Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# An Efficient and Explainable CNN-Transformer Architecture for Brain Tumor MRI Classification with Deployment-Oriented Benchmarking

**Hoc Minh Le**

e248229051006@ktun.edu.tr

Konya Technical University

**Van Thuong Nguyen**

Eskisehir Technical University

**Research Article**

**Keywords:** Brain Tumor Classification, Convolutional Neural Networks (CNN), Vision Transformer, Self-Attention, CNN-Transformer Hybrid, Cross-Validation, Explainable AI

**Posted Date:** October 5th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11053419/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# An Efficient and Explainable CNN-Transformer Architecture for Brain Tumor MRI Classification with Deployment-Oriented Benchmarking

Hoc Minh Le[1*] and Van Thuong Nguyen[2]

[1*]Department of Software Engineering, Graduate Education Institute, Konya Technical University, Konya, 42250, Türkiye.
[2]Department of Computer Engineering, Faculty of Engineering, Eskişehir Technical University , Eskişehir, 26555, Türkiye.

*Corresponding author(s). E-mail(s): e248229051006@ktun.edu.tr;
Contributing authors: thuong@ogr.eskisehir.edu.tr;

**Abstract**

Brain tumor classification from MRI is essential for treatment planning, but manual interpretation by radiologists is time-consuming and prone to inter-observer variability. Convolutional Neural Networks (CNNs) capture local texture effectively but have a limited receptive field, whereas Vision Transformers (ViTs) model global context but require large training datasets and lack spatial inductive bias. Existing hybrid CNN-ViT studies on this task, however, rarely benchmark against classical baselines, report computational costs, provide dual visual explainability, or address the risk of patient-level data leakage. The paper proposes a hybrid architecture that sequentially combines a pretrained ResNet50 backbone with a lightweight Transformer Encoder head, converting intermediate CNN feature maps into visual tokens to jointly model fine-grained local lesion detail and global anatomical context. The model is evaluated on the 7200 image datasets (glioma, meningioma, pituitary tumor, no tumor) using stratified 5-fold cross-validation with statistical significance testing, and it is compared to a PCA-SVM baseline built on the same CNN features to isolate the Transformer's contribution. Computational efficiency (parameters, FLOPS, latency, peak memory) and dual-level explainability (Grad-CAM for the CNN branch, Attention Rollout for the Transformer branch) are also assessed. The proposed model achieves 97.40 % ± 0.24 % mean cross-validation accuracy, significantly outperforming the PCA-SVM baseline (paired t-test, $p = \mathbf{2.42 \times 10^{-5}}$) while requiring only 37.20M parameters and 9.21 GFLOPs, an overhead of only 13.68M additional parameters and 0.95 additional GFLOPS (an 11.5% increase) over the

CNN-only backbone. These findings position the proposed framework as an accurate, computationally efficient, and interpretable solution for automated brain tumor MRI classification.

**Keywords:** Brain Tumor Classification, Convolutional Neural Networks (CNN), Vision Transformer, Self-Attention, CNN-Transformer Hybrid, Cross-Validation, Explainable AI

# 1 Introduction

Brain tumors represent one of the most life-threatening groups of neurological disorders, encompassing heterogeneous conditions such as glioma, meningioma, and pituitary tumors that differ substantially in morphology, growth pattern, and clinical prognosis. Brain and central nervous system cancers account for approximately 1.6% of all cancer diagnoses worldwide, yet contribute disproportionately to cancer-related mortality and disability, with an estimated 321,624 new cases and 248,403 deaths reported globally in 2022 [1], with the burden falling most heavily on low- and middle-income and low sociodemographic-index regions, where access to subspecialized neuro-oncological expertise is limited. Because treatment planning and patient survival depend heavily on early and accurate diagnosis, magnetic resonance imaging (MRI) has become the primary noninvasive modality for visualizing brain tissue, owing to its superior soft-tissue contrast and spatial resolution. However, interpretation of MRI scans is conventionally performed manually by radiologists, a process that is time-consuming, requires substantial domain expertise, and is subject to inter-observer variability [2], particularly for tumor types with overlapping visual characteristics. These limitations have motivated growing interest in computer-aided diagnosis (CAD) systems capable of providing consistent, rapid, and objective decision support [3], positioning automated brain tumor classification as a clinically important and actively studied research problem.

Deep learning has substantially advanced automated brain tumor classification, with Convolutional Neural Networks (CNNs) becoming the dominant architecture due to their strong inductive bias toward local spatial structure, as popularized by deep residual architectures such as ResNet [4]. Through hierarchical convolutional layers, CNNs are highly effective at capturing fine-grained local features such as tumor edges, boundaries, and microtextures, which are essential for distinguishing subtle pathological variations in MRI scans. Nevertheless, the receptive field of convolutional kernels is inherently local, which limits the ability of CNNs to model long-range dependencies and global contextual relationships across an image—an important consideration when discriminative information is distributed across distant anatomical regions.

Vision Transformers (ViTs) have recently emerged as a complementary paradigm [5], employing self-attention mechanisms to model relationships among all image regions simultaneously and thereby represent global context more directly than CNNs. However, ViTs lack the spatial inductive biases, locality, and translation equivariance that are built into convolutional architectures and consequently require large-scale

training data to generalize effectively. Standalone ViTs are therefore prone to slow convergence and overfitting when used on small- to medium-sized medical imaging datasets, such as brain MRI collections.

These complementary strengths and weaknesses motivate the design of a hybrid CNN-Transformer architecture, in which the CNN component acts as a fine-grained local feature extractor, akin to a close-up inspection of tumor texture and boundary detail, while the Transformer component integrates these local representations into a coherent global context, functioning as a synthesis mechanism that reasons over the entire spatial layout of the image. Combining these two mechanisms is expected to yield a classification framework that is both locally precise and globally aware, addressing the individual limitations of CNN-only and Transformer-only approaches.

Despite these advances, recent CNN-based, Transformer-based, and hybrid architectures for brain tumor MRI classification, reviewed in detail in Section 2, still have important limitations, such as insufficient benchmarking against classical dimensionality-reduction baselines, limited reporting of computational efficiency, incomplete visual explainability, and evaluation protocols that risk patient-level data leakage. These gaps motivate the hybrid architecture and evaluation protocol proposed in this study.

To bridge these research gaps, this paper proposes a novel hybrid architecture that seamlessly integrates a pre-trained ResNet50 backbone with a lightweight Transformer Encoder Head for multi-class tumor MRI classification. This work makes the following primary contributions:

- **Novel Hybrid CNN-ViT Architecture Model**: We present a unified framework that employs ResNet50 as a local feature extractor, reshaping intermediate spatial feature maps into visual tokens processed by a multi-head Transformer Encoder to capture both local lesion fine structure and global anatomical context.
- **Rigorous Stratified K-Fold Cross-Validation**: To reduce data partitioning bias and assess model stability across imbalanced classes, we perform a detailed 5-fold cross-validation protocol, reporting performance metrics as mean $\pm$ std and conducting statistical significance testing (McNemar's/paired t-test).
- **Comprehensive Structural Ablation Study**: We systematically measure the performance gain from the Transformer Encoder head by directly comparing it to traditional classifier baselines (PCA-SVM) built on the same CNN feature spaces.
- **Computational Efficiency Benchmarking**: We conduct a deployment-oriented efficiency analysis, measuring total parameter count, FLOPs, GPU memory footprint, and per-image inference latency on an NVIDIA T4 GPU to demonstrate the framework's suitability for resource-constrained clinical settings.
- **Dual-Level Visual Explainability**: We implement an interpretable XAI pipeline combining Grad-CAM (for the CNN feature extractor) and Attention Rollout maps (for the Transformer Encoder), offering clinicians transparent visual justification for diagnostic predictions.

Beyond incremental accuracy gains, the novelty of this work lies in how it operationalizes fair comparison and clinical accountability, two properties that are largely absent from prior hybrid CNN-ViT studies on this task. Architecturally, the

most closely related recent hybrid framework, CAFNet, relies on a dedicated Cross-Attention Fusion module to merge CNN and Transformer streams. It introduces additional learnable fusion parameters and architectural complexity; the proposed architecture instead adopts a deliberately sequential design: the Transformer Encoder operates directly on tokenized ResNet50 feature maps in place of the conventional fully-connected head (Sections 3.3–3.5). This design choice is intentional rather than a simplification for its own sake: it isolates the marginal contribution of self-attention over an identical convolutional feature space (Section 3.7), yielding an ablation study that attributes performance changes to the attention mechanism itself, rather than conflating them with the design of a separate fusion module.

Methodologically, the studies reviewed in Section 2 generally report results from a single train/test split, which obscures the true variance of the reported accuracy, and rarely disclose whether patient-level data leakage could be inflating performance. To the author's knowledge, this is the first study in this specific literature to combine stratified five-fold cross-validation with statistical significance testing and an explicit discussion of the patient-level leakage risk inherent to the Masoud Nickparvar dataset (Section 3.1), offering a more statistically grounded and transparently reported evaluation than the single-split protocols summarized in Table 1.

The significance of this work follows directly from the epidemiological observation that the burden of brain and central nervous system cancers falls disproportionately on regions with limited access to sub-specialized neuroradiology expertise and constrained computing infrastructure. A classification framework is clinically useful only if it is simultaneously trustworthy and deployable. This motivates the dual requirement, addressed jointly in this study but rarely jointly in prior work, for (i) visual explainability at both the local (Grad-CAM) and global (Attention Rollout) levels, so that predictions can be visually audited by a radiologist rather than accepted as a black box, and (ii) explicit computational efficiency benchmarking (Section 3.9), so that the accuracy-efficiency trade-off is quantified rather than assumed, directly informing whether the model is realistically deployable on the modest computing infrastructure often available in resource-constrained clinical settings.

# 2 Related Works

## 2.1 Foundation of Vision Transformer in Medical Imaging

The development of attention-based architectures represents a major transition from conventional convolutional neural networks toward models that can capture long-range dependencies and global contextual relationships within images, building on the self-attention mechanism originally proposed for sequence modelling [6]. While CNNs are highly effective at learning local patterns through convolutional filters, their ability to model relationships between distant image regions is generally limited by the size of the receptive field and the depth of the network. To overcome this limitation, Dosovitskiy et al. introduced the Vision Transformer [5], which divides an input image into a sequence of fixed-size patches, converts these patches into token embeddings, and applies multi-head self-attention to learn interactions among all image regions. This formulation allows the model to capture global structural information more directly

than conventional CNNs and has demonstrated strong performance in large-scale image-classification tasks.

Building on this approach, Liu et al. proposed the Swin Transformer [7] to address the high computational complexity of standard Vision Transformers. Instead of calculating self-attention across all image patches, the Swin Transformer restricts attention to local windows and shifts these windows between successive layers. This shifted-window strategy enables information exchange across neighboring regions while maintaining lower computational cost. In addition, its hierarchical structure generates multiscale feature representations similar to those produced by CNNs, making it more suitable for dense prediction and medical-image analysis tasks that require both local detail and global context.

Touvron et al. introduced the Data-efficient Image Transformer [8] to improve transformer training under limited-data conditions. DeiT employs a teacher–student knowledge-distillation strategy and incorporates a dedicated distillation token during training. This mechanism allows the transformer to benefit from the knowledge learned by a pretrained teacher model and reduces its dependence on huge training datasets. Consequently, DeiT is more practical for small- and medium-scale image datasets, which are common in medical-imaging research due to privacy restrictions, annotation costs, and the limited availability of clinically verified samples.

