

DEEP LEARNING

AI-Based Hybrid Deep Learning Framework for Accurate Brain Tumor Classification from MRI Images Using CNN and Vision Transformer

Sachin U. Kadam, Brijendra Gupta

Department of Information Technology, Siddhant College of Engineering (affiliated to Savitribai Phule Pune University),
Sudumbare, Pune, Maharashtra, India

Published October 2026

Abstract

Brain tumors remain among the most life-threatening neurological disorders, and their outcome depends heavily on how early and how accurately they are diagnosed. Magnetic Resonance Imaging (MRI) is the modality of choice for tumor evaluation, yet manual reading of MRI scans is slow and subject to inter-observer disagreement. This paper proposes NeuroFusionNet, a lightweight hybrid deep learning framework that fuses a MobileNetV2-based Convolutional Neural Network (CNN) branch with a Vision Transformer (ViT) branch through a Squeeze-and-Excitation (SE) gated fusion mechanism, allowing the model to combine fine-grained local texture cues with long-range global context in a single forward pass. The framework further incorporates a U-Net-based segmentation head to delineate tumor boundaries, a five-method Explainable AI (XAI) suite for transparent decision support, and a Large Language Model (LLM) module that converts model outputs into structured, human-readable clinical reports. Evaluated on the BRISC 2025 benchmark, which contains expert-annotated MRI scans spanning glioma, meningioma, pituitary tumor, and healthy classes, the proposed model achieves 96.2% test accuracy while using only 7.8 million parameters, indicating that competitive diagnostic accuracy can be obtained without the computational burden typical of large-scale transformer or ensemble models. The results suggest that NeuroFusionNet is a practical candidate for deployment in resource-constrained clinical settings that still require interpretable, end-to-end decision support.

Keywords: Brain Tumor; Deep Learning; MRI; CNN; Vision Transformer; Hybrid Model; Feature Fusion; Segmentation; Explainable AI; Knowledge Distillation; Medical Imaging

1 Introduction

Brain tumors are among the most aggressive and challenging neurological conditions, arising from abnormal and uncontrolled cell growth within brain tissue. This abnormal proliferation disrupts normal neurological function and, depending on tumor type and grade, can lead to severe cognitive, motor, and sensory impairments. Early and accurate diagnosis is therefore critical, as it directly influences treatment planning, surgical guidance, and ultimately patient survival rates. Among available imaging modalities, Magnetic Resonance Imaging (MRI) remains the preferred choice for tumor detection, owing to its superior soft-tissue contrast, high spatial resolution, and non-invasive nature. Nevertheless, manual interpretation of MRI scans presents significant limitations: the process is time-consuming, and diagnostic outcomes may vary between radiologists due to subjective judgment and inter-observer variability.

In recent years, deep learning has fundamentally reshaped the landscape of medical image analysis. Unlike conventional machine learning approaches, which depend on manually engineered features,

deep learning models automatically learn hierarchical representations directly from raw image data. This capability has proven particularly effective for brain tumor analysis, supporting three major tasks: tumor detection, classification of tumor type and grade, and segmentation of affected regions.

Early studies in this domain relied primarily on Convolutional Neural Networks (CNNs) for tumor classification and delineation, consistently outperforming traditional machine learning techniques such as Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN) in terms of accuracy. The subsequent availability of powerful computing hardware and openly accessible benchmark datasets—most notably the Multimodal Brain Tumor Segmentation Challenge (BRATS)—accelerated the adoption of advanced architectures such as U-Net, ResNet, and EfficientNet, which have since become standard baselines for tumor analysis.

More recently, Transformer-based models have emerged as a promising alternative. By capturing long-range dependencies and global contextual information within MRI scans, architectures such as hierarchical Vision Transformers have demonstrated greater robustness under challenging imaging conditions and, in several cases, improved interpretability compared to conventional CNNs. In parallel, Explainable Artificial Intelligence (XAI) techniques—including attention visualization and Bayesian uncertainty estimation—have been integrated into tumor detection pipelines to enhance transparency and foster trust among clinicians.

Hybrid frameworks also continue to evolve. Lightweight models that combine CNNs with Transformers, often enhanced through Knowledge Distillation (KD), pave the way for computationally efficient solutions suitable for real-world clinical environments. KD enables compact networks to approximate the performance of larger models with minimal accuracy loss, making deep learning feasible even in resource-constrained healthcare settings.

Malignant tumors of the brain and central nervous system (CNS) represent a significant global health burden, accounting for a small but disproportionately fatal share of all cancer diagnoses [1]. Gliomas, which originate from glial cells, constitute the majority of malignant brain tumors, with Glioblastoma Multiforme (GBM) — classified as WHO Grade IV — being the most aggressive and prevalent subtype [2].

Despite advances in surgical resection, radiotherapy, and chemotherapy, the long-term survival of GBM patients remains poor [3]. These figures underscore the urgent clinical need for early, precise, and reliable diagnostic tools capable of supporting timely therapeutic intervention.

Motivated by this need, the present work proposes NeuroFusionNet, a hybrid CNN-Transformer architecture designed specifically to balance diagnostic accuracy, computational efficiency, and clinical interpretability. Rather than choosing between a CNN's local sensitivity and a Transformer's global reasoning, NeuroFusionNet fuses both representations adaptively through a Squeeze-and-Excitation gated mechanism, and pairs the resulting classifier with a segmentation module, an explainability suite, and an automated report generator, so that the system can support a clinician from detection through to a written diagnostic summary.

2 Related work

Research on automated brain tumor analysis has progressed broadly along three fronts — classification, segmentation, and hybrid or resource-efficient frameworks — and Table 1 summarizes representative contributions from each of these fronts along with the datasets and metrics used to evaluate them.

Table 1. Representative deep learning approaches for brain tumor analysis.

Category	Study	Core Approach	Dataset	Reported Result
Classification	Shahin et al. [4]	Fine-tuned ResNet-34 with transfer learning	MRI (BRATS)	99.66% accuracy
Classification	Asif et al. [5]	Inception + Xception ensemble	BRATS 2021	99.92% accuracy
Classification	Sankari et al. [6]	Hierarchical Multiscale Vision Transformer (HMVT)	Private MRI set	98.7% accuracy
Classification	Ghasemi et al. [7]	Hybrid CNN–MLP, multimodal MRI+CT fusion	MRI + CT	97.5% accuracy
Segmentation	Lakshmi et al. [8]	MEDU-Net with Bayesian classifier (XAISS–BMLBT)	BRATS 2020	0.96 Dice
Segmentation	Rajendran et al. [9]	3D CNN–U-Net hybrid for glioma subregions	BRATS 2021	0.94 Dice
Segmentation	Pereira et al. [10]	Deep CNN with small convolution kernels	BRATS 2013–15	0.88 Dice
Hybrid / Efficient	Benfatto et al. [11]	Epigenetics-linked explainable AI framework	Epigenetic MRI set	Qualitative XAI
Hybrid / Efficient	Hinton et al. [12]	Teacher–student Knowledge Distillation (foundational)	Conceptual	—
Hybrid / Efficient	Feng et al. [13]	Lightweight KD for tumor classification	MRI (BRATS)	95% retained, 60% smaller

2.1 Convolutional and transfer-learning classifiers

Convolutional networks such as ResNet, Inception, and Xception have consistently produced strong classification results once adapted to MRI data through transfer learning [4, 5]. Ensembling multiple convolutional backbones tends to reduce overfitting and stabilizes predictions further, though at the cost of a larger overall parameter count and slower inference, which limits practicality on constrained hardware.

2.2 Vision-Transformer and multimodal approaches

Vision Transformer variants, including hierarchical multiscale designs, capture long-range spatial relationships that pure convolutional stacks tend to miss, and have reported classification accuracy comparable to, or exceeding, CNN baselines on several MRI datasets [6]. Multimodal frameworks that jointly process MRI and CT data through combined CNN–MLP pipelines have likewise shown improved cross-modality generalization [7], indicating that fusing complementary feature types is a productive direction — a principle that directly motivates the fusion strategy adopted in this work.

2.3 Segmentation architectures

For pixel-level tumor delineation, U-Net and its derivatives remain the dominant family of architectures because their encoder–decoder structure with skip connections preserves both semantic and spatial detail. Explainable variants that pair a modified U-Net encoder with a Bayesian classification head have achieved Dice scores as high as 0.96 while retaining some model transparency [8], and 3D extensions that reason across MRI slices have improved delineation of glioma subregions relative to plain 2D U-Net baselines [9]. Earlier CNN-based segmentation work using smaller convolution ker-

nels established that boundary detection can be improved with deeper stacks, although at the expense of interpretability and robustness to class imbalance [10].

2.4 Explainability, privacy, and model compression

A parallel line of work argues that accuracy alone is insufficient for clinical adoption. Frameworks that connect deep feature activations to interpretable biological or radiological markers aim to make model decisions traceable for clinicians [11]. Separately, Knowledge Distillation, first formalized as a teacher–student compression strategy [12], has since been applied specifically to brain tumor classification, where distilled student networks have been shown to retain more than 95% of a larger teacher model's accuracy while cutting parameter count by roughly 60% [13]. These two threads — explainability and lightweight design — are treated in the present work not as separate add-ons but as core design requirements of the proposed system.

2.5 Research gap

Taken together, the literature shows a clear trade-off: high-accuracy ensemble and pure-Transformer models tend to be computationally heavy, while lightweight CNN classifiers alone under-utilize global context. Few existing frameworks combine an adaptive CNN–Transformer fusion mechanism with a segmentation head, a multi-method explainability suite, and automated report generation within a single lightweight pipeline. This gap motivates the design of NeuroFusionNet, described in Section III.

3 Proposed methodology

The proposed system, NeuroFusionNet, is designed as an end-to-end pipeline that takes a raw MRI slice as input and produces four outputs: a tumor class label, a pixel-level segmentation mask, a set of visual explanations, and a natural-language clinical summary. The overall pipeline is organized into four stages, described below.

3.1 Dual-branch feature extraction

The input MRI slice is processed in parallel by two feature-extraction branches. The first branch is a MobileNetV2 backbone [15], chosen for its inverted-residual, depth-wise separable convolutions that extract local texture and edge information at low computational cost. The second branch is a Vision Transformer encoder [14], which splits the image into fixed-size patches, embeds them linearly, and applies multi-head self-attention across the resulting patch-token sequence to model long-range spatial dependencies that are otherwise invisible to a convolutional receptive field.

3.2 Squeeze-and-Excitation gated fusion

Rather than simply concatenating the two feature streams, NeuroFusionNet fuses them through a Squeeze-and-Excitation (SE) gated module [16]. The CNN and ViT feature maps are first aligned to a common channel dimension; a global-average-pooling “squeeze” step then summarizes each channel, and a small gating network learns per-channel weights that adaptively emphasize the more informative branch for a given input. This soft gating allows the network to rely more heavily on local texture cues for well-localized tumors and on global context for diffuse or boundary-ambiguous cases, rather than fixing the CNN/ViT contribution in advance.

3.3 Classification and segmentation heads

The fused feature representation is passed to a fully connected classification head that outputs probabilities over four classes — glioma, meningioma, pituitary tumor, and no tumor — matching the label set of the BRISC 2025 dataset. In parallel, an auxiliary U-Net-style decoder [17] reuses the CNN branch's encoder features through skip connections to reconstruct a pixel-level segmentation mask of the tumor region, so that classification and localization are learned jointly rather than as two disconnected models.

3.4 Explainability suite and automated reporting

To keep the model's reasoning transparent, NeuroFusionNet integrates a five-method XAI suite covering Grad-CAM and Grad-CAM++ saliency mapping [18], SHAP and LIME feature-attribution methods, and attention-rollout visualization drawn directly from the ViT branch's self-attention weights. The combined visual evidence, together with the predicted class, confidence score, and segmentation mask, is passed as structured input to a Large Language Model, which is prompted to generate a concise, templated clinical report describing the predicted tumor type, its approximate location and extent, and the visual evidence supporting the prediction — intended as a decision-support aid rather than a replacement for radiologist review.

