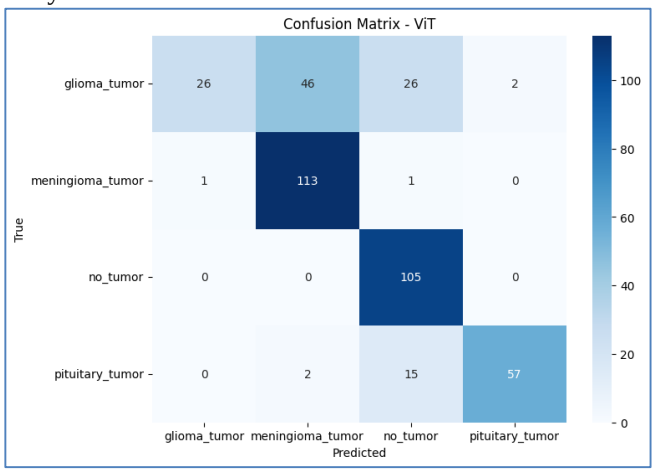

Figure 5: ViT confusion matrix

D. Comparative Summary

Tables VII and VIII synthesize the overall performance and generalization behavior across all three models.

TABLE VII.
OVERALL COMPARATIVE SUMMARY OF ALL MODELS

Model	Train Acc.	Test Acc.	W. Precision	W. Recall	W. F1
CNN	0.80	0.24	0.22	0.24	0.20
ResNet50	0.97	0.28	0.25	0.28	0.15
ViT	**0.76**	**0.76**	**0.82**	**0.76**	**0.73**

TABLE VIII.
OVERFITTING ANALYSIS: TRAIN–TEST ACCURACY GAP

Model	Train Accuracy	Test Accuracy	Train–Test Gap
CNN	0.80	0.24	0.56
ResNet50	0.97	0.28	0.69 (severe)
ViT	**0.76**	**0.76**	**< 0.01 (none)**

The ViT outperforms both CNN and ResNet50 across all metrics by a substantial margin. Critically, it is the only model that does not exhibit overfitting, achieving identical train and test accuracy (0.76). ResNet50, despite its pre-training advantage, produces the worst weighted F1-score (0.15) due to its complete failure on glioma and pituitary tumor classes.

III. DISCUSSION

A. Interpretation of Model Behaviours

The results reveal fundamentally different generalization behaviors across the three architectures. The CNN's poor test performance (accuracy = 0.24, weighted F1 = 0.20) reflects a key structural limitation: its convolutional filters capture only local spatial patterns within small receptive fields, which is insufficient for encoding the global structural features necessary to distinguish between visually similar tumor types in MRI. Furthermore, with 11.2M parameters, the majority concentrated in a single dense layer rather than distributed across convolutional blocks, the model is highly prone to overfitting on a dataset of 2,870 training images.

ResNet50's severe overfitting (train = 0.97, test = 0.28) is particularly noteworthy given its transfer learning advantage. The class collapse to no_tumor prediction demonstrates that the model learned a spurious statistical shortcut rather than genuine discriminative features. Two compounding mechanisms explain this failure: first, the domain gap between ImageNet (natural RGB images with diverse object categories) and brain MRI scans (grayscale images with domain-specific intensity patterns) substantially reduces the transferability of pre-learned representations; second, the limited training set (2,870 images) is insufficient to re-calibrate the 30 unfrozen layers during fine-tuning without severe overfitting.

The ViT's superior performance (test accuracy = 0.76, weighted F1 = 0.73, zero overfitting) is attributed to its multi-head self-attention mechanism, which computes pairwise dependencies between all 196 image patches simultaneously. This global receptive field enables the model to integrate morphological context, tumor boundary characteristics, and surrounding tissue patterns across the entire image, information that is diagnostically critical but spatially distributed. The pre-training on ImageNet-21k also provides a richer initialization that transfers more effectively than ResNet50, likely because the ViT's patch-level representation is more domain-agnostic.

B. Analysis of Glioma Classification Difficulty

The consistently poor recall for glioma across all models (CNN: 0.03, ResNet50: 0.00, ViT: 0.26) warrants specific attention, as glioma is clinically the most severe and heterogeneous tumor type in this dataset. Its visual presentation in MRI is highly variable: gliomas can appear iso-intense with surrounding tissue, enhance heterogeneously with contrast agents, or infiltrate neighboring structures in patterns that overlap significantly with meningioma and normal tissue. This intra-class variability, combined with 826 training samples, makes glioma the hardest class to learn. Future work should investigate class-imbalance correction strategies such as focal loss, targeted oversampling (SMOTE), or class-weighted training to specifically improve glioma recall.

C. Clinical Relevance and Deployment Constraints

A test accuracy of 0.76 and a glioma recall of 0.26 for the best-performing model are insufficient for clinical deployment. Medical AI diagnostic tools are typically subject to regulatory standards (e.g., FDA 510(k) pathway, EU MDR for Class IIb medical devices) requiring sensitivity and specificity exceeding 90% on large, demographically representative, multi-institutional datasets. The results reported here should therefore be interpreted as a research benchmark, not as a clinical readiness indicator. Any translation to clinical practice would require prospective validation studies on larger, multi-scanner, multi-site datasets with appropriate demographic diversity.

Furthermore, none of the three models were evaluated for interpretability, which is a significant limitation in a medical context. Clinicians require not only accurate predictions but also visual explanations identifying which image regions drove the model's decision. Gradient-weighted Class Activation Mapping (Grad-CAM) for CNN and ResNet50, and attention rollout or attention map visualization for ViT, represent natural next steps that would substantially enhance clinical trustworthiness and support regulatory review.

D. Methodological Limitations

Several methodological limitations must be acknowledged. First, the holdout validation strategy, using the test set for both epoch level monitoring and final evaluation, introduces optimistic bias. A stratified k-fold cross-validation (k = 5 or 10) would yield more reliable performance estimates. Second, the single-source dataset (Kaggle) does not capture scanner variability, imaging protocol differences, or patient demographic diversity across hospitals, limiting the generalizability of the findings. Third, the comparison would benefit from additional competitive baselines such as EfficientNet, DenseNet, or hybrid CNN-Transformer architectures, which have demonstrated strong performance in medical imaging benchmarks and would more rigorously contextualize the ViT's relative advantage.

IV. CONCLUSION

This study presented a systematic and unified comparative analysis of three deep learning architectures: a custom CNN, ResNet50 with two phase transfer learning and fine tuning, and a pre trained Vision Transformer, for brain tumor classification from MRI images. All models were evaluated on the same standardized four class dataset (3,264 images) under identical experimental conditions including preprocessing, batch size, optimizer, and evaluation protocol.

The experimental results demonstrate clear performance differentiation across the three architectures. The CNN achieves limited generalization (test accuracy = 0.24, weighted F1 = 0.20) due to its shallow convolutional structure and local feature extraction constraint. ResNet50 suffers from severe overfitting (train accuracy = 0.97, test accuracy = 0.28, train–test gap = 0.69), attributable to the domain gap between ImageNet pre-training and MRI data and to the limited training set size. The Vision Transformer achieves the best overall performance (test accuracy = 0.76, weighted F1 = 0.73) with zero overfitting, benefiting from its global dependency modeling via multi-head self-attention and rich ImageNet-21k initialization.

These findings confirm that attention-based architectures represent a promising direction for complex medical image classification tasks where global spatial reasoning is essential. However, the glioma recall of 0.26 in the best model, the absence of interpretability analysis, and the use of a single-source limited dataset collectively underscore the gap that remains before clinical deployment becomes feasible.

Future work will focus on five research directions: (1) incorporating hybrid CNN-Transformer architectures to jointly leverage local and global feature extraction; (2) applying Grad-CAM and attention map visualization to provide clinically interpretable predictions; (3) extending the evaluation to larger, multi-institutional, multi-scanner MRI datasets to assess generalizability; (4) implementing advanced regularization and class-balancing strategies (focal loss, mixup, SMOTE) to improve glioma classification; and (5) conducting prospective clinical validation studies to assess real-world diagnostic utility in collaboration with radiology departments.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 25, 2012, pp. 1097–1105.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," in Proc. International Conference on Learning Representations (ICLR), 2021.

[5] J. Cheng, W. Huang, S. Cao, R. Yang, W. Yang, Z. Yun, Z. Wang, and Q. Feng, "Enhanced Performance of Brain Tumor Classification via Tumor Region Augmentation and Partition," PLOS ONE, vol. 10, no. 10, 2015.

[6] M. M. Badža and M. Č. Barjaktarović, "Classification of Brain Tumors from MRI Images Using a Convolutional Neural Network," Applied Sciences, vol. 10, no. 6, p. 1999, 2020.

[7] S. Khan, N. Islam, Z. Jan, I. U. Din, and J. J. P. C. Rodrigues, "A Novel Deep Learning Based Framework for the Detection and Classification of Breast Cancer Using Transfer Learning," Pattern Recognition Letters, vol. 125, pp. 1–6, 2019.

[8] S. Deepak and P. M. Ameer, "Brain Tumor Classification Using Deep CNN Features via Transfer Learning," Computers in Biology and Medicine, vol. 111, p. 103345, 2019.

[9] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training Data-Efficient Image Transformers and Distillation through Attention," in Proc. International Conference on Machine Learning (ICML), 2021, pp. 10347–10357.

[10] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D Medical Image Segmentation," in Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 574–584.

[11] F. Oshasha, M. Kayembe et al., "EDCST-Rain: Enhanced Density-Aware Cross-Scale Transformer for Robust Object Classification Under Diverse Rainfall Conditions," 2026.

[12] R. Azad, A. Kazerouni, M. Heidari, E. K. Aghdam, A. Molaei, Y. Jia, A. Jose, R. Roy, and D. Merhof, "Advances in Medical Image Analysis with Vision Transformers: A Comprehensive Review," Medical Image Analysis, vol. 91, p. 103024, 2024.

[13] N. Takahashi, Y. Sugimoto, and T. Nakamura, "Comparison of Vision Transformers and Convolutional Neural Networks for Medical Image Classification," Journal of Medical Systems, vol. 48, no. 1, 2024.

[14] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," arXiv preprint arXiv:2102.04306, 2021.