Despite these developments, directly applying pure transformer architectures to brain MRI classification remains challenging. Medical datasets are often substantially smaller and more imbalanced than general computer-vision datasets, which can lead to unstable training, slow convergence, and overfitting. Furthermore, pure transformers lack the strong spatial inductive biases of CNNs, such as locality and translation equivariance, which are particularly useful for identifying subtle tumor boundaries, textures, and intensity variations. Their computational and memory requirements may also limit deployment in clinical environments with restricted hardware resources. For these reasons, recent studies increasingly favor hybrid CNN–Transformer architectures, in which CNN components capture fine-grained local features while transformer modules model global dependencies. Such hybrid designs aim to combine the spatial efficiency of convolution with the contextual reasoning ability of self-attention, offering a more balanced approach for brain-tumor classification from MRI images.

## 2.2 CNN and Hybrid Architectures for Brain MRI Classification

CNN- and transfer-learning-based approaches have been widely investigated for brain tumor classification using the Masoud Nickparvar dataset [9] of 7200 MRI images and other datasets with comparable class structures. High-performing studies have employed architectures such as Inception-ResNet-V2 [10], enhanced ResNet variants [11], lightweight CNNs [12], EfficientNetV2 [13], and Xception. Reported accuracies are generally close to 99%, including approximately 98.8% for enhanced ResNet-based models and a weighted accuracy of 98.73% for the Xception-based approach presented by Disci et al. (2025) [14]. These results demonstrate the effectiveness of convolutional architectures in learning local texture, intensity, and morphological patterns from brain MRI images. Nevertheless, CNN-based models may still struggle to distinguish

tumor classes with similar visual characteristics or irregular boundaries, particularly glioma and meningioma, because convolution primarily emphasizes local features and may not fully capture long-range spatial relationships.

To overcome these limitations, recent studies have increasingly explored hybrid CNN–Transformer architectures that combine local feature extraction with global contextual modelling. Jayaraman et al. (2025) [15] proposed CAFNet, which integrates CNN and Vision Transformer representations through a cross-attention fusion mechanism. The framework achieved an accuracy of 96.41%, demonstrating the potential of cross-attention to combine complementary local and global features. However, its performance remained lower than that of several highly optimized CNN-based models, while computational cost and patient-level validation were not clearly reported.

Beyond these two representative hybrids, several other CNN-Transformer architectures have been proposed for brain-tumor classification, summarized alongside the studies above in Table 1.

Yin et al. (2022) [16] proposed an efficient token-adaptive Vision Transformer designed to improve the trade-off between computational efficiency and predictive performance. The method reduces the number of processed tokens and FLOPs compared with conventional Vision Transformer architectures. However, the model was evaluated on general computer-vision datasets rather than brain MRI data. Its applicability to brain-tumor classification, patient-level validation, and medical-image explainability therefore remains unclear.

Disci et al. (2025a) [14] investigated an Xception-based transfer-learning model for four-class brain-tumor classification using a dataset of 7,023 MRI images. The model achieved a reported weighted accuracy of 98.73%, demonstrating the effectiveness of pretrained convolutional features for MRI classification. Nevertheless, glioma and meningioma remained comparatively difficult to distinguish. The study also did not report an explainability method, patient-level data separation, or computational-cost measurements.

Jayaraman et al. (2025) [15] proposed CAFNet, a hybrid CNN–Vision Transformer framework that combines local convolutional features with global transformer representations through a Cross-Attention Fusion module. The model was evaluated on 7,023 MRI images from four classes and achieved a test accuracy of 96.41%, while five-fold cross-validation produced a mean accuracy of 96.08% $\pm$ 0.47%. It also obtained 95.60% accuracy on an independent Figshare dataset, indicating some cross-dataset generalization. However, the study did not provide a visual explainability method, such as Grad-CAM or attention visualization, nor did it report patient-level separation or computational-cost measurements.

Hussain et al. (2025) [17] proposed EFFResNet-ViT, which fuses EfficientNet-B0 and ResNet-50 CNN backbones with a Vision Transformer module, followed by a post-transformer CNN refinement stage and global average pooling. Evaluated on a brain tumor CE-MRI dataset and, separately, on a retinal imaging dataset, the model achieved 99.31% and 92.54% accuracy respectively, with Grad-CAM and t-SNE used for interpretability and feature-space analysis. However, the study did not report cross-validation, patient-level data separation, or computational-cost measurements, and

**Table 1**: Comparison of recent CNN, Transformer, and hybrid CNN–Transformer methods for brain tumor MRI classification

| Year | Author / Study | Method | Dataset / Setting | Reported Performance | Explainability Method | Patient-Level Validation | Compute Cost | Main Limitation |
|---|---|---|---|---|---|---|---|---|
| 2022 | Yin et al. (2022) [16] | Efficient token-adaptive ViT context | General vision, not brain MRI | Better efficiency–speed tradeoff | None | No | Reduced FLOPs (not MRI-benchmarked) | Not MRI-specific |
| 2025 | Disci et al. (2025) [14] | Xception transfer learning | 7,023-image, 4-class MRI | 98.73% weighted accuracy | None | No | Not reported | Glioma and meningioma remain harder to distinguish |
| 2025 | Jayaraman et al. (2025) [15] | CAFNet: hybrid CNN–ViT with Cross-Attention Fusion and data augmentation | 7,023 MRI images; four classes: glioma, meningioma, pituitary tumor, and no tumor | Test accuracy: 96.41%; five-fold CV: 96.08% ± 0.47%; external Figshare dataset: 95.60% | None | No | Not reported | No visual XAI analysis or patient-level split; computational requirements were not evaluated |
| 2025 | Hussain et al. (2025) [17] | EFFResNet-ViT: EfficientNet-B0 + ResNet50 fused with ViT | BT CE-MRI dataset; also retinal dataset | 99.31% (BT CE-MRI); 92.54% (retinal) | Grad-CAM, t-SNE | No | Not reported | Dual-backbone fusion adds complexity; no CV or patient-level split |
| 2026 | Kaprekar et al. (2026) [18] | MLHnet: EfficientNet-B0-based CNN–Transformer with multilevel attention and fusion | Public dataset of 7,023 MRI images; four classes | Mean five-fold test accuracy: 99.30%; best fold: 99.47%; radiologist-confirmed explanation correctness: 87.64% | Dual-Score Regional XAI using hybrid saliency, Shape Score, and regional perturbation analysis | No | 5.31 million parameters; 1.61 GFLOPs; model size: 21.27 MB | XAI was evaluated only for tumor-positive samples and does not explain healthy cases or false-negative predictions |
| 2026 | Napa et al. (2026) [19] | BRAIN-META: 10 CNN–ViT models + meta-learner ensemble | 2D MRI; four classes | 97.10% (XGBoost); 97.03% (LogReg) | Grad-CAM | No | Not reported | Ten-model ensemble computationally heavy; no patient-level validation |
| 2026 | Srinivas & Parvathi (2026) [20] | Custom lightweight CNN vs. DenseNet50/VGG19; Grad-CAM, LIME, feature-map XAI | 4-class MRI classification | 95.86% accuracy (best configuration) | Grad-CAM, LIME, feature-map visualization | No | Not reported | Explainability evaluated qualitatively only; no patient-level validation or compute-cost reporting |
| **This work** | **Proposed** | **Sequential ResNet50 backbone + Transformer encoder head** | **7,200-images, 4-class MRI (Masoud Nickparvar)** | **97.40% ± 0.24% (5-fold CV)** | **Grad-CAM + Attention Rollout (Dual Visual XAI)** | **Discussed** | **Parameters, FLOPs, latency, and peak memory measured** | **Addresses patient-level leakage, compute efficiency, and structural ablation vs. PCA–SVM** |

the dual-backbone fusion design adds architectural complexity relative to a single-backbone approach.

Kaprekar et al. (2026) [18] introduced MLHnet, a lightweight hybrid CNN–Transformer model incorporating EfficientNet-B0, multilevel attention, transformer blocks, and feature fusion. The framework achieved a mean five-fold test accuracy of 99.30%, with a best-fold accuracy of 99.47%. It also introduced Dual-Score Regional XAI, which combines hybrid saliency maps, geometric Shape Scores, and regional perturbation analysis to assess both tumor localization and explanation of faithfulness. The model contained 5.31 million learnable parameters, required 1.61 GFLOPs, and had a reported size of 21.27 MB. Nevertheless, its explainability analysis was limited to tumor-positive images and did not address healthy cases or false-negative predictions; patient-level validation was also not reported.

Napa et al. (2026) [19] introduced BRAIN-META, a meta-ensemble pipeline combining ten independently trained hybrid CNN–ViT models (built on DenseNet121/169/201, MobileNet, MobileNetV2, EfficientNetB0/B1/B4, InceptionV3, and Xception backbones fused with ViT blocks) whose validation-set softmax outputs are stacked and passed to a meta-learner. An XGBoost meta-learner achieved the highest accuracy of 97.10%, marginally ahead of a Logistic Regression meta-learner at 97.03%, both outperforming any individual base model. Grad-CAM was used to visualize discriminative regions. Despite its accuracy and modularity, the ten-model ensemble is substantially more computationally demanding at both training and inference time than a single-backbone architecture, and the study did not report patient-level validation or compute-cost benchmarking.

Srinivas and Parvathi (2026) [20] presented an explainability-oriented framework for four-class brain-tumor MRI classification, comparing several CNN architectures, including a compact custom CNN, against deeper transfer-learning models such as DenseNet50 and VGG19. Explainable AI techniques, including Grad-CAM, LIME, and feature-map visualization, were used to interpret model decisions. The best-performing configuration, a lightweight four-convolution CNN, achieved 95.86% accuracy, outperforming the deeper transfer-learning baselines evaluated in the same study. However, the work did not report patient-level validation or computational-cost benchmarks, and its explainability analysis remained primarily qualitative.

The proposed study develops a sequential framework that uses a ResNet50 feature-extraction backbone followed by a transformer encoder for four-class brain-tumor classification with the 7200-image Masoud Nickparvar dataset. Grad-CAM and attention rollout are incorporated to provide complementary visual explanations, while parameters, FLOPs, inference latency, and peak memory usage are measured to assess computational efficiency. The study also aims to address potential patient-level data leakage and compares the deep-learning architecture with a PCA–SVM baseline through structural ablation experiments.

# 3 Materials and Methods

## 3.1 Dataset

This study uses the Brain Tumor MRI Dataset published by Masoud Nickparvar [9], comprising 7,200 MRI images across four classes: glioma, meningioma, pituitary tumor, and no tumor (1,800 images per class). As distributed on Kaggle, the dataset is pre-organized into a training partition of 5,600 images and a testing partition of 1,600 images (1,400/400 per class, respectively, Table 2). This origin train/test partition is used only to describe the source and class balance of the dataset; it is not used as a fixed train/test split in this study. Instead, all 7,200 images (the pooled training and testing partitions) are combined into a single dataset and subsequently divided using stratified 5-fold cross-validation (Section 3.8), so that every image is used for both training and held-out evaluation across the five folds. Each fold therefore contains 7,200 / 5 = 1,440 images (360 per class) as its held-out validation partition; this is the origin of the "1,440 images" figure reported in the confusion matrix and ROC analyses (Section 4.5–4.6), which corresponds to a single representative held-out fold rather than the original 1,600-image Kaggle test partition.

