

Explainable Multi-model Attention-based Deep Learning Algorithm for Brain Tumor Segmentation

Nayon Karmoker¹[0009–0003–6647–4276], Shrawman Majumder
Rudra¹[0009–0007–9278–2597], Fahima Tabassum¹[0000–0002–0183–3907], Risala
Tasin Khan ^{*}1[0000–0001–8236–5959], and M. Shamim
Kaiser¹[0000–0003–4721–8798]

Institute of Information Technology, Jahangirnagar University, Savar, Dhaka

Abstract. MRI-based brain tumor segmentation is essential for diagnosis, treatment planning, and monitoring. Despite advances in deep learning, combining long-range context with precise tumor boundaries remains difficult. We propose a Dual-Branch Attention Fusion Network that integrates an attention-enhanced U-Net for local detail with a Vision Transformer for global context. A Region-Guided Attention Fusion (RGAF) module adaptively combines both branches, while Explainable AI (XAI) highlights regions influencing segmentation. On BraTS2020, the framework achieves an average Dice score of 0.870 with low 95th-percentile Hausdorff distances, providing a robust and interpretable approach for AI-assisted brain tumor analysis.

Keywords: Brain Tumor Segmentation · MRI scans · Deep Learning · Dual-Branch Attention Fusion Network · Transformer · RGAF.

1 Introduction

Brain tumors are heterogeneous pathological growths affecting the brain or central nervous system [23]. They are classified as primary tumors arising in the brain or secondary metastatic tumors spreading from cancers elsewhere in the body [22]. Common primary tumors include gliomas, meningiomas, and pituitary adenomas, which differ in aggressiveness, mass effect, and systemic consequences [6]. Accurate MRI delineation is important for treatment planning, yet manual segmentation is time-consuming and prone to intra- and inter-observer variability, particularly during longitudinal monitoring [20]. Irregular shapes and variable boundaries further complicate diagnosis, although multi-sequence MRI provides detailed structural information. Deep learning can automate this analysis, reducing clinician workload while improving consistency and accuracy [16, 9]. Recent models still struggle to combine global context with precise local features, while explainability often relies on qualitative attention maps with limited quantitative validation [7, 26]. We therefore propose a Dual-Branch Attention

^{*} Corresponding author.

Fusion Network (DBAFN) combining CNN and Transformer representations, with Grad-CAM and Layer-wise Relevance Propagation (LRP) for interpretation, while retaining complementary local spatial detail and long-range contextual information for segmentation. This also reflects wider calls for explainable, privacy-preserving and human-centred clinical AI [13, 12, 11]. The main contributions of this paper are as follows:

- We develop a Dual-Branch Attention Fusion Network that couples a 3D U-Net branch with channel and spatial attention to a Vision Transformer branch for global context modeling.
- We design a Region-Guided Attention Fusion (RGAF) module in which an ROI map from the local branch guides cross-attention over the global branch, so that context is aggregated by tumor relevance rather than uniformly.
- We integrate Grad-CAM and Layer-wise Relevance Propagation (LRP) to visualize salient tumor regions and voxel-level relevance behind the model’s decisions.

2 Related Works

Research on brain tumor segmentation has largely progressed along three directions. Firstly, CNN-based segmentation improves U-Net variants by utilizing residual encoders, 3D convolutions, GLCM features, and YOLO detectors. This method raises Dice scores of more than 0.95 and classification accuracy up to 99%; however, it performs poorly for small tumors, morphological variability, and multi-center generalization. Secondly, Transformer attention (Swin [18], VSA-GCNN-BiGRU [17], and TransUNet [4]) and multimodal encoders (Pan [19], Pacal [18], Ting [24], among others) are utilized to acquire long-range context and address the issue of missing modalities. Related work on multimodal and explainable healthcare AI also highlights the value of combining heterogeneous representations while preserving interpretability [8, 27]. The execution of these algorithms is resource-intensive, and they often only function with a single dataset for validation purposes. Thirdly, hybrid CNN–transformer designs (Aboussaleh [2], Nguyen-Tat [15], Soh [21], and others) use attention-based feature fusion to merge local information with global dependencies. However, they are still unable to differentiate small lesions sufficiently and do not provide any clinical context. Across all three directions, a common limitation is the trade-off between accuracy, computational cost, and interpretability. CNN-based models are efficient but struggle with long-range context, transformer-based models capture global dependencies at a heavy computational expense, and hybrid designs improve accuracy yet remain difficult to deploy in resource-constrained clinical settings. Moreover, most of these methods treat segmentation as an isolated task, offering little insight into why a particular region is labeled as tumor. Where explainability is considered at all, it is usually restricted to coarse visual attention maps that highlight active regions without quantifying pixel-level contributions, which weakens clinical reliability and trustworthiness [7, 26, 25]. Overall, existing methods produce masks without transparent, clinically grounded interpretation,