4 Dataset and preprocessing

The proposed framework is trained and evaluated on BRISC 2025, an expert-annotated, contrast-enhanced T1-weighted MRI dataset released for brain tumor segmentation and classification [19]. The dataset contains approximately 6,000 MRI slices spanning glioma, meningioma, pituitary tumor, and no-tumor cases, collected across axial, coronal, and sagittal planes, with pixel-level segmentation masks reviewed by radiologists. The data is used with its published stratified split of roughly 5,000 training and 1,000 test images.

Prior to training, each slice undergoes intensity normalization and resizing to a fixed input resolution compatible with both the MobileNetV2 and ViT branches. Data augmentation — random rotation, horizontal flipping, and mild contrast jitter — is applied on the training split to reduce overfitting given the moderate dataset size. Segmentation masks are resized with nearest-neighbor interpolation to preserve label boundaries, and both branches share the same preprocessing pipeline so that classification and segmentation targets remain spatially aligned.

5 Experimental results and discussion

On the BRISC 2025 test split, NeuroFusionNet achieves an overall classification accuracy of 96.2% while using only 7.8 million parameters, a substantially smaller footprint than typical Transformer-only or multi-network ensemble designs. Table 2 situates this result against representative figures reported in the literature for related architecture families, noting that these external figures come from different datasets and are included only as context rather than a like-for-like comparison.

Table 2. NeuroFusionNet in the context of representative literature results.

Model / Family	Reported Accuracy	Approx. Parameters	Basis
Fine-tuned ResNet-34 [4]	99.66%	≈21 M	BRATS MRI
Inception + Xception ensemble [5]	99.92%	>40 M (combined)	BRATS 2021

Model / Family	Reported Accuracy	Approx. Parameters	Basis
Hierarchical Multiscale ViT [6]	98.7%	≈ 25–30 M	Private MRI set
NeuroFusionNet (proposed)	96.2%	7.8 M	BRISC 2025

The comparison indicates that the larger ResNet, ensemble, and Transformer models reach marginally higher accuracy on their respective benchmarks, but each does so with several times the parameter count of NeuroFusionNet. The proposed model's SE-gated fusion appears to recover most of the benefit of combining CNN and Transformer representations without requiring either branch to be scaled up, which is consistent with the broader Knowledge-Distillation literature showing that compact models can retain the large majority of a bigger model's accuracy [12, 13].

Qualitatively, misclassifications are concentrated among meningioma and no-tumor cases that show subtle contrast differences, which is a pattern consistent with reports elsewhere in the literature for similar four-class MRI settings. The auxiliary segmentation head produces boundary overlaps in line with the 0.88–0.96 Dice range reported for U-Net-family models on comparable tumor segmentation tasks [8–10], suggesting that the shared encoder does not sacrifice localization quality for classification performance. The attention-rollout and Grad-CAM outputs from the XAI suite generally highlight the tumor region itself rather than surrounding tissue, supporting the intended goal of producing explanations that a clinician can visually verify against the segmentation mask.

An informal ablation — removing the ViT branch and relying on the MobileNetV2 branch alone, or removing the CNN branch and relying on the ViT branch alone — is expected, consistent with prior hybrid-model findings [6, 7], to underperform the fused configuration, since each branch alone omits either the fine local texture cues or the long-range context that the other supplies. A full quantitative ablation across fusion variants is left as near-term future work.

6 Challenges and limitations

Despite the encouraging results, several limitations should be acknowledged. First, BRISC 2025, like most public brain-tumor datasets, is modest in scale relative to natural-image benchmarks, so generalization to scanners, protocols, and populations outside the dataset's source institutions remains untested. Second, class imbalance between tumor and non-tumor pixels can still bias the segmentation head toward the majority class despite augmentation. Third, while the five-method XAI suite improves transparency, it does not certify clinical correctness — saliency and attention maps indicate where the model looked, not that the model reasoned correctly. Fourth, the LLM-based reporting module depends on the quality of the upstream classification and segmentation outputs, so any upstream error can propagate into a fluent but incorrect textual report if not flagged for clinician review. Finally, deployment on genuinely low-resource hardware would benefit from further compression, for instance through the pruning and Knowledge Distillation techniques discussed in the related work, which were not applied to the present model.

7 Future scope

Future work will focus on five directions. First, validating NeuroFusionNet across multiple public and institutional datasets (e.g., BRATS, TCIA, Figshare) to test cross-domain generalization more rigorously. Second, applying Knowledge Distillation to compress the current model further for edge deployment in low-resource clinical settings. Third, extending the fusion mechanism to a federated

learning setting, allowing multiple hospitals to jointly improve the model without sharing raw patient scans. Fourth, expanding the multimodal input to include CT scans, genetic markers, or patient history alongside MRI, following the direction suggested by recent multimodal fusion studies [7]. Fifth, conducting a structured clinician evaluation of the LLM-generated reports to quantify their usefulness and safety in an actual radiology workflow before any real-world pilot deployment.

8 Conclusion

This paper presented NeuroFusionNet, a lightweight hybrid deep learning framework that fuses MobileNetV2 and Vision Transformer feature representations through a Squeeze-and-Excitation gated mechanism for brain tumor classification from MRI images. Complementary U-Net segmentation, a five-method explainability suite, and an LLM-based clinical reporting module extend the framework beyond a simple classifier into an end-to-end decision-support pipeline. On the BRISC 2025 dataset, the proposed model reaches 96.2% test accuracy with only 7.8 million parameters, indicating a favorable balance between diagnostic accuracy, model size, and interpretability compared with heavier CNN, ensemble, and Transformer-only baselines reported in the literature. Continued validation across additional datasets, model compression for edge deployment, and structured clinical evaluation of the generated reports represent the natural next steps toward real-world adoption of this framework.

References

- [1] Ostrom QT, et al. *CBTRUS statistical report: Primary brain and other central nervous system tumors diagnosed in the United States*. Neuro-Oncology. Central Brain Tumor Registry of the United States [annual report].
- [2] Louis DN, et al. *The WHO classification of tumors of the central nervous system*. Acta Neuropathologica / WHO Classification of Tumours series.
- [3] Stupp R, et al. *Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma*. New England Journal of Medicine.
- [4] Shahin A. Fine-tuned ResNet-34 model for MRI brain tumor classification. *Scientific Reports*. 2025;15(3):1–10.
- [5] Asif RN, et al. Ensemble deep learning framework for robust brain tumor classification in MRI. *Scientific Reports*. 2025;15(2):1–9.
- [6] Sankari C, et al. Hierarchical multiscale vision transformer model for accurate detection and classification of brain tumors in MRI-based imaging. *Scientific Reports*. 2025;15:1–12.
- [7] Ghasemi R, Hosseini A, Zare S. Hybrid CNN-MLP architecture for multimodal brain tumor classification. *Scientific Reports*. 2025;15(4):1–10.
- [8] Lakshmi K, Rajesh R, Nagarajan P. *Explainable artificial intelligence with U-Net-based segmentation and Bayesian machine learning for classification of brain tumors using MRI images*. Scientific Reports. 2025;15.
- [9] Rajendran S, et al. Automated brain tumor segmentation using 3D CNN-U-Net hybrid architecture. *Scientific Reports*. 2023;14:10582–10594.
- [10] Pereira S, et al. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans. Med. Imaging*. 2016;35(5):1240–1251.
- [11] Benfatto S, et al. *Integrating epigenetics and explainable deep learning in tumor analysis*. Nature Communications. 2025;16(3).
- [12] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. *arXiv:1503.02531 [Preprint]*. 2015.
- [13] Feng Y, Chen J, Zhang H. Lightweight brain tumor classification via knowledge distillation. *Biomedical Signal Processing and Control*. 2023;85:104967.
- [14] Dosovitskiy A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: *Proc. ICLR*. 2021.

-
- [15] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted residuals and linear bottlenecks. In: *Proc. IEEE CVPR*. 2018.
 - [16] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: *Proc. IEEE CVPR*. 2018.
 - [17] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: *Proc. MICCAI*. 2015.
 - [18] Selvaraju RR, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: *Proc. IEEE ICCV*. 2017.
 - [19] Fateh A, et al. BRISC: Annotated dataset for brain tumor segmentation and classification. *arXiv:2506.14318* [Preprint]. 2025.