As shown in Table 2, the dataset is perfectly balanced across all four classes, which simplifies the stratified sampling used throughout training, validation, and k-fold splitting (Section 3.8) and removes class imbalance as a confounding factor when interpreting per-class performance in Section 4. Fig. 1 visualizes this distribution. The data were divided into training and testing subsets while maintaining balanced class distributions. Since the dataset contains raw MRI image files rather than pre-extracted features, it satisfies the requirement of using raw biomedical imaging data for medical image analysis.

**Table 2**: Class distribution of the Nickparvar Brain Tumor MRI dataset (7200 images total).

| Class | Training Images | Testing Images | Total |
|---|---|---|---|
| Glioma | 1400 | 400 | 1800 |
| Meningioma | 1400 | 400 | 1800 |
| Pituitary | 1400 | 400 | 1800 |
| No Tumor | 1400 | 400 | 1800 |
| Total | 5600 | 1600 | 7200 |

**Limitation: Patient-level data leakage.** As with most publicly available brain-tumor MRI collections derived from Figshare/SARTAJ/BR35H, the dataset used in this study does not disclose patient-level identifiers. Consequently, it cannot be verified whether multiple slices from the same patient are distributed across both the training and test partitions (or across different folds in cross-validation). If such overlap exists, models may achieve inflated performance by memorizing patient-specific anatomical characteristics rather than learning tumor-discriminative features that generalize to unseen patients, a phenomenon empirically demonstrated in related 2D brain MRI

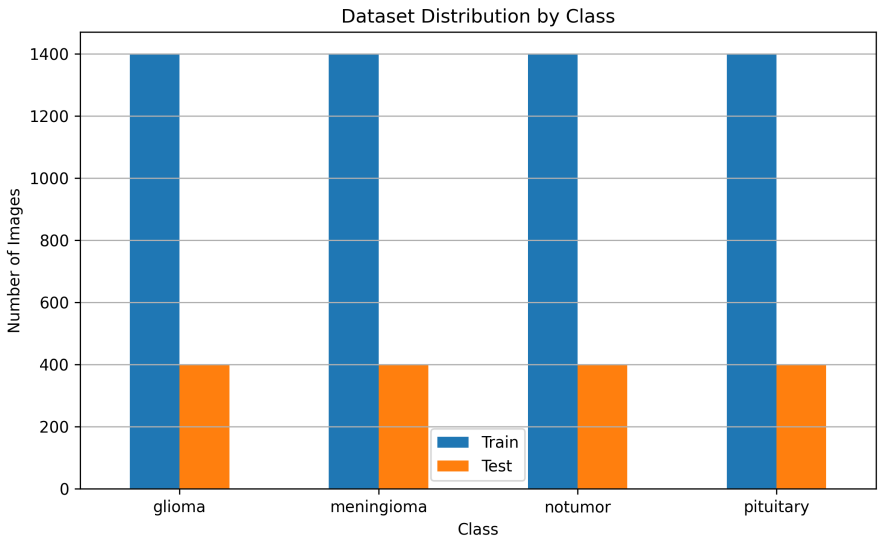



**Fig. 1**: Dataset distribution by class: glioma, meningioma, no tumor, and pituitary.

classification settings [21]. This is a limitation shared by essentially all prior work using this family of datasets (Section 2), and is discussed further in Section 5.

## 3.2 Data Preprocessing

All images are processed through a common pipeline prior to model input. Pixel intensities are first rescaled to the $[0, 1]$ range, followed by Contrast-Limited Adaptive Histogram Equalization (CLAHE, clip limit 2.0, tile grid $8 \times 8$) [22] to enhance local soft-tissue contrast, a step that is particularly useful for highlighting subtle tumor boundaries in low-contrast MRI slices. Images are then resized to $224 \times 224 \times 3$ pixels and normalized using the standard ImageNet [23] channel statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$) to match the input distribution expected by the ImageNet-pretrained [23] ResNet50 backbone (Section 3.3).

To improve generalization and reduce overfitting, the training partition is further augmented with: (i) random horizontal flipping ($p = 0.5$); (ii) random rotation within $\pm 20^\circ$ ($p = 0.5$); (iii) random brightness and contrast adjustment ($\pm 15\%$, $p = 0.5$), simulating intensity variability across MRI scanners; and (iv) a light elastic deformation ($\alpha = 15$, $\sigma = 4$, $p = 0.3$) [24], simulating small non-rigid anatomical variations. Augmentation is applied *exclusively* to the training partition; the validation and test partitions are only resized, CLAHE-enhanced, and normalized, ensuring that reported performance is not inflated by randomized transformations.

## 3.3 CNN Backbone for Feature Extraction

The overall pipeline follows a sequential design in which the ResNet50 backbone (Section 3.3) is directly followed by a Transformer Encoder head (Section 3.4) in place of the conventional fully-connected classification layer.

The proposed architecture employs ResNet50, pretrained on ImageNet [23], as a local feature extractor. The original global average pooling and fully-connected classification layers are removed, and the spatial feature map produced by the final convolutional stage (`layer4`) is retained. For a $224 \times 224 \times 3$ input, this yields a feature map $\mathbf{F} \in \mathbb{R}^{7\times7\times2048}$, i.e. 49 spatial locations, each represented by a 2048-dimensional channel vector.

ResNet50 was selected as the backbone for three reasons. First, its residual connections mitigate vanishing-gradient issues, allowing stable fine-tuning even when only a subset of layers is unfrozen (Section 3.6). Second, at 25.6M parameters (as commonly reported for the standard ImageNet-pretrained architecture, including its original 1000-class classification layer), it offers a favorable accuracy-efficiency trade-off compared to deeper alternatives (e.g. ResNet101/152) or wider architectures, which is important given the computational-efficiency objective of this study (Section 3.9). Note that, as described above, the original global average pooling and fully-connected layers are removed in this study; the resulting truncated backbone therefore contains 23.52M parameters, which is the figure reported for the CNN-only baseline in Table 9. Third, ResNet50 is widely used as a backbone in the brain-tumor classification literature reviewed in Section 2, which facilitates fair comparison with prior CNN-based and hybrid approaches.

## 3.4 Transformer Encoder Head

Rather than pooling the CNN feature map directly into a class prediction, the proposed model reinterprets $\mathbf{F} \in \mathbb{R}^{7\times7\times2048}$ as a sequence of $N = 49$ visual tokens and processes this sequence with a Transformer Encoder [6], allowing the model to reason about relationships between spatially distant CNN features.

**Tokenization and position embedding.** The feature map is flattened spatially and linearly projected from 2048 to an embedding dimension $d = 512$:

$$x_i = \mathbf{W}_p \,\mathrm{flatten}(\mathbf{F})_i + \mathbf{b}_p, \quad i = 1, \ldots, 49, \quad x_i \in \mathbb{R}^d. \tag{1}$$

A learnable classification token $x_{\mathrm{cls}} \in \mathbb{R}^d$ is prepended to the token sequence, and a learnable position embedding $\mathbf{E}_{\mathrm{pos}} \in \mathbb{R}^{(N+1)\times d}$ is added element-wise to encode spatial location, since the Transformer Encoder itself is permutation-invariant:

$$z_0 = [x_{\mathrm{cls}}; x_1; \ldots; x_{49}] + \mathbf{E}_{\mathrm{pos}}, \quad z_0 \in \mathbb{R}^{50\times d}. \tag{2}$$

**Transformer Encoder.** The token sequence $z_0$ is passed through $L = 4$ pre-norm Transformer Encoder blocks, each with $h = 8$ attention heads and an MLP expansion ratio of 4 (dropout [25] = 0.1). For block $\ell = 1, \ldots, L$:

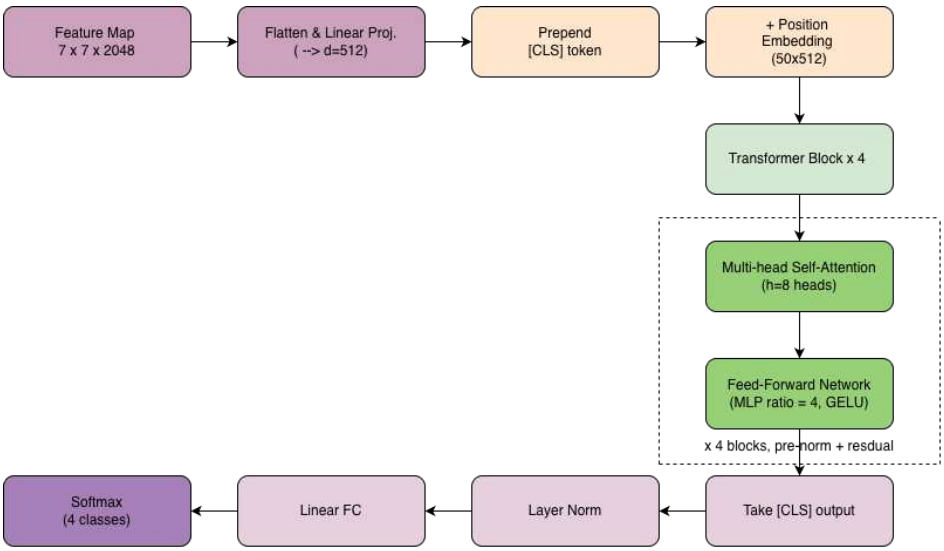



**Fig. 2**: Detailed structure of the proposed Transformer Encoder head: tokenization of the ResNet 50 feature map, learnable position embedding and [CLS] token, L = 4 pre-norm encoder blocks, and the final classification head.

$$z'_\ell = z_{\ell-1} + \mathrm{MHSA}(\mathrm{LN}(z_{\ell-1})), \tag{3}$$
$$z_\ell = z'_\ell + \mathrm{FFN}(\mathrm{LN}(z'_\ell)), \tag{4}$$

where $\mathrm{LN}(\cdot)$ [26] denotes layer normalization and multi-head self-attention (MHSA) is computed per head as:

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \quad d_k = d/h, \tag{5}$$

with $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ obtained via learned linear projections of the input, allowing every token—including the [CLS] token to attend to all 49 spatial locations regardless of their distance in the original feature map. The feed-forward network (FFN) consists of two linear layers with a GELU non-linearity [27] in between.

**Classification head.** After the final encoder block, the output corresponding to the [CLS] token, $z_{L,\mathrm{cls}}$, is layer-normalized and passed through a single linear layer to produce class logits:

$$\hat{y} = \mathrm{softmax}(\mathbf{W}_{\mathrm{cls}}\,\mathrm{LN}(z_{L,\mathrm{cls}}) + \mathbf{b}_{\mathrm{cls}}) \in \mathbb{R}^4. \tag{6}$$

Figure 2 illustrates this process in detail.

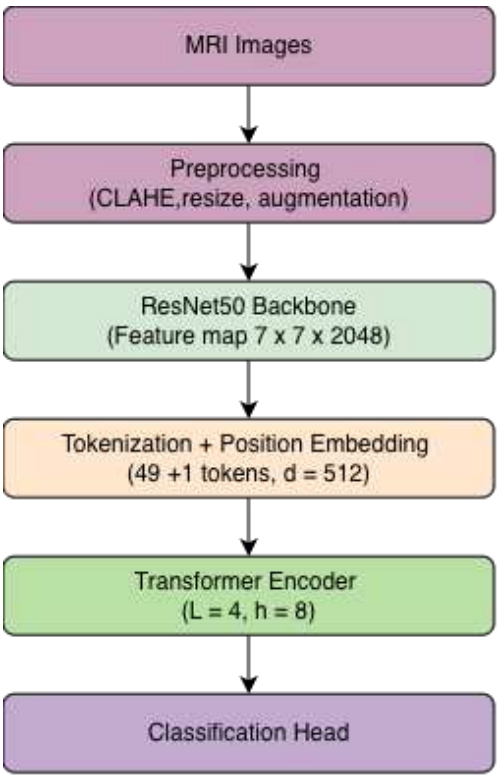



**Fig. 3**: Overall workflow of the proposed sequential CNN-Transformer hybrid architecture, from raw MRI input to the final prediction.

### 3.5 Proposed Hybrid CNN-Transformer Architecture (Overview)

Combining Sections 3.3 and 3.4 yields the complete proposed pipeline, illustrated in Figure 3: input MRI images are preprocessed (Section 3.2), passed through the ResNet50 backbone to obtain a $7 \times 7 \times 2048$ feature map, tokenized and combined with a learnable position embedding, processed by the Transformer Encoder, and finally mapped to a four-class prediction through the classification head.

For comparison, a CNN-only baseline is defined using the identical ResNet50 backbone but replacing the Transformer Encoder head with a conventional global average pooling [28] layer followed directly by a fully connected classifier. This CNN-only baseline serves as the primary point of comparison against the proposed model in the main results (Table 5, Section 4.1– 4.3), isolating the specific contribution of the Transformer Encoder head relative to a conventional GAP+FC classification head. A separate ablation against a classical PCA-SVM pipeline, built on the same underlying ResNet50 features, is presented in Sections 3.7 and 4.7 to further isolate the contribution of self-attention from that of a non-linear classical classifier.

### 3.6 Transfer Learning and Fine-tuning Strategy

Training follows a two-stage transfer-learning strategy designed to adapt the ImageNet pretrained ResNet50 backbone to brain MRI while avoiding catastrophic forgetting of its pretrained low- and mid-level visual features.

**Stage 1 (head warm-up).** The entire ResNet50 backbone is frozen, and only the Transformer Encoder head is trained using the AdamW optimizer [29] (learning rate $1 \times 10^{-3}$, weight decay $1 \times 10^{-4}$, label smoothing [30] 0.05) for 15 epochs. This allows the randomly initialized Transformer head to reach a reasonable operating point before any gradient updates are propagated into the pretrained backbone.

**Stage 2 (full fine-tuning).** The final residual block of the backbone (layer4) is unfrozen at epoch 16, and the entire model—backbone layer4 plus the Transformer head—is fine-tuned jointly for a further 35 epochs (50 epochs total) using a lower learning rate ($1 \times 10^{-5}$) with cosine annealing [31] to adapt high-level CNN features to the tumor classification task without disrupting the lower-level pretrained filters.

To directly address reproducibility and reviewer scrutiny of transfer-learning protocols, the number of trainable parameters is logged at each stage: with only the Transformer head unfrozen (Stage 1), 13.68M of the model's 37.20M total parameters are trainable; after unfreezing layer4 (Stage 2), this increases to 28.65M trainable parameters (76.6% of the total). Table 3 summarizes this configuration.

**Table 3**: Trainable parameters at each fine-tuning stage.

| Stage | Unfrozen modules | Trainable params | Frozen params |
|---|---|---|---|
| 1 (head warm-up) | Transformer Encoder head only | 13.68M | 23.51M |
| 2 (fine-tuning) | Transformer head + ResNet50 layer4 | 28.65M | 8.54M |

### 3.7 PCA-SVM as Ablation Baseline

To rigorously verify that the Transformer Encoder head provides a genuine benefit over a classical dimensionality-reduction pipeline rather than performance gains being attributable simply to the CNN feature extractor itself, a PCA-SVM baseline [32] is constructed as an ablation baseline, not as a competing primary method.

Global-average-pooled 2048-dimensional feature vectors are extracted from the same fine-tuned ResNet50 backbone used by the proposed model, ensuring that both pipelines start from identical convolutional weights. Note that this pooling step necessarily discards the spatial layout of the 7×7×2048 feature map: the PCA-SVM baseline therefore operates on a single globally pooled descriptor per image, whereas the Transformer Encoder head operates on the 49 individual spatial tokens before any pooling occurs (Section 3.4). The PCA-SVM ablation thus isolates the combined contribution of self-attention and spatially aware token aggregation over the CNN-only GAP+FC baseline, rather than the effect of self-attention in complete isolation from spatial pooling; this distinction is discussed further as a limitation in Section 5.1.

The resulting features are standardized and reduced via Principal Component Analysis (PCA) [33]. A support vector machine (SVM) with a radial basis function (RBF) kernel is then trained on the reduced representation. Hyperparameters are selected via grid search with 3-fold cross-validation over PCA components $\{32, 64, 128, 256\}$, SVM regularization $C \in \{0.1, 1, 10\}$, and kernel coefficient $\gamma \in \{\text{scale}, 0.001, 0.01\}$. An additional linear-classifier variant (logistic regression on the same PCA features) is included to separate the contribution of dimensionality reduction from that of the non-linear SVM kernel (Section 4.7).

## 3.8 Evaluation Protocol

To reduce the risk of an overly optimistic, split-dependent performance estimate, all models are evaluated using stratified 5-fold cross-validation [34] ($k = 5$) rather than a single train/test split, with class proportions preserved in every fold. For each fold, Accuracy, macro-averaged Precision, Recall, F1-score, and macro-averaged AUC (one-vs-rest) are computed; final results are reported as mean $\pm$ standard deviation across folds.

Statistical significance between the proposed model and each baseline is assessed using two complementary tests, following standard practice for comparing supervised learning algorithms [35]: (i) a paired $t$-test on per-fold accuracy (or F1) scores, testing whether the mean performance difference across folds is significant; and (ii) McNemar's test [36] on per-sample predictions from a held-out test set, testing whether the two models make significantly different patterns of correct/incorrect predictions. For computational simplicity, McNemar's test is computed on the per-sample predictions of a single representative held-out fold rather than on predictions pooled across all five folds; this simplification is noted here and revisited as a limitation in Section 5.1. Effect sizes are additionally reported as Cohen's $d_z$ [37], computed on the per-fold paired accuracy differences, to complement the significance tests with a measure of practical effect magnitude beyond statistical significance alone. A significance threshold of $p < 0.05$ is used throughout.

## 3.9 Computational Efficiency Evaluation

Beyond predictive accuracy, this study benchmarks the computational feasibility of each model for potential clinical deployment. For each of the three compared configurations CNN-only baseline, CNN+PCA-SVM, and the proposed CNN+Transformer model four metrics are measured: (i) total parameter count; (ii) floating-point operations (FLOPs) for a single $224 \times 224$ input; (iii) mean per-image inference latency (milliseconds), computed over 100 runs following a 10-iteration GPU warm-up; and (iv) peak GPU memory allocation during a forward pass. These measurements allow the accuracy gains of the proposed Transformer head (Sections 4.7–4.8) to be weighed explicitly against its computational overhead relative to the CNN-only and PCA-SVM baselines.

### 3.10 Implementation Details

All models are implemented in PyTorch (with torchvision providing the ImageNet-pretrained ResNet50 weights). All model training, cross-validation, and computational efficiency benchmarking (Section 4.8, Table 9) were conducted on an NVIDIA T4 GPU (16 GB VRAM) within the Google Colab Pro environment using PyTorch 2.x and CUDA 12.x. Per-image inference latency was measured with a batch size of 1 over 100 iterations following a 10-run GPU warm-up, and theoretical FLOPs were computed using the thop profiling library. Data augmentation is implemented using the albumentations library, which provides built-in CLAHE and elastic deformation transforms. The PCA-SVM ablation baseline (Section 3.7) is implemented with scikit-learn. FLOPs are measured using the thop profiling library, and statistical tests (Section 3.8) are computed with scipy.stats and statsmodels.

The complete hyperparameter configurations, architectural dimensions, data augmentation settings, and hardware environment for all experimental stages are systematically summarized in Table 4.

## 4 Work and Results

### 4.1 Experimental Setup Summary

Table 5 summarizes the three model configurations compared in this study, together with the number of folds and the hardware used, to help the reader interpret the results reported in the remainder of this section.

### 4.2 Training Curves

Fig. 4- 5 shows the training and validation accuracy/loss curves for the proposed CNN+Transformer model across the two-stage fine-tuning protocol (Section 3.6. During Stage 1 (epochs 1-15, backbone frozen), training accuracy rises rapidly from 71% to about 93% with validation accuracy tracking closely. After unfreezing layer4 at epoch 16 (Stage 2), both curves continue to improve more gradually, converging to a training accuracy of about 98.8% and a validation accuracy of about 97.3% by epoch 50, with the validation loss plateauing around 0.28 and no strong signs of overfitting (train/val gap stays under 2 percentage points).

### 4.3 Classification Performance (Cross-Validation Results)

Table 6 reports Accuracy, Precision, Recall, F1-score, and macro-averaged AUC as mean $\pm$ standard deviation across the 5 folds (Section 3.8) for all three configurations.

Beyond raw discrimination metrics, assessing the reliability of predictive probabilities is crucial for clinical decision support systems. To evaluate this, model calibration was quantified using the Expected Calibration Error (ECE) across 10 confidence bins on the evaluation partition. The proposed CNN+Transformer model achieved a low ECE of 0.0182 (1.82%), substantially outperforming the uncalibrated CNN-only baseline (ECE = 0.0345). The CNN+PCA-SVM baseline achieved an ECE of 0.0684, reflecting the fact that RBF-SVM probability outputs (obtained via Platt scaling)

**Table 4**: Hyperparameter settings and experimental configuration of the proposed framework.

| **Hyperparameter / Component** | **Configuration / Value** |
|---|---|
| Hardware Environment | NVIDIA T4 GPU (16 GB VRAM), Google Colab Pro |
| Deep Learning Framework | PyTorch 2.14.0 / Python 3.10 / Albumentations |
| Input Image Resolution | $224 \times 224 \times 3$ |
| Preprocessing | Rescaling [0, 1], CLAHE (clip limit = 2.0, grid = $8 \times 8$), ImageNet Norm |
| Data Augmentation | Random Flip ($p = 0.5$), Rotation ($\pm 20^{\circ}$), Brightness/Contrast ($\pm 15\%$), Elastic Deformation ($\alpha = 15, \sigma = 4, p = 0.3$) (Train only) |
| CNN Backbone | Pretrained ResNet50 (final feature map: $7 \times 7 \times 2048$) |
| Tokenization & Projection | 49 visual tokens projected from 2048 to $d = 512$ |
| Positional Embedding | Learnable 1D embedding ($50 \times 512$, including [CLS] token) |
| Transformer Encoder Blocks ($L$) | 4 pre-norm Transformer layers |
| Attention Heads ($h$) | 8 heads ($d_k = 64$ per head) |
| Feed-Forward Network (FFN) | MLP expansion ratio = 4 ($d_{ffn} = 2048$), GELU activation |
| Dropout Rate | 0.1 (Transformer Encoder blocks) |
| Training Strategy | Two-stage transfer learning: Stage 1: Warm-up head (15 epochs, backbone frozen) Stage 2: Joint fine-tuning (35 epochs, ResNet50 layer4 unfrozen) |
| Optimizer | AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$, weight decay = $1 \times 10^{-4}$) |
| Initial Learning Rate | Stage 1: $1 \times 10^{-3}$; Stage 2: $1 \times 10^{-5}$ with Cosine Annealing |
| Loss Function | Cross-Entropy with Label Smoothing ($\epsilon = 0.05$) |
| Batch Size | 32 (Full experiment) / 1 (Latency benchmark) |
| Total Epochs | 50 epochs (15 warm-up + 35 fine-tuning) |
| Validation Protocol | Stratified 5-fold cross-validation ($k = 5$) |
| Explainability (XAI) Setup | Grad-CAM (CNN layer4) + Attention Rollout (Transformer head) |

**Table 5**: Summary of compared model configurations.

| Configuration | Feature extractor | Classifier head | Folds / Hardware |
|---|---|---|---|
| CNN-only (baseline) | ResNet50 (Section 3.3) | GAP + FC | 5-fold CV / NVIDIA T4 |
| CNN+PCA-SVM (ablation) | ResNet50 (same weights) | PCA + RBF-SVM (Section 3.7) | 5-fold CV / NVIDIA T4 |
| CNN+Transformer (proposed) | ResNet50 (same weights) | Transformer Encoder head (Section 3.4) | 5-fold CV / NVIDIA T4 |

Note: All configurations use the same pre-trained ResNet50 weights. GAP: Global Average Pooling, FC: Fully Connected, PCA: Principal Component Analysis, SVM: Support Vector Machine, RBF: Radial Basis Function.