and where explainability is addressed it is rarely tied to quantitative evidence. In contrast to hybrid designs that combine the two streams by concatenation or a fixed-weight sum applied uniformly across the volume, the proposed RGAF module derives an ROI map from the local branch and uses it to gate cross-attention over the global branch, so that global context is aggregated in proportion to estimated tumor relevance rather than uniformly. This design also aligns with wider work advocating explainable, inclusive, privacy-aware and human-centred healthcare AI [13, 12, 11].

3 Methodology

3.1 Overview

The proposed approach emphasizes brain tumor segmentation via a cohesive multi-stage pipeline, as depicted in Fig. 1. Initially, the BraTS 2020 MRI dataset was employed to ensure variation in tumor kinds, imaging circumstances, and data distribution, hence facilitating the model’s acquisition of more generalized representations. The MRI images obtained were then fed into the pre-processing stage where normalization was applied to make intensity values in different scans consistent. To deal with class imbalance and maintain the representation of tumor and non-tumor regions, tumor-centered three-dimensional patches were sampled from the cropped volumes. After preprocessing, the patches have been fed into the suggested segmentation model, which is based on Dual-Branch Attention Fusion Network. This architecture consists of a branch based on the Transformer, which captures global contextual relationships, and a branch based on CNN (U-Net), which captures fine-grained local spatial features. In order to effectively fuse the features from both branches and highlight the tumor-related regions, a Region-Guided Attention Fusion (RGAF) module was introduced. Channel and spatial attention were also added to enhance feature representation. To enhance interpretability, methods were applied to explainable AI (XAI) like Grad-CAM and LRP [7, 26].

3.2 Dataset

To assess the robustness and generalizability of the proposed model, the publicly available benchmark dataset BraTS 2020 was employed. It is a widely used benchmark for brain tumor segmentation using multi-parametric MRI scans, including four imaging modalities—T1, T1-weighted contrast-enhanced (T1CE), T2, and FLAIR—acquired from multiple institutions. Each subject is accompanied by expert-annotated ground truth masks delineating three tumor subregions: enhancing tumor (ET), tumor core (TC), and whole tumor (WT). Each MRI volume standardized to a spatial resolution of $240 \times 240 \times 155$ voxels and provided in NIfTI format with corresponding segmentation masks. As a multi-site dataset, it introduces variability in scanner types and acquisition protocols while remaining a relatively controlled and structured benchmark, ensuring consistent input representation for reliable model training and comparative evaluation [14]. Details about the dataset used are presented in Table 1.



Fig. 1: Workflow of the proposed methodology, illustrating dataset collection, MRI preprocessing, dual-branch attention-based segmentation, and integration of explainable AI (XAI) techniques.

Table 1: Description of the used dataset, including source, type, format, size, sequences, and number of images

Attribute	BraTS 2020
Image Source	Public benchmark
Image Type	3D multi-parametric MRI
Image Format	NIFTI
Image Size	$240 \times 240 \times 155$
Modalities	T1, T1CE, T2, FLAIR
Task	Tumor segmentation

3.3 Data Preprocessing

Preprocessing is essential to make the MRI data uniform, noise free, and consistent across subjects and institutions to prepare the data for deep learning-based

segmentation. The six steps specified below represent the suggested pipeline shown in Figure 2. First, the four modalities (T1, T1CE, T2, and FLAIR) are all loaded at the same time and stacked within the subject to ensure robust data for each modality; the volumes are distributed already co-registered to a common anatomical template and skull-stripped, so no further alignment is required. Each modality is then separately z-scored to zero mean and unit variance over the intra-cranial voxels, thus allowing learning to focus on structural differences and not scanner specific artifacts. Each volume is subsequently cropped to the bounding box of its non-zero region—removing the surrounding margin devoid of brain tissue to lower computational needs without discarding any anatomy. The ground truth subregions (WT, TC and ET) are retained as three overlapping binary channels rather than merged, so that the model is optimized directly on the regions on which it is evaluated. Three-dimensional patches of size 128^3 voxels are then sampled from each volume, with tumor-centred oversampling so that the patches in every batch are centred on a foreground voxel, in order to expand the effective dataset size and counteract the severe class imbalance between tumor and non-tumor tissue. Finally, the normalized modalities are stacked channel-wise and linked with their region masks in a PyTorch-compatible format, producing high-quality, intensity-normalized, and spatially aligned volumes that ensure segmentation robustness and reproducibility.

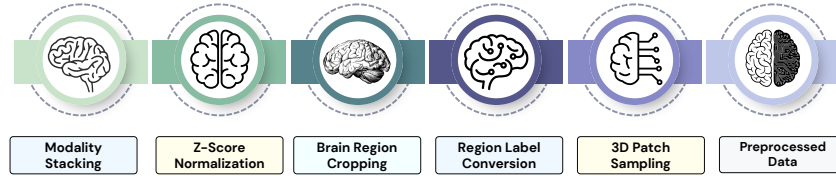


Fig. 2: Image Preprocessing Techniques used in this work

3.4 Proposed System Architecture

In order to efficiently extract both local fine-grained tumor borders and global contextual dependencies from MRI scans, the suggested segmentation approach uses a dual-branch attention fusion network operating on three-dimensional patches. Both branches receive the same input patch in parallel: a convolutional branch produces high-resolution local features, and a Transformer branch models long-range dependencies. A Region-Guided Attention Fusion (RGAF) module adaptively combines the two representations before sending them to a unified decoder for final mask generation. The general architecture of the suggested model is shown in Figure 3.

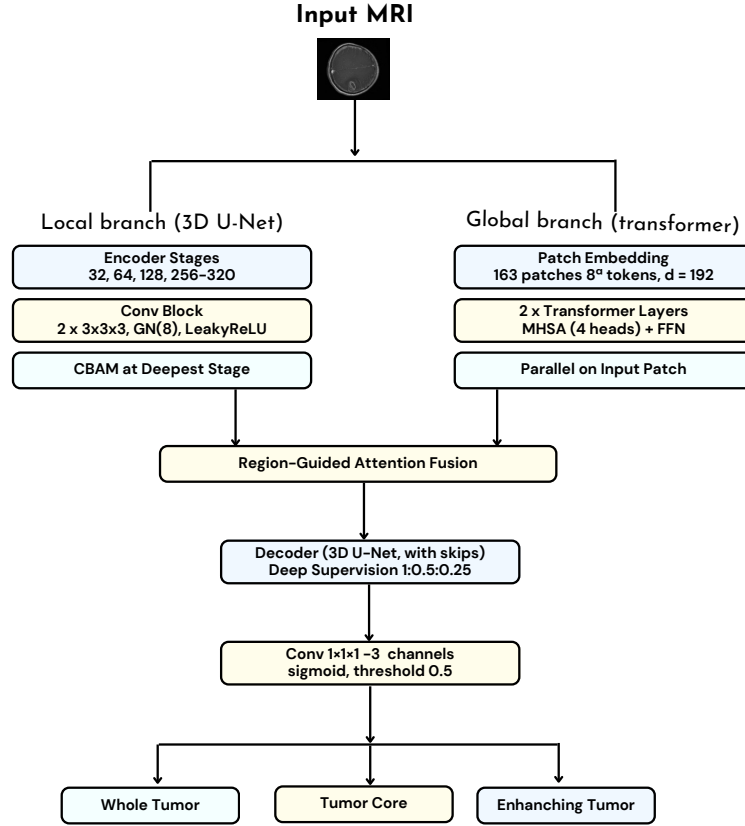


Fig. 3: Proposed Architecture of Segmentation Model

Branch 1: Attention U-Net (Local Feature Extraction) The first encoder branch is a three-dimensional U-Net backbone. Group normalization (GN), LeakyReLU activation, and two $3 \times 3 \times 3$ convolutional layers make up each encoder block. The convolution operation can be expressed as:

$$F = \sigma(\text{GN}(W * X + b)) \quad (1)$$

where X is the input feature volume, W represents the convolutional weights, b is the bias, and σ denotes the LeakyReLU activation function. The feature volumes are gradually downsampled across four encoder stages of 32, 64, 128 and 256 channels followed by a 320-channel bottleneck, with each stage halving

the spatial resolution:

$$F_{down}(u, v, w) = \max_{(i,j,k) \in \Omega(u,v,w)} F(i, j, k) \quad (2)$$

where $\Omega(u, v, w)$ denotes the pooling window centred on output location (u, v, w) . A Convolutional Block Attention Module (CBAM) refines the deepest encoder feature volume F using channel and spatial attention mechanisms:

$$F' = M_c(F) \odot F \quad (3)$$

$$F'' = M_s(F') \odot F' \quad (4)$$

where M_c and M_s denote channel and spatial attention maps, and $\odot$ represents element-wise multiplication with broadcasting. This attention mechanism enables the network to selectively emphasize tumor-relevant features while suppressing irrelevant regions. Features from all encoder stages are forwarded to the decoder through skip connections, preserving the high-resolution local detail essential for precise segmentation.

Branch 2: Transformer (Global Context Modeling) The second encoder branch operates in parallel on the same input patch and uses a Transformer-based architecture to capture global dependencies. The patch is divided into non-overlapping 16^3 sub-patches, which are flattened and linearly projected into an 8^3 token grid of dimension 192:

$$z_i = E \cdot x_i + p_i \quad (5)$$

where x_i is the i -th sub-patch, E is the embedding matrix, and p_i represents positional encoding. The core of the Transformer is the multi-head self-attention mechanism:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \quad (6)$$

where Q , K and V are query, key and value matrices, and d_k is the dimensionality of the key vectors. The output of multi-head attention is:

$$\text{MHA}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O \quad (7)$$

Each attention layer is followed by a feed-forward network:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (8)$$

with residual connections and layer normalization around both sub-layers. Two such layers with four heads produce a globally contextualized representation F_g , which is reshaped back onto the volumetric grid.

Region-Guided Attention Fusion (RGAF) Module The RGAF module adaptively integrates local and global features into a unified representation. Let F_l denote local features from the U-Net branch and F_g denote global features from the Transformer. A single-channel ROI attention map is first generated from the local features:

$$A_{roi} = \sigma(W_r * F_l) \quad (9)$$

This map gates the local features, which then form the queries of a four-head cross-attention operation, while keys and values are formed from the global features:

$$Q = W_Q(A_{roi} \odot F_l), \quad K = W_K(F_g), \quad V = W_V(F_g) \quad (10)$$

$$F'_g = \text{Attention}(Q, K, V) \quad (11)$$

so that each spatial location attends to global context in proportion to its estimated relevance to the tumor region. Because F'_g is projected back onto the same grid and channel count as F_l , the two are combined as:

$$F_{fused} = \alpha F_l + (1 - \alpha) F'_g \quad (12)$$

where $\alpha = \sigma(\hat{\alpha})$ is obtained from a single learnable scalar. Finally, the fused feature volume is passed through a decoder with deep supervision to reconstruct the three region masks:

$$Y = \sigma(\text{Decoder}(F_{fused})) \quad (13)$$

Each of the three output channels is thresholded at 0.5 to yield the WT, TC and ET masks. This fusion strategy ensures both spatial precision and global semantic consistency.

3.5 Evaluation Metrics

Overlap-based, boundary-based and region-based measurements are used to evaluate the suggested segmentation framework to ensure that segmentation accuracy and clinical importance are comprehensively assessed. In particular, the Dice Similarity Coefficient (DSC) and 95th percentile Hausdorff Distance (HD95) are reported. Together, these metrics provide a holistic evaluation of the volumetric overlap, boundary conformity, and accuracy at the pixel level.

4 Experiments and Results

4.1 Implementation Details

The network is trained on random 128^3 patches (batch size 2, 70% tumor-centered, random flips along all axes) using the loss $\mathcal{L} = \mathcal{L} * \text{BCE} + \mathcal{L} * \text{Dice}$, averaged over the three tumor regions, with deep supervision weighted 1:0.5:0.25. Optimization uses AdamW (learning rate 3×10^{-4} , weight decay 3×10^{-5}) with cosine annealing over 300 epochs, gradient clipping at 12.0, and mixed-precision

training on two T4 GPUs. The encoder comprises four stages (32, 64, 128, and 256 channels) followed by a 320-channel bottleneck, using Group Normalization, LeakyReLU activations, and a CBAM block at the deepest stage. A parallel transformer processes 16^3 voxel patches as an 8^3 token grid (embedding dimension 192, depth 2, 4 heads), while RGAF employs 4-head cross-attention with a learnable sigmoid fusion weight initialized to 0.5. Table 2 reports best-epoch values on the original split, for comparability with published figures. Section 4.4 additionally reports subject-level 3-fold cross-validation with checkpoint selection on an inner split of the training folds, giving a more conservative estimate. During inference, Gaussian-weighted sliding-window prediction (50% overlap) with three-axis flip test-time augmentation is applied, followed by thresholding at 0.5, largest-component filtering for the whole-tumor mask, and removal of enhancing-tumor components smaller than 50 voxels. HD95 follows the standard BraTS convention for empty masks.

4.2 Comparison with State-of-the-Art Methods

Table 2: Comparison of segmentation performance on the BraTS2020 dataset (Baseline results are as reported in the original publications, under their own splits and implementations; they were not re-run under our protocol, so differences should be read with caution.)

Method	Year	Whole Tumor		Tumor Core		Enhancing Tumor		Average	
		Dice	HD95	Dice	HD95	Dice	HD95	Dice	HD95
Vox2Vox [1]	2020	0.872	6.44	0.811	24.36	0.787	18.95	0.823	16.58
U-Net++ [29]	2020	0.897	6.30	0.855	5.48	0.798	4.33	0.850	5.37
TransBTS [28]	2021	0.901	6.16	0.817	12.79	0.787	19.70	0.835	12.83
ACMINet [30]	2023	0.906	4.45	0.847	8.63	0.811	17.50	0.854	10.19
ADHDC-Net [10]	2023	0.899	5.25	0.833	9.82	0.780	29.34	0.837	14.80
3DUV-NetR+ [2]	2024	0.919	4.90	0.828	6.00	0.817	3.80	0.854	4.90
DAUNet [5]	2024	0.898	5.40	0.830	9.80	0.786	27.60	0.838	14.27
LATUP-Net [3]	2025	0.880	3.19	0.838	4.24	0.736	3.97	0.818	3.80
Proposed	2026	0.914	3.606	0.877	2.828	0.820	1.637	0.870	2.690

The model proposed in this paper achieves consistent accuracy, as seen in Table 2. It compares favourably with recent state-of-the-art techniques. Specifically, the proposed model obtains Dice scores of 0.914 for Whole Tumor (WT), 0.877 for Tumor Core (TC), and 0.820 for Enhancing Tumor (ET), giving an average of 0.870, the highest average among all methods in the table. The TC and ET scores are the best reported, while for WT the proposed model remains competitive with 3DUV-NetR+, which achieves 0.919. The model also has low 95th percentile Hausdorff Distance (HD95) values of 3.606 mm, 2.828 mm, and 1.637

mm for WT, TC, and ET, respectively, with the smallest mean HD95 (2.690 mm), indicating that it has successfully achieved low segmentation errors and accurate boundaries. The findings show that the proposed algorithm can achieve better Dice scores compared to transformer-based models like TransBTS, which mainly consider global dependency. Similarly, the proposed technique can exploit the transformer branch, enabling it to model long-range dependency and the complex structure of the tumors as well, which is better than CNN-based architectures like U-Net++ and ADHDC-Net. The enhancement is particularly significant in terms of boundary accuracy for the Enhancing Tumor (ET) region, which is usually difficult to delineate because it is small and irregular in shape or intensity; the HD95 value of this subregion has been reduced to 1.637 mm, the lowest value among the compared methods. The best margin over the compared methods is achieved for the Tumor Core in terms of Dice. Furthermore, the Region-Guided Attention Fusion (RGAF) module is applied with the aim of promoting consistency of the segmentations across all subregions.

4.3 Training Stability and Convergence Analysis

In order to evaluate the stability and convergence of the proposed framework, the training and validation Dice scores of the WT, TC, and ET regions are presented in Fig. 4. The optimization of all three subregions is steep in the first few epochs and then gradually improves before it stabilizes, with the training loss falling smoothly, flattening out towards the later epochs, suggesting stable optimization without divergence. As illustrated in Fig. 4(a), the Whole Tumor (WT) curves converge most rapidly and reach the highest Dice score as the overall segmentability of this bigger and more distinct region is easier. The heterogeneous intensity distribution of the Tumor Core (TC) region likely causes slower and noisier ascent of the validation curve in Fig. 4(b), but this curve remains close to the training curve. In Fig. 4(c), the Enhancing Tumor (ET) curves display the most gradual rise and the largest epoch-to-epoch fluctuation, consistent with the difficulty of delineating this small and irregularly shaped subregion; nevertheless, the oscillations dampen progressively and both curves stabilize in the final epochs. Across all three subregions, the small and consistent margin between the training and validation curves indicates minimal overfitting, demonstrating the robustness and generalization ability of the proposed methodology.

4.4 Cross-Validation and Ablation

The model was retrained under subject-level 3-fold cross-validation, with checkpoint selection driven by an inner split of the training folds so the held-out fold was never used for model selection. Held-out Dice scores were 0.838, 0.819 and 0.830 (mean 0.829 ± 0.008); selection scores exceeded held-out scores by 0.02–0.05 in every fold, quantifying the optimism of best-epoch selection. Three variants were trained under an identical protocol on 74 held-out subjects with CBAM enabled throughout (Table 3), evaluated on raw thresholded predictions without the post-processing used for Table 2, so absolute values are lower but

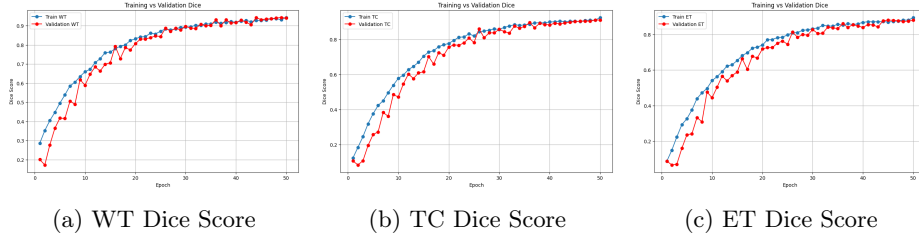


Fig. 4: Training vs Validation Dice scores for (a) Whole Tumor (WT), (b) Tumor Core (TC), and (c) Enhancing Tumor (ET) on the BraTS2020 dataset.

comparable across variants. RGAF improves average Dice by 0.023 over the backbone (+2.8%), which does not reach significance (Wilcoxon, $p = 0.127$) and is therefore suggestive rather than established. HD95 falls further, from 17.41,mm to 10.59,mm (39%), while concatenation fusion *increases* HD95 despite adding the same global branch indicating the gain derives from region-guided gating rather than added capacity. The largest per-region gain is for enhancing tumour (+0.038 Dice), at a cost of 0.41,M parameters (2%) and no measurable inference overhead.

Table 3: Ablation of the fusion mechanism

Fusion	WT	TC	ET	Avg.	HD95	Params (M)
Backbone	0.884	0.803	0.731	0.806	17.41	20.77
Concat	0.883	0.824	0.722	0.810	19.33	20.98
RGAF	0.886	0.832	0.769	0.829	10.59	21.18

4.5 Visual Interpretations using Explainable AI

To characterise which input regions influence the model’s output, Explainable AI (XAI) techniques are incorporated to visualize the model’s decision-making process. Grad-CAM and Layer-wise Relevance Propagation (LRP) are utilized to interpret the learned representations and highlight tumor-relevant regions, providing complementary insights by identifying salient regions and quantifying pixel-level contributions [7, 26, 25]. Grad-CAM creates coarse localization maps by calculating the gradients of the target class with respect to the feature maps of the last convolutional layers. As shown in Fig. 5, the visualizations distinctly emphasize the Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET) regions, indicating that the model successfully learns features related to tumor borders and internal structures. Unlike conventional CNN-based methods, the incorporation of attention mechanisms allows for improved localization in intricate and diverse tumor areas. The prediction score, on the other hand,

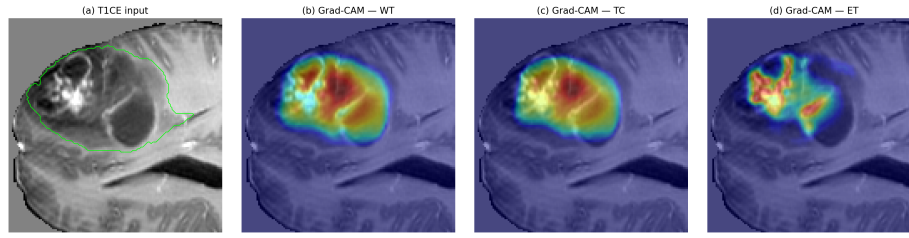


Fig. 5: Grad-CAM visualization highlighting important regions contributing to tumor segmentation. The heatmaps indicate the model’s focus on tumor-relevant areas across MRI slices.

propagates back through the network with LRP to assign relevance values to individual pixels, giving a more detailed and quantitative interpretation of the model’s predictions. The relevance maps shown in Fig. 6 reveal that the system focuses on areas of the tumor and on areas without a tumor, thus highlighting the fact that the model reflects the spatial and contextual information required for accurate segmentation. The relevance distribution is also the same for all types of tumors, further highlighting the robustness and interpretability of the model.

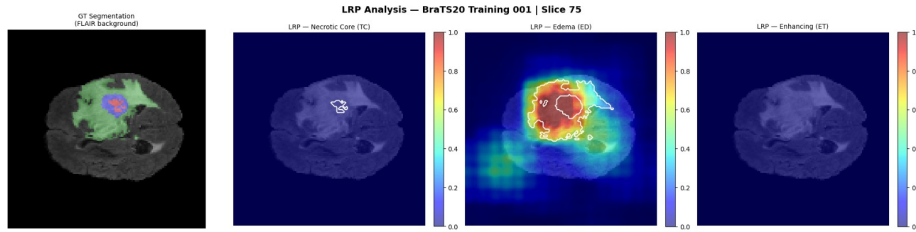


Fig. 6: LRP-based relevance visualization showing pixel-level contribution to segmentation output. Brighter regions indicate higher relevance to the model’s prediction.

4.6 Quantitative Evaluation of Explainability

Explanations were evaluated for localisation (Dice overlap of the top-decile relevance region with the ground-truth mask), fidelity (deletion score) and stability (SSIM between relevance maps under input perturbation), Table 4. LRP concentrates relevance on the tumour roughly five times more tightly than Grad-CAM, consistent with voxel-level relevance propagation versus coarse gradient maps. Deletion scores are comparable (0.239–0.249), indicating similar fidelity,

and both methods are highly stable. LRP is reported for WT and TC; the enhancing-tumour case is left for future work.

Table 4: Quantitative evaluation of XAI explanations

Method	Region	Top- k	Dice	Deletion	SSIM
Grad-CAM	WT	0.161	0.242	1.000	
Grad-CAM	TC	0.135	0.249	1.000	
Grad-CAM	ET	0.016	0.246	1.000	
LRP	WT	0.817	0.249	1.000	
LRP	TC	0.664	0.239	1.000	

5 Discussion and Conclusion

In this work, an explainable deep learning approach to brain tumor segmentation from multi-modal MRI was introduced. The Dual-Branch Attention Fusion Network combines a CBAM-enhanced U-Net encoder with a Vision Transformer encoder, with the connections between the two branches made through a Region-Guided Attention Fusion (RGAF) module that uses an ROI map from the local branch to guide the cross attention of the global branch, thus obtaining the context information by tumor relevance. Average Dice is 0.829 ± 0.008 under 3-fold cross validation where the held-out part is never used to select the checkpoint. Combined, the ablation isolates the fusion mechanism, yielding an improvement in average Dice of 0.023 over the backbone, although this improvement is not statistically significant at the subject level ($p = 0.127$), and reducing HD95 from 17.41,mm to 10.59,mm; concatenation fusion, on the other hand, worsens HD95 despite adding the same global branch, indicating that the boundary gain comes from the fusion of regions and not from the added capacity. LRP localises relevance more tightly than Grad-CAM, with comparable fidelity and stability. Limitations remain. Considering the small sample size and the fact that the ablation was applied in one split, the Dice advantage has not yet been proven statistically but the boundary-accuracy gain is more supported. Enhancing tumor performance remains lowest, due to its small size and irregular shape. Evaluation is only performed at BraTS2020 and baseline comparisons are made based on data from the original publications and not reruns with the same protocol. These explanations flag the regions that affected the output, but they are not the same as clinical relevance—that is, no reader study of the radiologists was conducted and no analysis was performed of the relationship between RGAF’s internal attention and post hoc relevance. Natural next steps are external validation, expert assessment of the explanations, comparison of RGAF’s internal attention with post-hoc relevance, per-step quantification of the inference-time post-processing, and richer gating formulations for the fusion weight. Responsible translation should additionally consider privacy, accessibility, human oversight and inclusive deployment [13, 12, 11].

Code and Data Availability

The BraTS-2020 dataset used in this study is publicly available through the data request portal <https://www.med.upenn.edu/cbica/brats2020/registration.html>. The code implementing the proposed framework is available at <https://github.com/Nayons/Brain-tumor-Segmentation>.

References

1. Vox2vox: 3d-gan for brain tumour segmentation. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer
2. Aboussaleh, I., Riffi, J., El Fazazy, K., Mahraz, A.M., Tairi, H.: 3duv-netr+: A 3d hybrid semantic architecture using transformers for brain tumor segmentation with multimodal mr images. *Results in Engineering* **21**, 101892 (2024)
3. Alwadee, E.J., Sun, X., Qin, Y., Langbein, F.C.: Latup-net: A lightweight 3d attention u-net with parallel convolutions for brain tumor segmentation. *Computers in Biology and Medicine* **184**, 109353 (2025)
4. Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M.P., Zhang, S., Xing, L., Lu, L., Yuille, A., Zhou, Y.: Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis* **97**, 103280 (2024)
5. Feng, Y., Cao, Y., An, D., Liu, P., Liao, X., Yu, B.: Daunet: A u-shaped network combining deep supervision and attention for brain tumor segmentation. *Knowledge-Based Systems* **285**, 111348 (2024)
6. Hasan, M.R., et al.: 3d mri reconstruction and brain tumor diagnosis using deep learning with explainable ai. *Expert Systems with Applications* p. 131513 (2026)
7. Hassija, V., et al.: Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation* **16**(1), 45–74 (2024)
8. Jahan, S., et al.: Explainable ai-based alzheimer's prediction and management using multimodal data. *PLoS ONE* **18**(11) (2023)
9. Javed, A.R., Saadia, A., Mughal, H., Gadekallu, T.R., Rizwan, M., Maddikunta, P.K.R., Mahmud, M., Liyanage, M., Hussain, A.: Artificial intelligence for cognitive health assessment: State-of-the-art, open challenges and future directions. *Cognitive Computation* **15**(6), 1767–1812 (2023)
10. Liu, H., Huo, G., Li, Q., Guan, X., Tseng, M.L.: Adhdc-net: Multiscale lightweight 3d segmentation algorithm with attention mechanism for brain tumor image segmentation. *Expert Systems with Applications* **214**, 119166 (2023)
11. Mahmud, M., Baslyman, M., et al.: Inclusive ai for digital human modelling: A perspective on explainable multimodal digital twins for accessibility and inclusion. In: *Proc. HCII*. pp. 1–16. Springer (2026)
12. Mahmud, M., Shaffi, N., Kaiser, M.S., et al.: Privacy-preserving multi-agent systems for human-centred inclusive healthcare ai. *AI Magazine* **47**(3), e70087 (2026)
13. Mahmud, M., et al.: Towards explainable and privacy-preserving artificial intelligence for personalisation in autism spectrum disorder. *Lecture Notes in Computer Science* **13309**, 356–370 (2022)
14. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., et al.: The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging* **34**(10), 1993–2024 (2015)

15. Nguyen-Tat, T.B., Nguyen, T.Q.T., Nguyen, H.N., Ngo, V.M.: Enhancing brain tumor segmentation in mri images: A hybrid approach using unet, attention mechanisms, and transformers. *Egyptian Informatics Journal* **27**, 100528 (2024)
16. Noor, M.B.T., Zenia, N.Z., Kaiser, M.S., Mamun, S.A., Mahmud, M.: Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of alzheimer's disease, parkinson's disease and schizophrenia. *Brain Informatics* **7**(1) (2020)
17. oshi, K.P., et al.: Vsa-gcn: Attention guided graph neural networks for brain tumor segmentation and classification. *Big Data and Cognitive Computing* **9**(2), 29 (2025)
18. Pacal, I.: A novel swin transformer approach utilizing residual multi-layer perceptron for diagnosing brain tumors in mri images. *International Journal of Machine Learning and Cybernetics* **15**(9), 3579–3597 (2024)
19. Pan, D., Shen, J., Al-Huda, Z., Al-qaness, M.A.A.: Vcanet: Vision transformer with fusion channel and spatial attention module for 3d brain tumor segmentation. *Computers in Biology and Medicine* **186**, 109662 (2025)
20. Pei, L., Bakas, S., Vossough, A., Reza, S.M.S., Davatzikos, C., Iftekharruddin, K.M.: Longitudinal brain tumor segmentation prediction in mri using feature and label fusion. *Biomedical Signal Processing and Control* **55**, 101648 (2020)
21. Soh, W.K., Yuen, H.Y., Rajapakse, J.C.: Hut: Hybrid unet transformer for brain lesion and tumour segmentation. *Heliyon* **9**(12), e22412 (2023)
22. Tanzhu, G., et al.: Metastatic brain tumors: From development to cutting-edge treatment. *MedComm* **6**(1) (2025)
23. Thenuwara, G., Curtin, J., Tian, F.: Advances in diagnostic tools and therapeutic approaches for gliomas: A comprehensive review. *Sensors* **23**(24), 9842 (2023)
24. