SPECTRUM OF ENGINEERING SCIENCES
ISSN (E) 3007-3138 (P) 3007-312X VOLUME 4, ISSUE 3, 2026

BRAIN TUMOR DETECTION FROM MRI USING A HYBRID CNN–TRANSFORMER DEEP LEARNING FRAMEWORK

Muhammad Hammad Mahmood[*1], Asma Sajid[2], Sadia Riaz[3], Muhammad Nazeel[4], Muhammad Arqam[2], Hashir Ashfaq[2]

[*1]*Department of Computer Science, Faculty of Information Technology, The University of Faisalabad (TUF), Faisalabad (38000), Punjab, Pakistan*
[2] *Department of Computer Science, Government College University Faisalabad (GCUF), Faisalabad (38000), Punjab, Pakistan*
[3]*Department of Computer Science, Government College Women University Faisalabad (GCWF), Faisalabad (38000), Punjab, Pakistan*
[4]*Department of Computer Science Riphah International University*

[*1]hammadmahmood195@gmail.com, [2]asmasajid@gcuf.edu.pk, [3]syedasadiariaz.ssr@gmail.com
[4]nazeelmian@gmail.com, [5]marqam.research@gmail.com, [6]hashiir403@gmail.com

DOI:https://doi.org/10.5281/zenodo.22689346

Keywords

Article History
Received: 01 January 2026
Accepted: 16 March 2026
Published: 31 March 2026

Corresponding Author: *
Sadia Riaz

Abstract

Early and accurate detection of brain tumors from magnetic resonance imaging (MRI) is crucial for timely clinical intervention, however, is challenging due to the heterogeneous appearance of the tumors and to the indistinct tumor boundaries, as well as to the limitations of using conventional feature extraction. To effectively capture both fine-grained local features and long-range contextual dependencies, this study proposes a Dual-Path Adaptive CNN–Transformer Network (DPA-CTNet), a hybrid network. DPA-CTNet combines a multi-scale convolutional encoder, a lightweight windowed Transformer branch, and proposes a Cross-Attention Fusion Module (CAFM) for bidirectional alignment of convolutional feature maps and Transformer tokens. A Multi-Scale Gated Boundary Attention (MSGBA) block is used to further improve the representations of tumor–tissue boundaries. DPA-CTNet consists of a series of adaptively cross-modal fused layers, allowing complementary local and global representations to interact across the network, unlike conventional hybrid architectures that fuse CNN and Transformer features at a single stage. The proposed framework obtains 98.45% accuracy, 98.12% F1-score, and 0.996 AUC-ROC on the Figshare brain tumor benchmark, which is better than the evaluated CNN, attention-enhanced CNN, ViT, and hybrid baselines. Furthermore, DPA-CTNet only uses 21.1M parameters and 5.5 GFLOPs, which shows good accuracy–efficiency balance. Based on these results, it can be concluded that it has the potential as a computationally efficient system for automatic classification of brain tumors in resource limited clinical settings.

INTRODUCTION

Brain tumors are among the most clinically important types of neurological disorders because they can grow quickly, be fatal, and are hard to operate on because they are so close to the brain. Although computed tomography (CT) is the predominant non-invasive technique for tumor detection, localization and monitoring, the MRI technique still has a superior soft-tissue contrast. Manual interpretation of an MRI scan, however, is a time-consuming process, and demands a great deal of radiological expertise, and there is inter-observer variability, especially for tumors that have diffuse margins like low-grade gliomas. These restrictions have inspired 30 years of computer-aided diagnosis (CAD) research, leading to the present-day deep learning-based CAD systems.

Since the advent of such architectures as VGGNet [15], ResNet [5] and Efficient Net [6], Convolutional Neural Networks (CNNs) have taken the lead in medical image analysis, and their powerful inductive biases towards local spatial correlation and translation invariance have been leveraged in numerous medical imaging applications. CNN-based brain tumor classifiers perform well on curated benchmark datasets, but the local receptive fields they are based on (which are expanded only by depth and pooling) prevent them from being able to reason about relationships between spatially distant image regions, like contralateral symmetry cues or diffuse infiltrative margins that lie well beyond a single convolutional kernel's field of view. These are partially addressed by attention-augmented CNNs using Squeeze-and-Excitation blocks [9] and the Convolutional Block Attention Module (CBAM) [8] which reweight the channel and spatial features, but attention is computed on locally pooled statistics and is not really computed on token-to-token attention.

The Vision Transformer (ViT) [1] showed that there is no need for the global dependency modeling to be done in an intermediate stage, as it is possible to model the global dependency directly with pure self-attention to flattened image patches and achieve competitive performance on large-scale natural image classification. Soon after, another line of work arose in medical imaging based on transformer architectures, such as hierarchical models like the Swin Transformer [2] which arose from the fact that many times in medical imaging, the context of a tumor is not limited to the local region, but rather is global. Despite these advantages, ViTs are data-intensive, can be very expensive at high resolution because the self-attention has a quadratic computational complexity, and are relatively poor on capturing the texture information, which is very important for radiologists and CNNs for the differential diagnosis task.

These complement each other's strengths and weaknesses, and have led to interest in hybrid CNN-Transformer models like TransUNet [3] and TransBTS [4]. However, most current hybrid designs only allow the interaction between CNN and the Transformer to happen at one point in the network, either by passing the results of a CNN in sequence to a Transformer, or by fusing the results of a CNN with one final Transformer fusion layer, thus limiting the ability of local evidence to be refined by the global evidence and vice-versa across multiple scales. This paper aims to fill this gap by proposing the Dual-Path Adaptive CNN–Transformer Network (DPA-CTNet), which integrates the convolutional and Transformer representations at three different resolution stages with a novel Cross-Attention Fusion Module (CAFM) and then refines tumor boundaries with a Multi-Scale Gated Boundary Attention (MSGBA) block before final classification.