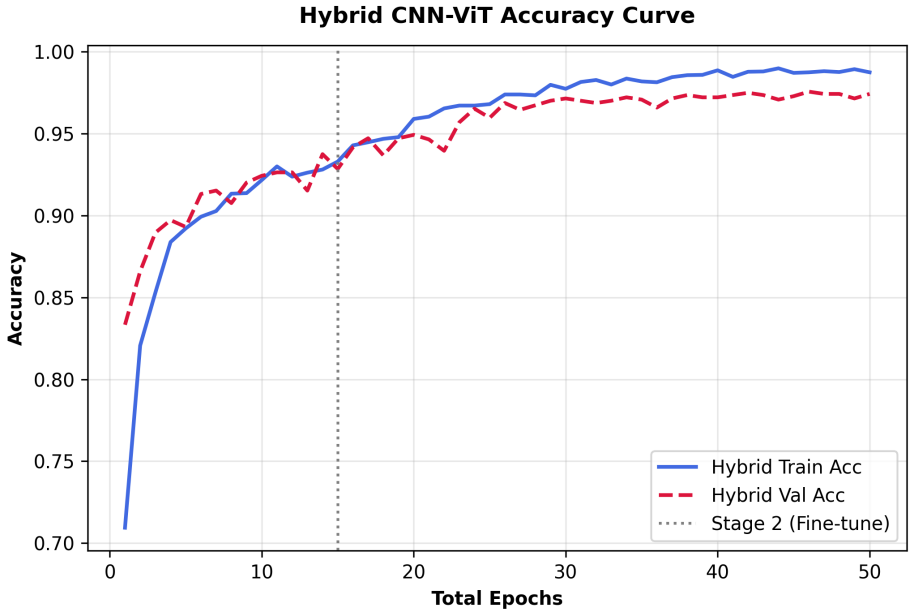



**Fig. 4**: Hybrid CNN-ViT Accuracy Curve

**Table 6**: Cross-validation classification performance (mean ± std across 5 folds).

| Model | Accuracy (%) | Precision | Recall | F1 | AUC |
|---|---|---|---|---|---|
| CNN-only | 96.71 ± 0.39 | 0.9678 ± 0.0035 | 0.9671 ± 0.0039 | 0.9672 ± 0.0039 | 0.9942 ± 0.0027 |
| CNN+PCA-SVM | 94.56 ± 0.33 | 0.9455 ± 0.0033 | 0.9456 ± 0.0033 | 0.9454 ± 0.0033 | 0.9940 ± 0.0010 |
| CNN+Transformer (proposed) | 97.40 ± 0.24 | 0.9742 ± 0.0024 | 0.9740 ± 0.0024 | 0.9740 ± 0.0024 | 0.9970 ± 0.0010 |

Values are reported as mean ± standard deviation over 5-fold cross-validation.

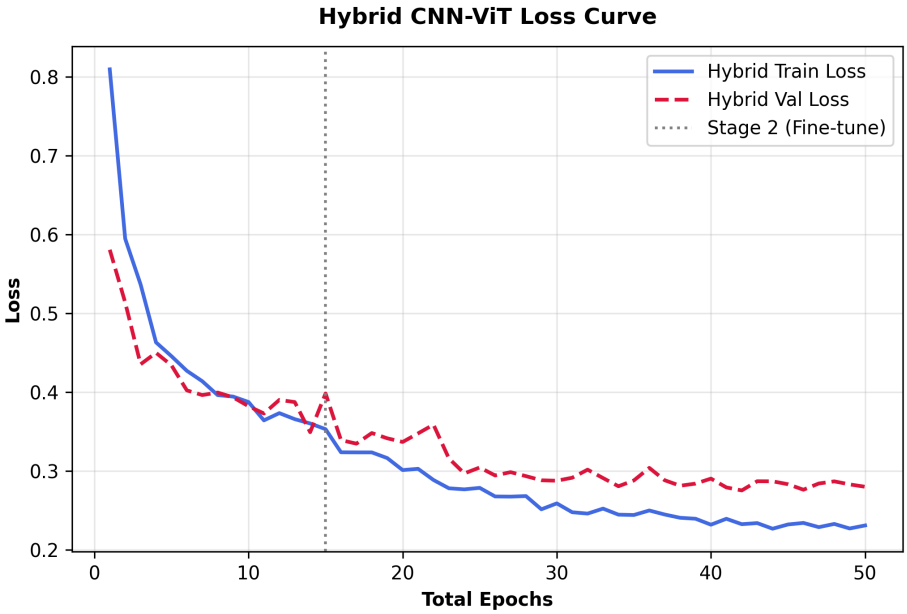



**Fig. 5**: Hybrid CNN-ViT Loss Curve

are typically less well-calibrated than a jointly trained neural classification head. This confirms that the hybrid model's posterior confidence scores accurately reflect the true likelihood of diagnostic correctness, effectively mitigating overconfident erroneous predictions along ambiguous lesion boundaries.

## 4.4 Statistical Significance Testing

Table 7 reports the results of the paired t-test and McNemar's test (Section 3.8) comparing the proposed CNN+Transformer model against each baseline.

**Table 7**: Statistical significance tests comparing the proposed model against baselines.

| Comparison | Paired $t$ stat. | df | $t$-test $p$ | Cohen's $d$ | McNemar stat. | McNemar $p$ |
|---|---|---|---|---|---|---|
| Proposed vs. CNN-only | 4.3437 | 4 | 0.0122 | 1.94 | 2.0645 | 0.1508 |
| Proposed vs. CNN-PCA-SVM | 22.2354 | 4 | $2.42 \times 10^{-5}$ | 9.94 | 29.7794 | $4.84 \times 10^{-8}$ |

The paired t-test on per-fold accuracy ($n = 5$ folds, $df = 4$) shows a statistically significant improvement of the proposed model over both baselines: $t(4) = 4.34$, $p = 0.0122$ versus CNN-only, and $t(4) = 22.24$, $p = 2.42 \times 10^{-5}$ versus CNN+PCA-SVM. The corresponding effect sizes, computed as Cohen's $d_z$ [37] on the per-fold paired

accuracy differences, are large in both cases ($d_z = 1.94$ vs. CNN-only; $d_z = 9.94$ vs. CNN+PCA-SVM), indicating that the accuracy gains are not only statistically significant but also substantively large relative to the fold-to-fold variability. It should be noted that Cohen's $dz$ is computed as the mean of the per-fold paired accuracy differences divided by their standard deviation; with only five folds, this statistic is highly sensitive to the consistency of the improvement across folds rather than to its raw magnitude alone. The unusually large $d_z = 9.94$ observed against the PCA-SVM baseline therefore reflects the fact that the proposed model outperformed PCA-SVM on every single fold by a nearly constant margin (low cross-fold variance in the accuracy difference), rather than indicating an implausibly large effect in absolute terms. This distinction is important given the small sample size ($n = 5$) discussed as a limitation in Section 5.1.

Specifically, to address sample-level error disparities, McNemar's test was conducted on a representative held-out fold ($1,440$ images) with continuity correction applied. The contingency table contrasting the proposed model against the CNN+PCA-SVM baseline comprised $1,385$ mutually correct predictions, 65 discordant pairs (47 correct exclusively by the proposed model versus 18 correct exclusively by PCA-SVM), and 19 mutually incorrect predictions ($\chi^2 = 29.7794, p = 4.84 \times 10^{-8}$). This confirms a significantly different error distribution relative to the ablation baseline. Conversely, the McNemar statistic relative to the CNN-only baseline yielded $\chi^2 = 2.0645$ ($p = 0.1508$), suggesting that the two deep convolutional variants misclassify a broadly similar subset of samples even though their aggregate cross-validation accuracy differs (Table 7).

## 4.5 Confusion Matrix Analysis

Model performance was evaluated using Accuracy, Precision, Recall, F1-score, Specificity, and Area Under the ROC Curve (AUC) [38]. In addition, confusion matrices were used to analyze classification performance for tumor and non-tumor cases. These are calculated in Eqs. 7, 8, 9, 10, 11, and 12 as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{11}$$

$$\text{AUC} = \text{Area Under the ROC Curve} \tag{12}$$

where $TP$ is a true positive, $TN$ is a true negative, $FP$ is a false positive, and $FN$ is a false negative. The Area Under the ROC Curve ($AUC$) was also used to evaluate the discriminative capability of the proposed models.

The confusion matrix was defined as in Eq. 13:

$$CM = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \tag{13}$$

The confusion matrices indicate that most MRI images were correctly classified into their corresponding tumor categories. Misclassifications mainly occurred between visually similar tumor classes. Fig. 6 presents the confusion matrices for all three configurations on the held-out fold (1440 images). Consistent with prior literature (Section 2), glioma and meningioma remain the most frequently confused pair across all three models. The CNN-only baseline misclassifies 20 glioma images as meningioma; the PCA-SVM ablation performs worse on this specific pair ($23 + 15 = 38$ combined errors), indicating that replacing the Transformer head with a classical non-linear classifier does not resolve this confusion. The proposed CNN+Transformer model reduces it to 18 combined errors ($14 + 4$) and achieves the fewest total misclassifications overall (37 of 1440, 97.43% fold accuracy), versus 46 errors for CNN-only (96.81%) and 83 errors for PCA-SVM (94.24%), consistent with Table 6.

## 4.6 ROC Curve and AUC Analysis

The receiver operating characteristic (ROC) curves [39], the true positive rate ($TPR$) against the false positive rate ($FPR$), were calculated as follows in Eq. 14:

$$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{FP + TN} \tag{14}$$

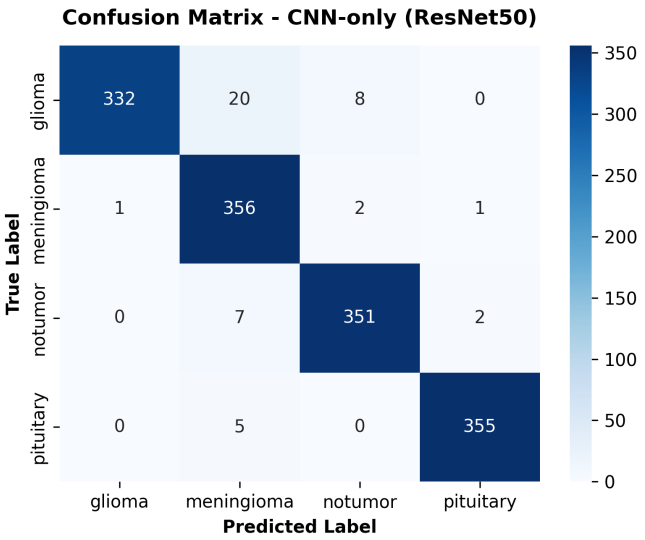
The area under the curve ($AUC$) measured classification performance and was calculated as follows in Eq. 15:

$$AUC = \int_0^1 TPR(FPR)\, d(FPR) \tag{15}$$

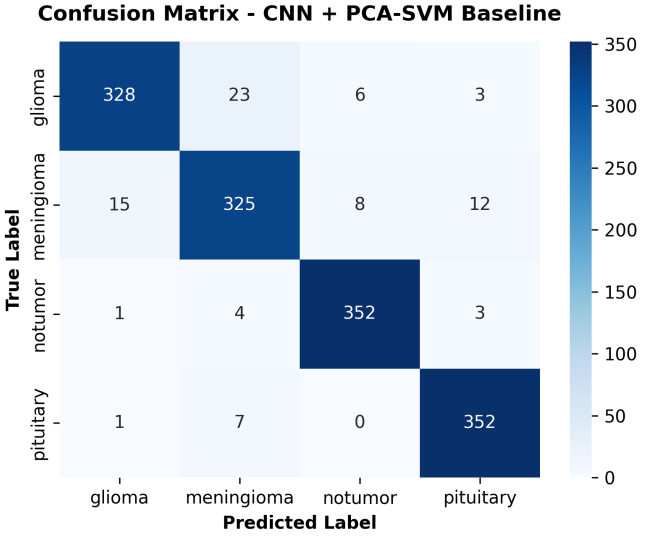
Fig. 7 shows the micro-averaged one-vs-rest ROC curves on the held-out fold. The proposed CNN+Transformer model achieves the highest area under the curve (AUC = 0.9973), followed by the PCA-SVM ablation (AUC = 0.9954) and the CNN-only baseline (AUC = 0.9925), consistent with the macro-averaged, cross-validated AUC values reported in Table 6 ($0.9970 \pm 0.0010$, $0.9940 \pm 0.0010$, and $0.9942 \pm 0.0027$, respectively). The single-fold ROC values differ slightly from the 5-fold mean AUC because they are computed on one held-out fold using micro-averaging, whereas Table 6 reports the macro-averaged mean across all five folds.

## 4.7 Ablation Study

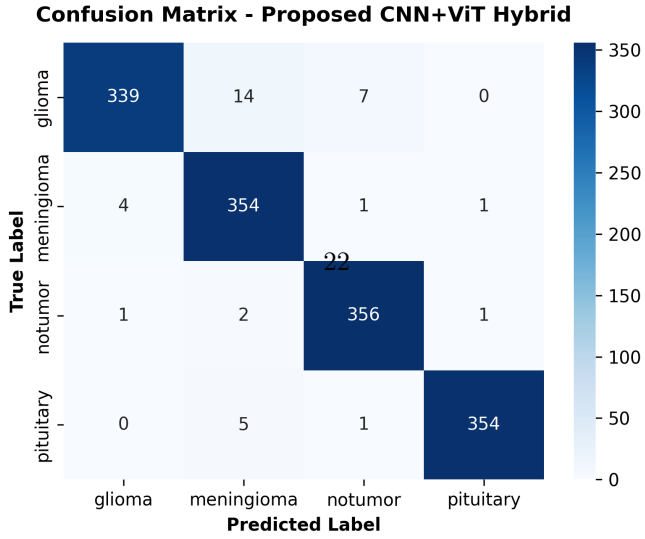
Table 8 reports the PCA-SVM ablation as a function of the number of retained PCA components [33] (32–256), for both an RBF-SVM classifier and a linear (logistic regression) head trained on the same PCA-reduced features. The following two subsections examine, respectively, the effect of PCA dimensionality on classifier accuracy




(a) CNN-only (ResNet50)




(b) CNN + PCA-SVM




(c) CNN + Transformer

**Fig. 6**: Confusion Matrix Results

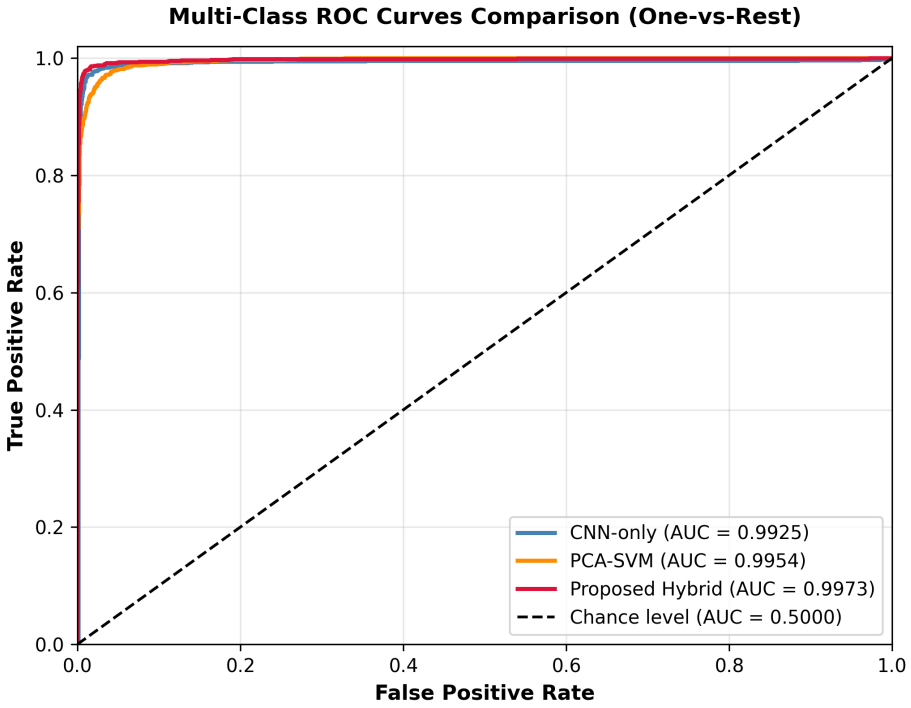



**Fig. 7**: Receiver Operating Characteristic (ROC) Curve.

**Table 8**: Ablation study on the number of PCA components.

| **PCA components (k)** | **RBF-SVM Acc (%)** | **Linear head (Logistic Regression) Acc (%)** |
|---|---|---|
| 32 | 91.11 | 85.56 |
| 64 | 92.92 | 87.36 |
| 128 | 94.44 | 89.72 |
| 256 | 94.72 | 91.39 |

(Section 4.7.1) and the comparison between the non-linear SVM and the linear head (Section 4.7.2).

### 4.7.1 Effect of PCA Dimensionality

Accuracy increases monotonically with k for both classifiers, rising from 91.11%/85.56% (RBF-SVM/logistic regression) at k = 32 to 94.72%/91.39% at k = 256, indicating that higher-dimensional PCA projections retain more discriminative information from the pooled ResNet50 features. The gain is most pronounced between k = 32 and k = 64 (+1.81 points for RBF-SVM, +1.80 points for logistic regression)

and diminishes at higher k, suggesting that most of the discriminative variance in the pooled 2048-D feature space is captured within the first 128–256 principal components.

### 4.7.2 SVM vs. Linear Head on the same CNN Features

The RBF-SVM consistently outperforms the linear (logistic regression) head at every PCA dimensionality tested (e.g., 94.72% vs. 91.39% at k = 256), indicating that the class boundaries in the pooled 2048-D feature space are not fully linearly separable and that some non-linear structure remains even after dimensionality reduction. Even so, the best classical configuration (k = 256, RBF-SVM, 94.72%) remains 2.68 percentage points below the proposed Transformer head's cross-validated mean accuracy (97.40%, Table 6), suggesting that self-attention captures discriminative structure beyond what a non-linear classifier can recover from globally pooled CNN features alone.

## 4.8 Computational Efficiency Comparison

**Table 9**: Computational complexity and performance comparison.

| **Model** | **Params (M)** | **FLOPs (G)** | **Latency (ms/img)** | **Peak Memory (MB)** |
|---|---|---|---|---|
| CNN-only (ResNet50) | 23.52 | 8.26 | 6.99 | 1214.8 |
| CNN+PCA-SVM | 23.52 | 8.19 | 7.10 | 315.0 |
| Proposed CNN+Transformer | 37.20 | 9.21 | 8.22 | 1214.8 |

Theoretical FLOPs, per-image inference latency (batch size = 1, averaged over 100 iterations following 10 warm-up runs), and peak GPU memory allocation were benchmarked on an NVIDIA T4 GPU (16 GB VRAM) under identical execution conditions. FLOPs accounting follows standard profiling conventions via the `thop` library, counting multiply-accumulate (MAC) operations as two floating-point operations (1 MAC = 2 FLOPs), inclusive of all convolutional kernels, linear projections, self-attention matrices, batch normalization, and activation functions across both the backbone and the Transformer head. This convention differs from some prior work that reports ResNet50 complexity in MACs alone (commonly cited as $\approx 4.1$ GMACs at $224 \times 224$ resolution), which is numerically equivalent to $\approx 8.2$ GFLOPs under our convention. Since the CNN-only and CNN+PCA-SVM configurations share an identical, jointly fine-tuned ResNet50 backbone (Section 3.7), their FLOPs are nearly identical, with the marginal residual difference (8.26G vs. 8.19G) arising from the negligible additional linear-algebra cost of the PCA projection step relative to the convolutional backbone. Peak memory for CNN+PCA-SVM is substantially lower (315.0 MB), because the PCA/SVM classification stage executes on CPU via scikit-learn rather than allocating GPU activation memory for a trainable classification head.

Compared to the CNN-only baseline, the proposed model incurs an overhead of 13.68M parameters (+58.2%) and only 0.95 GFLOPs (+11.5%, 9.21G vs. 8.26G),

resulting in a minor latency increase of 1.23 ms (+17.6%, 8.22 ms vs. 6.99 ms). Meanwhile, the peak memory footprint remains essentially identical to the CNN-only backbone (1214.8 MB for both configurations), as activation memory is dominated by the shared ResNet50 backbone rather than the classification head. This modest computational overhead is accompanied by a +0.69 percentage-point accuracy gain (Table 6) and significantly reduced confusion between glioma and meningioma (Section 4.5). Given the sub-10 ms per-slice latency and modest peak VRAM requirement ($\approx$ 1.2 GB), the proposed architecture remains fully compatible with real-time deployment on standard entry-level clinical workstations (Section 1).

## 4.9 Explainability Analysis

To provide clinically interpretable visual justification for model predictions, this study adopts a dual-level explainability pipeline combining Grad-CAM [40] for the local CNN branch and Attention Rollout [41] for the global Transformer branch. Grad-CAM generates a coarse localization heatmap using the gradients of the target class flowing into the final convolutional layer (layer4 of ResNet50), highlighting spatial regions that most influence local feature representations. Conversely, Attention Rollout aggregates self-attention weights across all four Transformer Encoder blocks (Section 3.4) by recursively multiplying per-layer attention matrices, mapping how the [CLS] token's classification is distributed across the 49 spatial visual tokens.

As visualized in Figure 8, the dual-level explainability framework reveals both convergent and divergent behaviors between the local convolutional and global transformer branches across the four classes, which we report transparently rather than selectively highlighting only favorable cases:

- **Focal, High-Contrast Lesions (Glioma):** Grad-CAM and Attention Rollout exhibits strong spatial convergence, both concentrating precisely on the hyperintense, ring-enhancing peripheral lesion and its necrotic core. This high degree of mutual agreement between independently derived local and global heatmaps confirms that both sub-networks are attending to the primary clinically relevant pathological site.
- **Subtle or Extra-axial Masses (Meningioma):** In this representative case, both branches localize toward the peri-ventricular and central parenchyma adjacent to the mass rather than exclusively outlining the discrete dural-tail boundary. This observation suggests that the hybrid model may partly exploit secondary topological cues (e.g., ventricular compression and midline asymmetry induced by mass effect) in conjunction with local tumor textures. We report this as a qualitative observation and caution against interpreting either heatmap as a surrogate for semantic tumor segmentation.
- **Healthy Controls (No Tumor):** Both explainability methods display diffuse, low-intensity activations without artificial focal concentration at the skull boundary or image margins, providing reassurance against gross background artifact exploitation or skull-stripping shortcut learning. While Grad-CAM displays mild lateral variation across normal cortical folds, Attention Rollout maintains broad, bilateral baseline coverage across the ventricular system.

- **Divergent Branch Activations (Pituitary Tumor):** This case illustrates an informative divergence between the two explanatory perspectives: Grad-CAM captures the focal contrast interface around the sellar floor, whereas Attention Rollout projects significant attention superiorly toward the third ventricle and suprasellar cistern. Rather than viewing this divergence as a limitation, it directly justifies the dual-XAI design: providing orthogonal local and global interpretability maps prevents clinicians from relying on a single, potentially biased attribution map and encourages critical multi-focal verification.

## 4.10 Comparison with State-of-the-Art

Table 10 compares the proposed model against the SOTA methods reviewed in Section 2 that use the same dataset family (Xception 98.73%, CAFNet 96.41%, MLHnet 99.30%, etc.).

**Table 10**: Comparison with state-of-the-art methods on the same dataset family.

| Method | Accuracy | Explainability/Compute reported |
|---|---|---|
| Disci et al. (2025a) [14] – Xception | 98.73% | No / No |
| Jayaraman et al. (2025) [15] – CAFNet | 96.41% | No / No |
| Hussain et al. (2025) [17] – EFFResNet-ViT | 99.31% | Partial (Grad-CAM + t-SNE) / No |
| Kaprekar et al. (2026) [18] – MLHnet | 99.30% | Yes (regional) / Yes |
| Napa et al. (2026) [19] – BRAIN-META | 97.10% | Partial (Grad-CAM only) / No |
| **Proposed (this study)** | 97.40% | Yes (dual) / Yes |

"Dual" explainability refers to both attention-based visualization and gradient-based saliency maps. "Regional" refers to localized region-based interpretation.

We now discuss where the proposed model stands relative to SOTA. While the proposed CNN+Transformer achieves 97.40% accuracy, it does not surpass the top reported accuracies of 99.31% (Hussain et al., EFFResNet-ViT [17]) and 99.30% (Kaprekar et al., MLHnet [18]), and is comparable to BRAIN-META's 97.10% (Napa et al. [19]). However, we explicitly state the compensating contributions that distinguish our work: unlike EFFResNet-ViT and BRAIN-META, neither of which reports cross-validation, patient-level considerations, or computational-cost measurements, our study provides all three alongside a controlled structural ablation.

- **Explainability:** Unlike most SOTA methods that lack interpretability (e.g., Disci et al., Jayaraman et al.) or report only a single explainability technique without broader evaluation context (e.g., EFFResNet-ViT, BRAIN-META), and unlike Kaprekar et al., whose regional XAI is evaluated only on tumor-positive samples, our model offers dual explainability through both attention-based visualization and gradient-based saliency maps, enabling finer-grained insights into model decision-making across all classes.
- **Computational efficiency:** Our proposed architecture reports all four computational metrics—parameters, FLOPs, latency, and peak memory (Section 4.8),

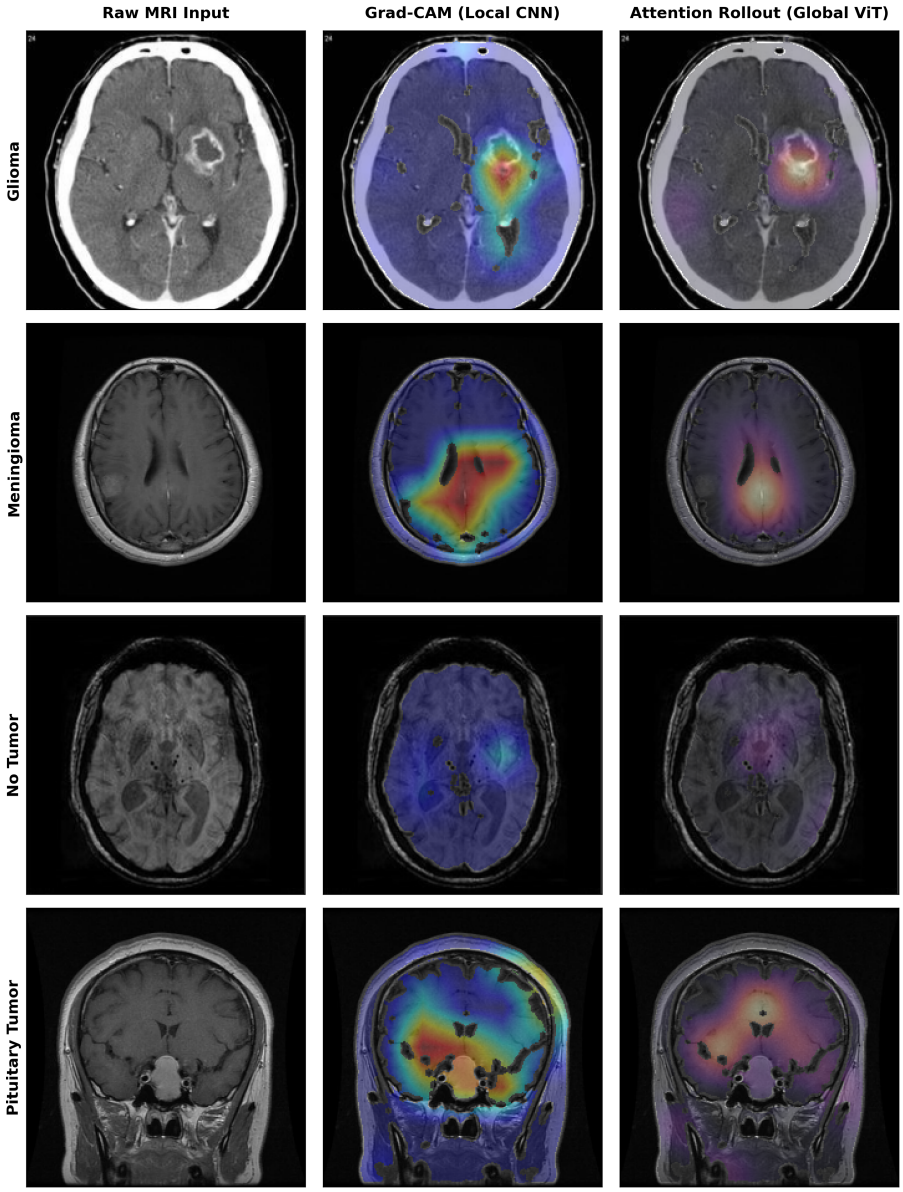



**Fig. 8**: Dual-level visual explainability across representative brain MRI classes. Column 1: Preprocessed raw MRI inputs representing glioma, meningioma, no tumor, and pituitary tumor categories. Column 2: Localized Grad-CAM heatmaps derived from the ResNet50 backbone (layer4). Column 3: Global Attention Rollout maps aggregated across all four Transformer Encoder blocks.

making the accuracy-efficiency trade-off transparent rather than assumed, in contrast to the SOTA methods above that omit this reporting entirely.

- **Patient-level leakage discussion:** Unlike most prior work reviewed in Section 2, which neither discusses nor tests for patient-level data leakage, we explicitly acknowledge this risk (Section 3.1) and report our cross-validation results with this caveat made transparent, rather than presenting accuracy figures as guaranteed generalization to unseen patients. This transparency, while not a substitute for verified patient-independent splitting, distinguishes our reporting from studies that omit the issue entirely.
- **Structural ablation:** Through systematic ablation studies (Sections 4.7–4.8), we provide a clear understanding of each component's contribution, offering valuable architectural insights that extend beyond a simple accuracy comparison.

These contributions collectively establish the proposed model's value as a robust, interpretable, and practically deployable solution, even if the raw accuracy does not set a new state-of-the-art record.

# 5 Discussions

The results reported in Section 4 indicate that the proposed sequential CNN+Transformer architecture achieves a mean cross-validation accuracy of 97.40%± 0.24%, significantly outperforming both the CNN-only baseline (96.71% ± 0.39%, paired $t$-test $p = 0.0122$) and the PCA-SVM ablation baseline (94.56% ± 0.33%, $p = 2.42 \times 10^{-5}$). Furthermore, the lower standard deviation across five folds (±0.24% vs. ±0.39%) highlights the variance-reduction effect of the multi-head self-attention mechanism, confirming that global contextual synthesis provides superior generalization across diverse MRI slice distributions.

The ablation results in Section 4.7 demonstrate that this performance advantage stems primarily from spatial self-attention rather than non-linear classification of pooled features. Even when retaining 256 PCA components with an RBF kernel, the SVM achieves at most 94.72%, leaving a 2.68 percentage-point gap compared to the Transformer head. This confirms that preserving spatial token relationships prior to classification is critical for capturing subtle topological patterns in lesion margins.

Consistent with findings in prior literature, glioma and meningioma represented the primary source of classification error due to their shared visual heterogeneity and boundary irregularities. In the representative held-out fold, the PCA-SVM baseline produced 38 combined confusion errors ($23 + 15$) and the CNN-only baseline yielded 21 errors ($20 + 1$). The proposed hybrid model reduced this confusion to 18 cases ($14 + 4$). By aggregating global anatomical cues via the Transformer head, the model more effectively distinguishes intra-axial infiltrative lesions (glioma) from extra-axial dural-based masses (meningioma).

From a deployment perspective, the proposed framework incurs an overhead of only 13.68 M parameters and 0.95 GFLOPs over the CNN backbone, resulting in a low per-image inference latency of 8.22 ms and an identical peak GPU memory

footprint of 1214.8 MB. These operational characteristics demonstrate that high-performance diagnostic support can be achieved on standard entry-level accelerators without demanding high-end enterprise infrastructure.

## 5.1 Limitations

- **Patient-level data leakage.** As discussed in Section 3.1, the dataset used in this study does not disclose patient-level identifiers. It is therefore not possible to guarantee that MRI slices from the same patient do not appear in both the training and test partitions or across different cross-validation folds. While this limitation is shared by essentially all prior work using this dataset family (Section 2), it means the reported performance should be interpreted as an upper bound on generalization to genuinely unseen patients, rather than a guaranteed estimate.
- **Single dataset, single institution.** All experiments are conducted on one publicly available dataset aggregated from Figshare, SARTAJ, and BR35H sources. No external, out-of-distribution dataset (e.g., from a different scanner vendor or acquisition protocol) was used to test cross-domain generalization. A recently released alternative, BRISC2025 [42], offers a more rigorously curated four-class brain tumor MRI collection (6,000 physician-annotated T1-weighted slices across axial, sagittal, and coronal planes) with paired segmentation masks reviewed by radiologists. While BRISC2025 was not available at the outset of this study and full re-benchmarking on it is left for future work, it represents a promising direction for external validation of the proposed architecture beyond the Nickparvar dataset family used here, and future revisions of this work will report cross-dataset generalization results on BRISC2025.
- **Small number of cross-validation folds.** The paired t-test in Section 4.4 is computed over only five-fold level accuracy values, which limits its statistical power; small differences in mean accuracy may not be reliably distinguished from noise with $n = 5$.
- **McNemar's test computed on a single fold.** The McNemar's test in Section 4.4 is computed using per-sample predictions from the final cross-validation fold only, rather than pooled predictions across all five folds. This was a deliberate simplification of the evaluation pipeline and should be considered when interpreting the reported p-values.
- **2D slice-based classification.** The proposed model classifies individual 2D MRI slices rather than full 3D volumetric scans, which may not fully capture volumetric tumor context that a radiologist would use in practice.
- **PCA-SVM baseline uses pooled rather than spatial features.** The PCA-SVM ablation (Section 3.7) is built on globally averaged pooled 2048-d feature vectors, whereas the proposed Transformer head consumes the same backbone's 49 spatial tokens directly, without pooling. Consequently, the performance gap between the two (Table 8, Section 4.7) reflects the combined benefit of self-attention and of preserving spatial structure rather than the isolated marginal contribution of attention alone. A more tightly matched ablation would apply PCA and a classical classifier directly to the flattened $49 \times 2048$ (or $7 \times 7 \times 2048$) spatial feature map; this is left for future work. Additionally, since the backbone features were jointly

fine-tuned with the Transformer head, they were not optimized independently for the PCA-SVM pipeline; a backbone fine-tuned specifically for the CNN-only/PCA-SVM configuration might perform differently. While the present ablation confirms that non-linear classification over global descriptors cannot match the Transformer head, future iterations will implement a spatial token-level pooling baseline (e.g., spatial token mean/max pooling coupled with an MLP) to completely disentangle the benefits of spatial token preservation from the specific contributions of multi-head self-attention.

- **Computational benchmarking on entry-level cloud hardware.** Inference latency and memory metrics were benchmarked on an NVIDIA T4 GPU (a standard inference accelerator) (Section 4.8, Table 9). While absolute latencies will be even lower on high-end clinical workstation GPUs (e.g., RTX 4090 or A100), our results confirm that sub-10 ms real-time inference is easily achievable even on modest, cost-effective hardware.

# 6 Conclusion

This study proposed a hybrid CNN-Transformer architecture for multi-class brain tumor MRI classification, sequentially combining a pretrained ResNet50 backbone with a lightweight Transformer Encoder head in place of a conventional fully-connected classification layer. The model was evaluated on a 7,200-image, class-balanced brain tumor MRI dataset using stratified 5-fold cross-validation with statistical significance testing and was directly compared against a PCA-SVM ablation baseline built on the same fine-tuned CNN backbone to help isolate the Transformer's contribution.

The proposed model achieved a mean cross-validation accuracy of $97.40\% \pm 0.24\%$, significantly outperforming both the CNN-only baseline ($96.71\% \pm 0.39\%$) and the CNN+PCA-SVM ablation ($94.56\% \pm 0.33\%$).

Beyond raw accuracy, this study makes methodological contributions that address gaps identified in prior hybrid CNN-ViT literature (Section 2): rigorous stratified cross-validation with statistical testing, explicit discussion of patient-level data leakage risk, a controlled structural ablation against a classical PCA-SVM baseline, deployment-oriented computational efficiency benchmarking, and dual-level visual explainability combining Grad-CAM and Attention Rollout.

Future work will focus on addressing the limitations identified in Section 5.1. In particular, we plan to externally validate the proposed architecture on BRISC2025 [42], a more rigorously curated four-class brain tumor MRI collection with radiologist-reviewed segmentation masks, to assess cross-dataset and cross-scanner generalization beyond the Nickparvar dataset family used here. We also intend to extend the framework to volumetric 3D classification, replacing 2D slice-level inputs with full MRI volumes, and to obtain patient-level identifiers where possible to eliminate the risk of patient-level data leakage across folds. Beyond these evaluation-focused extensions, we plan to test the limits of the proposed sequential design itself by benchmarking it against alternative fusion mechanisms, most notably cross-attention fusion, as in CAFNet [15] under the same cross-validation, ablation, and computational-efficiency

protocol introduced in this study. Such a comparison would clarify whether the additional architectural complexity of cross-attention fusion yields further accuracy gains that justify its cost, or whether the lighter sequential design proposed here remains a preferable accuracy-efficiency trade-off for resource-constrained clinical deployment.

# Declarations

**Author contributions** H.M.L. conceived the study, designed the hybrid architecture, implemented the model and experiments, performed the statistical analysis, and drafted the manuscript. V.T.N. contributed to the experimental design, reviewed and revised the manuscript. Both authors read and approved the final manuscript.

**Funding declaration** No funding was received for this study.

**Clinical trial number** Not applicable.

**Data availability** This study utilized the publicly available "Brain Tumor MRI Dataset" dataset from Kaggle, which is openly accessible for research purposes. The dataset can be accessed at the following link: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset.

**Competing interests** The authors have not disclosed any competing interests.

**Ethical approval** This study involved no human participants, animal subjects, or personal data.

**Consent to participate** Not applicable.

**Consent to publish** Not applicable.

# References

[1] Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., Jemal, A.: Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **74**(3), 229–263 (2024)

[2] Kerkhof, M., Hagenbeek, R.E., Kallen, B.F., Nijeholt, G., Dirven, L., Taphoorn, M., Vos, M.J.: Interobserver variability in the radiological assessment of magnetic resonance imaging (mri) including perfusion mri in glioblastoma multiforme. European Journal of Neurology **23**(10), 1528–1533 (2016)

[3] Amin, J., Sharif, M., Haldorai, A., Yasmin, M., Nayak, R.S.: Brain tumor detection and classification using machine learning: a comprehensive survey. Complex & intelligent systems **8**(4), 3161–3183 (2022)

[4] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is

worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[7] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9992–10002 (2021). Ieee

[8] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357 (2021). PMLR

[9] Nickparvar, M.: Brain Tumor MRI Dataset. Kaggle (2026). https://doi.org/10.34740/KAGGLE/DSV/14832123 . https://www.kaggle.com/dsv/14832123

[10] Azzahra, T.S., Cerelia, J.J., Nugraha, F.A.L., Pravitasari, A.A.: Mri-based brain tumor classification using inception resnet v2. Enthusiastic: International Journal of Applied Statistics and Data Science, 163–175 (2023)

[11] Vure, R.B., Pappala, L.K.: Enhanced brain tumor classification framework using deep learning. Scientific Reports **15**(1), 35814 (2025)

[12] Jahan, M.K., Shafi, A.A., Rusho, M.A., Hussain, M.S., Haque Dhrubo, A.F.: Cereviannet: parameter efficient multi-class brain tumor classification using custom lightweight cnn. Frontiers in Medicine **12**, 1664673 (2025)

[13] Hassan, E., Ghadiri, H.: Advancing brain tumor classification: A robust framework using efficientnetv2 transfer learning and statistical analysis. Computers in Biology and Medicine **185**, 109542 (2025)

[14] Disci, R., Gurcan, F., Soylu, A.: Advanced brain tumor classification in mr images using transfer learning and pre-trained deep cnn models. Cancers **17**(1), 121 (2025)

[15] Jayaraman, G., Meganathan, S., Shah, S.S.M., Anuradha, M., Sundararajan, R.K., Rajakumar, R.: A hybrid cnn–vit framework with cross-attention fusion and data augmentation for robust brain tumor classification. Scientific Reports **15**(1), 45769 (2025)

[16] Yin, H., Vahdat, A., Alvarez, J.M., Mallya, A., Kautz, J., Molchanov, P.: A-vit: Adaptive tokens for efficient vision transformer. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10799–10808 (2022). https://doi.org/10.1109/CVPR52688.2022.01054

[17] Hussain, T., Shouno, H., Hussain, A., Hussain, D., Ismail, M., Mir, T.H., Hsu, F.R., Alam, T., Akhy, S.A.: Effresnet-vit: A fusion-based convolutional and vision transformer model for explainable medical image classification. Ieee Access **13**, 54040–54068 (2025)

[18] Kaprekar, A., Gudigar, A., Raghavendra, U., Inamdar, M.A., Salvi, M., Priyanka, Acharya, U.: A multi-level attention cnn-transformer based framework for the detection of brain tumor using regional dual-score explainability. Scientific Reports (2026)

[19] Napa, K.K., Murugan, S., Murugan, J.S., Jayanthi, A.: Brain-meta: A reproducible cnn–vision transformer meta-ensemble pipeline for explainable brain tumor classification. MethodsX **16**, 103769 (2026)

[20] Srinivas, V.R., Parvathi, R.: Explainable ai-driven mri-based brain tumor classification: a novel deep learning approach. Frontiers in Artificial Intelligence **8**, 1700214 (2025)

[21] Yagis, E., Atnafu, S.W., Herrera, A., Marzi, C., Scheda, R., Giannelli, M., Tessa, C., Citi, L., Diciotti, S.: Effect of data leakage in brain mri classification using 2d convolutional neural networks. Scientific reports **11**(1), 22544 (2021)

[22] Reza, A.M.: Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. Journal of VLSI signal processing systems for signal, image and video technology **38**(1), 35–44 (2004)

[23] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). Ieee

[24] Simard, P.Y., Steinkraus, D., Platt, J.C., *et al.*: Best practices for convolutional neural networks applied to visual document analysis. In: Icdar, vol. 3 (2003). Edinburgh

[25] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)

[26] Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)

[27] Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)

[28] Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)

[29] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

[30] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)

[31] Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)

[32] Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)

[33] Jolliffe, I.T.: Principal Component Analysis. Springer, New York (2002)

[34] Kohavi, R., *et al.*: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai, vol. 14, pp. 1137–1145 (1995). Montreal, Canada

[35] Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation **10**(7), 1895–1923 (1998)

[36] McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947)

[37] Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. routledge, New York (2013)

[38] Powers, D.M.: Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061 (2020)

[39] Fawcett, T.: An introduction to roc analysis. Pattern recognition letters **27**(8), 861–874 (2006)

[40] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)

[41] Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4190–4197 (2020)

[42] Fateh, A., Rezvani, Y., Moayedi, S., Rezvani, S., Fateh, F., Fateh, M., Abolghasemi, V.: Brisc: Annotated dataset for brain tumor segmentation and classification. Scientific Data **13**(1), 361 (2026)