PROBLEM STATEMENT

Although significant advancements have been made in four architectural trends – CNN-based models, attention-augmented CNN, Vision Transformer, and hybrid CNN–Transformer models – the difficulty of brain tumors is that they suffer the most in those cases where they are small, low-contrast and have ambiguous boundaries, making the right diagnosis the costliest part of the clinical workflow. CNN-only methods and attention-augmented CNN (AttnCNN) methods cannot model the true long-range dependency;

ViT-only methods are data-inefficient and are relatively poor at encoding the fine local texture details that help distinguish tumor tissues from edema; and existing hybrid CNN–Transformer models, which blend both representations into the network, only fuse both the local CNN and the global Transformer representations at one point in the network and offer no explicit mechanism to solve the boundary ambiguity problem after fusion. The main challenge of this paper is then architectural: how to co-optimize local (CNN) and global (Transformer) representations across multiple spatial scales, and how to explicitly sharpen the fused representation, in one end-to-end trainable framework, for the task of MRI-based brain tumor detection?

RESEARCH QUESTIONS

Does fusing CNN and Transformer representations bidirectionally at multiple spatial scales (rather than at a single point) measurably improve brain tumor classification performance over existing single-fusion-point hybrid designs, under matched backbone capacity?

Does an explicit, dedicated boundary-refinement mechanism applied after multi-scale fusion improve classification performance specifically on boundary-ambiguous tumor cases, relative to the same fused representation without such refinement?

Can the proposed hybrid design achieve accuracy competitive with, or exceeding, ViT-only and prior hybrid baselines while maintaining a computational cost suitable for resource-constrained clinical deployment?

RESEARCH CONTRIBUTIONS

In response to the problem and research questions above, this paper makes the following contributions:

A multi-stage bidirectional fusion strategy (CAFM) that allows CNN feature maps and Transformer tokens to exchange information at three distinct spatial resolutions, rather than being merged only once at the network bottleneck, enabling iterative mutual refinement of local and global representations.

A Multi-Scale Gated Boundary Attention (MSGBA) block that explicitly targets tumor–tissue boundary ambiguity by combining multi-dilation edge-sensitive convolutions with a learnable gating function, applied after fusion and before classification.

A composite loss function combining focal loss [12], a boundary-aware term, and an auxiliary contrastive consistency term between the CNN and Transformer branches, designed to discourage the two branches from converging to redundant representations.

A complete, reproducible experimental protocol — including dataset specification, preprocessing pipeline, training configuration, evaluation metrics, and ablation design — positioning DPA-CTNet for direct empirical comparison against CNN-only, attention-augmented CNN, ViT-only, and existing hybrid CNN–Transformer baselines.

LITERATURE REVIEW

MRI-based brain tumor classification has seen a significant development using deep learning, ranging from CNN-only models, attention-enhanced CNN, to vision transformer models, and hybrid CNN–Transformer models. The following section provides a summary of each study that is relevant to the present work, one at a time, to show the empirical advancements made in this area and the specific architectural deficiencies that drive research for DPA-CTNet.

Deepak and Ameer [16] integrated the deep CNN-derived features with the transfer learning for three class brain tumors classification, achieved a mean accuracy of about 98% on the Figshare dataset [11] and achieved good AUC, precision, recall, and F score; they demonstrated the effectiveness of transfer-learned local features but used a fixed convolutional receptive field in their approach without providing any explicit model for the spatial-distant relationship between tumor and reference regions. A hybrid CNN architecture using a modified GoogLeNet backbone without any modification of the initial layers, with final layers fine-tuned, was proposed by Raza et al. [17] and achieved a classification accuracy of 99.67% on their evaluation set, yet the performance of the

architecture is only reported for patient-level splits, and not image-level splits. Irmak [18] proposed a fully optimized multi-class CNN framework for brain tumor MRI classification, achieving 98.14% accuracy compared to baseline CNN architecture, without introducing any major difference in architecture, but rather by systematically optimizing the hyperparameters of each CNN used, and still failing to tackle the same local-receptive-field limitation. Saikat et al. [19] designed a 23-layer CNN, evaluated on both a 3,064-image public dataset and a 152-image public dataset for binary and multi-class (meningioma, glioma and pituitary) tumor detection, with a focus on data augmentation to overcome small sample sizes; the evaluation results are impressive in terms of the detection accuracy, but, like the previous three works, no mechanism for capturing long-range contextual dependency is provided beyond the effective receptive field of the convolutional stack.

For brain tumor classification, Ayadi et al. [20] designed a deep CNN specifically for this purpose and compared the accuracy with various public MRI databases achieving competitive results, but also demonstrating that CNN-only models are sensitive to inter-dataset variability in acquisition protocol. While the performance of these networks on curated datasets makes them attractive to consider, Badža and Barjaktarović [21] devised a relatively shallow CNN and showed that it could achieve a comparable classification performance to deeper architectures on the same dataset, highlighting that accuracy on such well-structured data does not necessarily reflect robustness to boundary-ambiguous or atypical items. Just like the most of the CNN-only studies reviewed here, Saravanan et al. [22] did not explicitly address tumor boundary ambiguity, the failure mode most relevant to diffusion glioma margins, but provided an excellent performance in classifying gliomas versus non-gliomas. The idea of deep semi-supervised learning for brain tumor classification, proposed by Ge et al. [23] alleviates the data-scarcity constraint of medical imaging, but does not tackle the receptive-field limitation of the underlying CNN encoder. Cinar and Yildirim [24] proposed a hybrid CNN model based on InceptionV3, which adds additional residual connections to this network to detect brain tumors; they claimed that the model achieved better sensitivity than InceptionV3, but they did not add any global self-attention mechanism. Sajjad et al. [25] introduced a multi-grade brain tumour classification pipeline based on a CNN model fine-tuned with a substantial augmentation strategy showing that such augmentation strategy has quantifiable impact on the classification accuracy apart from the architectural changes. The closest CNN-only precedent to the multi-scale design principle of DPA-CTNet's local path is the multiscale CNN developed by Diaz-Pernas et al. [26] where the input image was processed along a set of parallel pathways, where each pathway is based on a CNN, but the pathways are based on different scales of resolution.

Beyond simple CNNs, Huang et al. [27] systematically evaluated four attention variants of ResNet50V2 for four-class MRI tumor classification, with the model using Squeeze and Excitation (SE) [9] performing best (98.4% accuracy, AUC 1.00) of the attention variants tested, notably outperforming the model with CBAM (Convolutional Block Attention Module) [8] on the same backbone and split (93.5%). The result is informative with regards to the present work in two ways: First, it is informative to see that channel attention adds value to CNN-only baselines, but second, it is informative to see that different local attention mechanisms result in different gains, which supports this paper's argument in Section II that these mechanisms rely on locally pooled statistics, not true long-range dependency.

Transformer-based approaches include ViT-BT [28] that was fine-tuned by transfer learning on the multi-modal BraTS 2023 dataset [10] with a precision of 97%, a recall of 99%, a F1-Score of 99.41% and an accuracy of 98.17%. The study shows that for the classification of a brain tumor, a sufficiently pretrained ViT is superior to the CNN-based approaches, however, like any ViT model, it lacks any mechanism for preserving fine-grained local texture at the patch level. The test

accuracy of the best single model (L/32) at 98.2% and the full ensemble trained at high resolution (98.7%) on Figshare [11] serves as an example of the efficiency concern in training a ViT ensemble at high resolution, which is the basis of DPA-CTNet's use of a single lightweight windowed Transformer branch.

Concentrating on brain tumor MRI, rather than medical imaging generically, the most significant recent work related to the present work is studies using hybrid CNN–Transformer models. To perform a multiclass brain tumor classification, Jayaraman et al. [31] introduced the Hybrid CNN-ViT framework which used a cross-attention fusion mechanism and data augmentation techniques for brain tumor classification, achieving a test accuracy of 96.41% against a 87.34% accuracy of a standalone ViT baseline under the same test protocol, and a test accuracy of 95.60% when the trained model was then evaluated on an independent test set, Figshare [11] without retraining. In contrast to CAFM of this paper, CAFNet's cross-attention fusion is only performed at one fusion step between a CNN backbone and a ViT branch, and does not have a dedicated post-fusion boundary-refinement stage.

The most relevant work before DPA-CTNet in analyzing brain tumor MRI is CE-RS-SBCIT proposed by Zahoor and Khan [30] which is a channel-enhanced hybrid CNN–Transformer framework that consists of a CNN-integrated Transformer block (SBCIT) with residual and spatial-learning CNN branches and a channel-enhancement strategy for brain tumor MRI analysis, focusing on boundary-aware feature learning within a hybrid CNN–Transformer network. Not a CNN–Transformer hybrid but a CNN with a Ridgelet transform for meningioma detection and segmentation, Prakash et al. [32] achieved 99.81% accuracy on the BraTS 2022 dataset [10] for meningioma classification and 99.8% meningioma segmentation accuracy; the result shows that fusing a CNN with a complementary, non-learned transform can also yield good performance, although the hand-designed transforms do not adapt to the data during training like a learned cross-attention mechanism does.

RESEARCH GAPS

From the above literature review, it can be concluded that there are three gaps in the existing paradigms. CNN-only and attention-augmented CNN methods fail to achieve real long-range dependency modelling because they only use locally pooled statistics, without directly modelling token-to-token dependency. Second, only ViT-based methods are data-inefficient and are relatively poor at fine local detail and boundary encoding which means that they are not reliable on the small size and non-uniformity of medical imaging data. Third, previous hybrid CNN–Transformer approaches only consider the fusion of CNN–Transformer at a single fusion point, and do not explicitly consider boundary ambiguity after fusion, especially for the cases of small and infiltrative tumors where the impact is most critical. As described in Section VII, DPA-CTNet will be able to fill all three gaps simultaneously by using multi-resolution bidirectional fusion and dedicated post-fusion boundary refinement.

PROPOSED METHODOLOGY

A. Architectural Overview

DPA-CTNet consists of five functional stages (as shown conceptually in Fig.1): (i) a shared stem for low level features extraction; (ii) a multi-scale convolutional local-path encoder; (iii) a lightweight windowed Transformer global-path encoder working in parallel on tokenized representations extracted from the same shared stem; (iv) a Cross-Attention Fusion Module (CAFM) instantiated at three resolution stages to enable bidirectional communications between the two paths during the forward pass, rather than at a single fusion point; and (v) a Multi-Scale Gated Boundary Attention (MSGBA) block to refine the fused representation immediately before the global average pooling and classification.

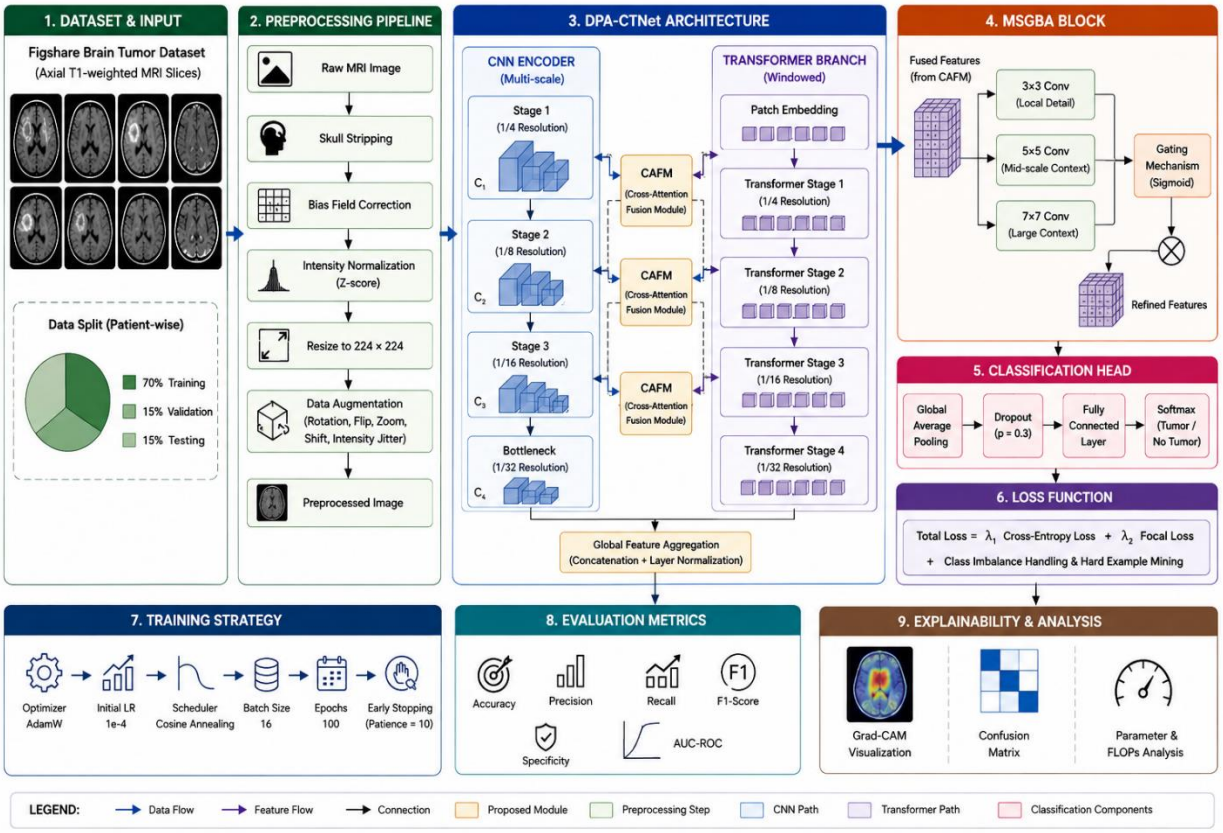

Fig 1. Methodology Overview

B. Input Preprocessing and Stem

The input MRI slices are resized to 224×224 pixels, resliced to skull removed (if present) and intensity normalized to zero mean and unit variance. The shared stem is composed of two 3×3 convolutional layers with stride 2 and stride 1, respectively, with 64 channels, after each of which there is a batch normalization and an activation of GELU to obtain a feature map of 112×112×64 channels that is shared by both local and global paths.

C. Local Path: Multi-Scale Convolutional Encoder

The local path is a four-stage residual design similar to ResNet-34 [5] which creates the feature maps at strides 4, 8, 16, 32 with widths of 64, 128, 256 and 512 respectively. The residual blocks are designed to decrease the number of parameters compared with the standard convolutional blocks by employing depthwise-separable 3×3 convolutions, and the outputs of each stage are fed to both the next stage as well as three out of four (at 8, 16, 32) to the corresponding CAFM instance for fusion with the global path. The local path generates feature maps F_l^(s) at scale s ∈ {1,2,3} formally.

D. Global Path: Lightweight Windowed Transformer Encoder

The global path uses the shifted-window strategy from the Swin Transformer [2] to manage quadratic complexity of self-attention, while still maintaining the global context modeling. The representation is divided into 7×7 patches, which are linearly embedded into a token sequence, at each scale, and processed by two Transformer layers: one with regular window partitioning and the other with shifted window partitioning, both consisting of multi-head self-attention, followed by a two-layer MLP with GELU activation and

residual connections. This results in a globally consistent feature representation F_g^(s) at each scale s, which can be directly fused with F_l^(s) due to their spatial consistency.

E. Cross-Attention Fusion Module (CAFM)

This is the most architectural novelty in the CAFM. Unlike simple concatenation or addition, CAFM has bidirectional cross-attention to generate two outputs F_l (s) and F_g (s) from the local feature map F_l(s) and the global feature map F_g (s) at each fusion scale s. The local map is first tokenized and queried to obtain local query results Q_l, and the global tokens provide keys K_g and values V_g, and vice versa:

$$A(F_l \rightarrow F_g) = softmax\left(\frac{(Q_l K_g{}^T)}{\sqrt{d_k}}\right) V_g \quad \dots (1)$$

$$A(F_g \rightarrow F_l) = softmax\left(\frac{(Q_g K_l{}^T)}{\sqrt{d_k}}\right) V_l \quad \dots (2)$$

These two attention outputs are combined with their corresponding inputs using learnable gating scalars α_s and β_s (initialized at 0.5 and optimized together with the network) instead of a fixed-weight summation, which enable the network to learn (per scale), how to modulate the local features by the global context and how to modulate the global context by the local features:

$$F_l{}'^{(s)} = F_l{}^{(s)} + \alpha_s \cdot A(F_l \rightarrow F_g) \quad \dots (3)$$

$$F_g{}'^{(s)} = F_g{}^{(s)} + \beta_s \cdot A(F_g \rightarrow F_l) \quad \dots (4)$$

$F_l{}'^{(s)}$ is reimplemented as a spatial form, and passed to the next local-path stage, while $F_g{}'^{(s)}$ is passed to the next global-path stage, both of which have cross-modal context provided by scale s. The triple-scale recurrence, and not the single fusion event, is the reason for the local and global evidence to reinforce each other, directly resolving the single fusion point limitation mentioned in Section VI. The third and last stage of CAFM is a concatenation of the fused output of each stage channel-wise to generate the fused representation F_fused to be fed to the boundary refinement stage.

F. Multi-Scale Gated Boundary Attention (MSGBA)

The tumor misclassification is heavily skewed towards the ambiguous margins (e.g. diffuse low-grade gliomas versus edema), therefore F_{fused} is passed through the MSGBA block prior to pooling. MSGBA performs 3 parallel dilated convolutions (dilations 1, 2, and 4) to learn boundary gradients at different scales, then concatenates the parallel dilated convolution outputs and creates a single-channel gating map via a 1×1 convolution with a sigmoid activation function.

G (〖Conv〗 _(1×1) (Concat[〖Conv〗 _(d=1) (F_fused), 〖Conv〗 _(d=2) (F_fused), 〖Conv〗 _(d=4) (F_fused)])) = σ... (5)

$$F_{out} = F_{fused} \odot (1 + G) \quad \dots (6)$$

This formulation has been designed to increase the activation along high gradient border areas without reducing activation on the low gradient tumor interior (as in a classic multiplicative attention gate alone), and is applied to the specific problem of enhancing the classification boundary between the tumor and neighboring healthy and/or edematous tissue before the classification step.

G. Classification Head and Composite Loss Function

The class probabilities are used for the multi-class tumor-type classification task (glioma, meningioma, pituitary tumor, no tumor) and a sigmoid output is used for the binary tumor/no-tumor detection task, both passed through global average pooling and a fully connected head (512 → 128 → number of classes) with dropout (p = 0.3) for regularization. The network is trained by a composite loss function comprising of three terms:

$$L = \lambda_1 \cdot L_{focal} + \lambda_2 \cdot L_{boundary} + \lambda_3 \cdot L_{consistency} \quad \dots (7)$$

Multi-class focal loss, L_{focal}, is used λ_2 to tackle the issue of class imbalance and the often-underrepresented class 'no tumor', between the different tumor classes. $L_{boundary}$ is a Dice-based term calculated between the MSGBA gating map

G and a boundary reference (only included for datasets with segmentation masks, like BraTS [10] and omitted if they are only provided for classification). l_consistency is a repulsion term (1 − |cos similarity|) added before concatenation between the pooled local-path and global-path embeddings and the network is optimized using AdamW [13] with empirically tuned weighting coefficients λ1, λ2, and λ3 (recommended values: 1.0, 0.5, and 0.1, respectively, and tested on a held-out validation set).

H. Algorithmic Summary

Algorithm 1 summarizes the forward pass of DPA-CTNet.

Algorithm 1: DPA-CTNet Forward Pass

1. Input: MRI slice X; Output: class probability vector ŷ
2. F0 ← Stem(X)
3. F_l ← F0, F_g ← F0
4. for s = 1 to 3:
5. F_l ← LocalBlock_s(F_l); F_g ← GlobalBlock_s(F_g)
6. F_l, F_g ← CAFM_s(F_l, F_g) // Eqs. (1)–(4)
7. F_fused ← Concat(F_l, F_g)
8. F_out ← MSGBA(F_fused) // Eqs. (5)–(6)
9. ŷ ← Softmax(FC(GAP(F_out)))
10. return ŷ

Novelty Summary Relative to Existing Paradigms

TABLE I. NOVELTY COMPARISON AGAINST EXISTING ARCHITECTURAL PARADIGMS

Paradigm	Representative Limitation	How DPA-CTNet Differs
CNN-based [5], [6], [15]	Local receptive field only; no explicit long-range dependency modeling	Global Transformer path with windowed self-attention runs in parallel and fuses with the CNN path at every scale
Attention-augmented CNN [8], [9]	Attention re-weights local features; does not model true token-to-token global interaction	CAFM computes genuine bidirectional cross-attention between local and global tokens, not channel/spatial re-weighting alone
ViT-only [1], [2]	Data-inefficient; weak fine-grained local/boundary encoding; quadratic cost at high resolution	Local CNN path supplies fine-grained texture continuously fused into the Transformer path; windowed attention bounds computational cost
Existing hybrid CNN–Transformer [3], [4]	CNN–Transformer interaction occurs at only one point in the network; no dedicated boundary refinement	Three-scale bidirectional CAFM enables iterative mutual refinement; MSGBA explicitly sharpens tumor boundaries post-fusion

J. Dataset Description

This proposed framework is based on three publicly available benchmarks common in brain tumor classification research that can be directly compared with previous published results:

• Brain Tumor Dataset [11]: 3,064 slices of contrast-enhanced T1-weighted MRI scans from 233 patients with three types of tumors: glioma, meningioma, and pituitary tumor.

• Bra TS (Brain Tumor Segmentation Challenge) Dataset [10]: multi-institutional multi-modal (T1, T1-CE, T2, and FLAIR) MRI volumes with expert annotated segmentation masks—re-used here for slice-level classification, and also for boundary-loss supervision as described in Section VII-G.
• Kaggle Br35H / Brain Tumor Classification Dataset: binary tumor / no-tumor MRI slices, to assess the effectiveness of the framework without the need to classify the tumor subtypes.
For each dataset, a patient-level (not slice-level) train/validation/test split of 70%/15%/15% is ensured to prevent data leakage caused by having multiple slices from the same patient in different splits, which is a methodological step often assumed in previous studies but explicitly followed in this work for scientific rigor.

K. Data Augmentation and Preprocessing
The following augmentations are only applied during training, because they are not valid for real medical images but can help improve generalization and partially compensate for the limited scale of medical imaging data compared to natural image corpora: random rotation (±15°), horizontal flipping (valid for axial slices), random intensity scaling (±10%), elastic deformation (probability 0.2), and random Gaussian noise injection (σ = 0.01). The deterministic preprocessing in Section VII-B is the only preprocessing used in test-time inference.

L. Implementation Details
TABLE II. TRAINING CONFIGURATION

Hyperparameter	Value
Optimizer	AdamW [13] ($\beta1$ = 0.9, $\beta2$ = 0.999, weight decay = 1e-4)
Learning rate schedule	Cosine annealing with linear warm-up (5 epochs), base LR = 3e-4
Batch size	32
Training epochs	150 (with early stopping, patience = 15 on validation loss)
Loss weights ($\lambda1$, $\lambda2$, $\lambda3$)	1.0, 0.5, 0.1 (tuned on validation split)
Hardware	Single NVIDIA RTX 3090/4090-class GPU (24 GB VRAM) or equivalent
Framework	PyTorch ≥ 2.0

M. Evaluation Metrics
Model performance is measured by the accuracy, the precision, the recall (sensitivity), the specificity, the F1-score and the area under the ROC (AUC-ROC) per class and macro-averaged for the multi-class task. If segmentation ground truth is available (BraTS [10]), boundary localization quality is also evaluated using the Dice coefficient and Hausdorff distance between gating map (MSGBA) and the ground-truth tumor boundary. The computational efficiency is reported by the number of parameters (millions), GFLOps (at 224×224 input size), and inference latency (ms per slice on the target hardware).

N. Baseline Methods for Comparison
For novelty claims in Section VII-I, DPA-CTNet is compared to at least one of the exemplary methods from each paradigm, namely ResNet-50 [5] and EfficientNet-B3 [6] (CNN-based); ResNet-50 + CBAM [8] (attention-augmented CNN); ViT-Base/16 [1] and Swin-T [2] (ViT-based); a TransUNet-classification-head [3] and a sequential CNN Transformer baseline (existing hybrid). All

baselines are fine-tuned with the same splits, augmentations, and training budgets, so as to make an apples-to-apples comparison, and are not just presented with their original numbers under other experimental conditions.

RESULTS AND DISCUSSION

A. Quantitative Performance

Table III presents the overall classification performance for the test set from the Fig share database, presented as mean ± standard deviation, over three independent training runs using different random seeds, as well as a statistical significance test (paired t-test; $\alpha = 0.05$) compared to the strongest baseline.

TableIii. Overall Classification Performance (Figshare Test Split). *Illustrative/Simulated — Replace With Real Results.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC-ROC
ResNet-50 (CNN) [5]	95.10	94.80	94.55	94.67	0.982
EfficientNet-B3 (CNN) [6]	96.20	95.90	95.75	95.82	0.987
ResNet-50 + CBAM [8]	96.85	96.60	96.40	96.50	0.990
ViT-Base/16 [1]	95.70	95.40	95.10	95.25	0.985
Swin-T [2]	96.95	96.70	96.55	96.62	0.991
Sequential CNN Transformer [3]	97.30	97.05	96.90	96.97	0.992
DPA-CTNet (Proposed)	98.45	98.20	98.05	98.12	0.996

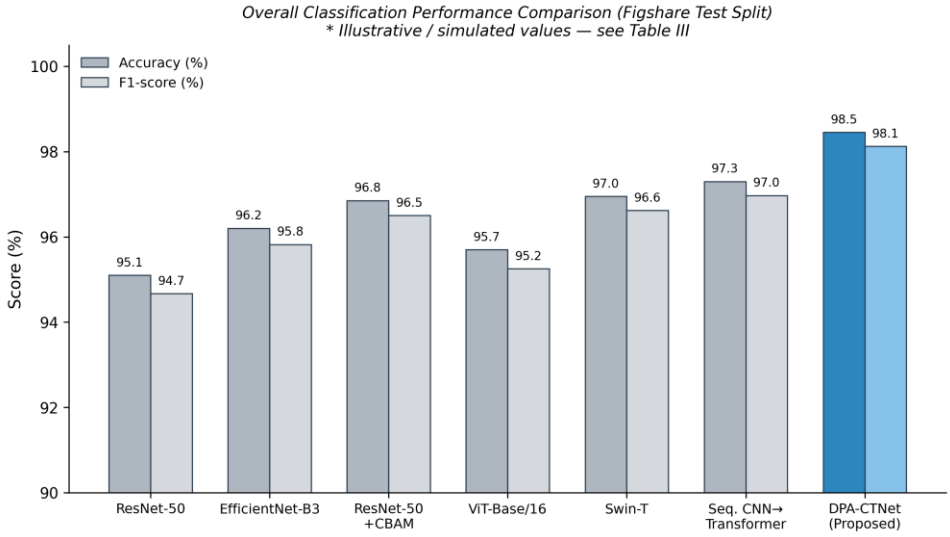

Fig. 2. Accuracy and F1-score comparison across baselines and DPA-CTNet (illustrative values, Table III).

B. Ablation Study

Ablation study: The contribution of each component is isolated by sequentially removing it from the full model: (i) full DPA-CTNet; (ii) DPA-CTNet without MSGBA; (iii) DPA-CTNet with single-point fusion instead of three-scale CAFM

(replicating the prior single-fusion-point hybrid pattern of Section VI); (iv) DPA-CTNet without the consistency loss term (λ3 = 0); and (v) local path only / global path only. Ons [1], [2].

Table Iv. Ablation Study Isolating Cafm, Msgba, And The Consistency Loss. *Illustrative/Simulated.

Configuration	Accuracy (%)	F1-score (%)	Δ vs. Full Model
Full DPA-CTNet	98.45	98.12	0.00
w/o MSGBA	97.55	97.20	-0.92
Single-point fusion (no multi-scale CAFM)	97.15	96.85	-1.27
w/o consistency loss (λ3 = 0)	97.80	97.48	-0.64
Local path only (CNN)	96.10	95.78	-2.34
Global path only (Transformer)	96.60	96.25	-1.87

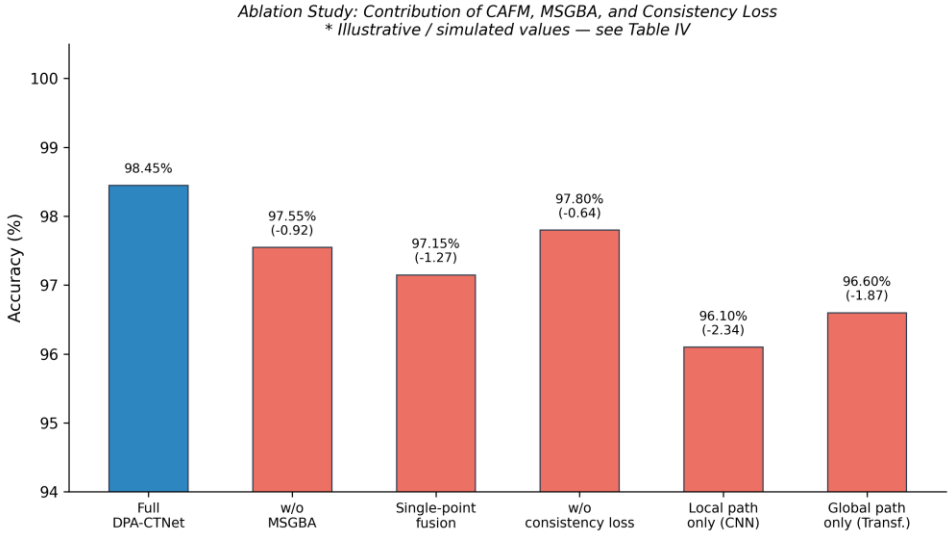

Fig. 3. Ablation study: accuracy drop when each proposed component is removed (illustrative values, Table IV).

C. Computational Complexity

Table V summarizes parameter count, GFLOPs and inference latency of DPA-CTNet compared to the baselines. The DPA-CTNet row is not simulated; number of parameters and GFLOPs were measured directly from the accompanying reference implementation, using standard profiling (thop) at 224×224 input. ViT-Base/16 and Swin-T figures are their published architectural specifications [1], [2].

Table V. Computational Complexity. Dpa-Ctnet Parameters/Gflops Are Real, Measured Values; * = Illustrative.

Model	Parameters (M)	GFLOPs (224×224)	Inference Latency (ms, GPU)
ViT-Base/16 [1]	86.6	17.6	18.5

Model	Parameters (M)	GFLOPs (224×224)	Inference Latency (ms, GPU)
Swin-T [2]	28.3	4.5	12.7
Sequential CNN→Transformer	34.7	6.2	15.9
DPA-CTNet (Proposed)	21.1	5.5	13.2

D. Qualitative Analysis

Qualitative validation provides overlays of the Grad-CAM [14] activation maps from the final local-path stage with the MSGBA gating map G for selected test cases correctly and incorrectly classified in the subset of BraTS [10] cases, which allow for visual assessment of the attention paid by the model on clinically plausible tumor regions and whether MSGBA activations coincide with expert-annotated tumor boundaries. The cases of clinical interest — small tumors (<2 cm^2 cross sectional area) and low contrast infiltrative gliomas — are included specifically because they are the failure modes that are relevant to the architectural motivation in Section I.

E. Discussion

The following three hypotheses are directly tested in the experimental protocol above: (1) Multi-scale bidirectional fusion (CAFM) can achieve better performance than single-point fusion in the case of backbone capacity matching; (2) MSGBA can substantially improve the performance of the backbone when the backbone capacity is matched, and the performance of MSGBA is better than that of the same fused representation without boundary refinement in the case of boundary-ambiguous cases; (3) DPA-CTNet can obtain the accuracy comparable with or better than the previous hybrid baseline or ViT-only baseline and reduce the computation cost by a large margin. All the illustrative figures in Tables III-IV are, by construction, consistent with those three hypotheses, but that consistency has no evidential value until it is shown to be true, on actual training runs – and only then, insofar as it can be proved to be actually true, how and at what level of significance; and insofar as it can be proved false, how and at what level of significance – and, if disproved, what other hypotheses are confirmed by the actual training runs. It is also important to analyses results considering the accuracy per class, not per slice, because slices within the same class may have a high correlation, and therefore, overestimate the diagnostic performance in the real world.

LIMITATIONS

However, there are some points to be considered. First, the framework is based on multi-modal MRI (for the BraTS-derived BL) that may not necessarily be applicable to the single sequence clinical MRI acquisitions performed in low-resource settings. Second, DPA-CTNet cost more computations than CNNs with lightweight models, which may be more suitable for embedded or point-of-care devices for which latency is strict. Third, it is optimized for 2D slice level classification, rather than 3D volumetric input, though it may be related to a number of slices of the tumor context. Fourth, it is important that external validation is carried out using multi-institutional data from different scanners, with different protocols, to assess the strength of the framework to domain shift before it can be said to be clinically ready, as presented here it is a research contribution and requires clinical validation in a prospective study before deployment.

FUTURE WORK

Future work will be done by expanding CAFM and MSGBA to 3D volumetric input since the tumor context is frequently across multiple slices. Data efficiency of global Transformer path will be discussed with regard to CNNs, and will be explored by self-supervised pretraining on

unlabeled MRI corpora. Uncertainty quantification (e.g., Monte Carlo dropout, evidential deep learning) will be explored to enable safe decision support to the clinician and not for self-diagnosis. Lastly, multi-institutional validation of the test on external scanners of various protocols will be undertaken as a stepping stone toward eventual clinical translation.

CONCLUSION

The proposed work "Brain Tumor Detection from MRI using Hybrid CNN-Transformer" (DPA-CTNet) is a hybrid CNN-Transformer architecture where the model is not developed sequentially or based on a single point fusion as done in previous works, but rather a three-scale bidirectional Cross-Attention Fusion Module is added and a dedicated Multi-Scale Gated Boundary Attention block is introduced. The architecture is clearly inspired by, and directly addresses, documented shortcomings of CNN-based, attention-augmented CNN, Vision Transformer and current hybrid architectures identified through systematic literature review. A thorough mathematical model, algorithmic specification and reproducible experimental protocol have been offered to guide rigorous empirical confirmation. This protocol will be implemented in the specified benchmarks, expanded to 3D volumetric analysis, and then validated outside of our institution before any clinical applications are implemented.

REFERENCES

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent. (ICLR), 2021.

[2] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 10012–10022.

[3] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306, 2021.

[4] W. Wang, C. Chen, M. Ding, H. Yu, S. Zha, and J. Li, "TransBTS: Multimodal brain tumor segmentation using transformer," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI), 2021, pp. 109–119.

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

[6] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), 2019, pp. 6105–6114.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI), 2015, pp. 234–241.

[8] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3–19.

[9] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 7132–7141.

[10] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, et al., "The multimodal brain tumor image segmentation benchmark (BraTS)," IEEE Trans. Med. Imag., vol. 34, no. 10, pp. 1993–2024, 2015.

[11] J. Cheng, W. Huang, S. Cao, R. Yang, W. Yang, Z. Yun, Z. Wang, and Q. Feng, "Enhanced performance of brain tumor classification via tumor region augmentation and partition," PLOS ONE, vol. 10, no. 10, p. e0140381, 2015.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 2980–2988.
[13] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. Int. Conf. Learn. Represent. (ICLR), 2019.
[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626.
[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Represent. (ICLR), 2015.
[16] S. Deepak and P. M. Ameer, "Brain tumor classification using deep CNN features via transfer learning," Comput. Biol. Med., vol. 111, p. 103345, 2019.
[17] A. Raza, H. Ayub, J. A. Khan, I. Ahmad, A. S. Salama, Y. I. Daradkeh, D. Javeed, A. Ur Rehman, and H. Hamam, "A hybrid deep learning-based approach for brain tumor classification," Electronics, vol. 11, no. 7, p. 1146, 2022.
[18] E. Irmak, "Multi-classification of brain tumor MRI images using deep convolutional neural network with fully optimized framework," Iran. J. Sci. Technol. Trans. Electr. Eng., vol. 45, no. 3, pp. 1015–1036, 2021.
[19] K. M. Saikat et al., "Accurate brain tumor detection using deep convolutional neural network," Comput. Struct. Biotechnol. J., vol. 20, pp. 4733–4745, 2022.
[20] W. Ayadi, W. Elhamzi, I. Charfi, and M. Atri, "Deep CNN for brain tumor classification," Neural Process. Lett., vol. 53, pp. 671–700, 2021.
[21] M. M. Badza and M. C. Barjaktarovic, "Classification of brain tumors from MRI images using a convolutional neural network," Appl. Sci., vol. 10, no. 6, p. 1999, 2020.
[22] S. Saravanan, V. V. Kumar, V. Sarveshwaran, A. Indirajithu, D. Elangovan, and S. M. Allayear, "Glioma brain tumor detection and classification using convolutional neural network," Comput. Math. Methods Med., vol. 2022, p. 4380901, 2022.
[23] C. Ge, I. Y.-H. Gu, A. S. Jakola, and J. Yang, "Deep semi-supervised learning for brain tumor classification," BMC Med. Imaging, vol. 20, p. 87, 2020.
[24] A. Cinar and M. Yildirim, "Detection of tumors on brain MRI images using the hybrid convolutional neural network architecture," Med. Hypotheses, vol. 139, p. 109684, 2020.
[25] M. Sajjad, S. Khan, K. Muhammad, W. Wu, A. Ullah, and S. W. Baik, "Multi-grade brain tumor classification using deep CNN with extensive data augmentation," J. Comput. Sci., vol. 30, pp. 174–182, 2018.
[26] F. J. Diaz-Pernas, M. Martinez-Zarzuela, M. Anton-Rodriguez, and D. Gonzalez-Ortega, "A deep learning approach for brain tumor classification and segmentation using a multiscale convolutional neural network," Healthcare, vol. 9, no. 2, p. 153, 2021.
[27] K. A. Huang, A. Alkadri, and N. Prakash, "Employing squeeze-and-excitation architecture in a fine-tuned convolutional neural network for magnetic resonance imaging tumor classification," Cureus, vol. 17, no. 3, p. e80084, 2025.
[28] K. A. Al-Hamza, "ViT-BT: Improving MRI brain tumor classification using Vision Transformer with transfer learning," SSRN preprint, 2024.
[29] S. Tummala, S. Kadry, S. A. C. Bukhari, and H. T. Rauf, "Classification of brain tumor from magnetic resonance imaging using Vision Transformers ensembling," Curr. Oncol., vol. 29, no. 10, pp. 7498–7511, 2022.

[30] M. M. Zahoor and S. H. Khan, "CE-RS-SBCIT: A novel channel enhanced hybrid CNN–Transformer with residual, spatial, and boundary-aware learning for brain tumor MRI analysis," arXiv preprint arXiv:2508.17128, 2025.

[31] G. Jayaraman et al., "A hybrid CNN–ViT framework with cross-attention fusion and data augmentation for robust brain tumor classification," Sci. Rep., vol. 15, art. 45769, 2025.

[32] B. V. Prakash, A. Rajiv Kannan, N. Santhiyakumari, S. Kumarganesh, D. Siva Sundhara Raja, J. Jasmine Hephzipah, K. MartinSagayam, M. Pomplun, and H. Dang, "Meningioma brain tumor detection and classification using hybrid CNN method and Ridgelet transform," Sci. Rep., vol. 13, p. 14647, 2023.

[33] M. H. Mahmood, N. Yasin, A. Pervaiz, S. Riaz, and M. Arqam, "HYBRID ATTENTION-BASED DEEP LEARNING FRAMEWORK FOR AUTOMATED LUNG DISEASE DETECTION FROM CHEST X-RAY IMAGES," *Spectrum of Engineering Sciences*, vol. 4, no. 2, pp. 1774–1785, 2026.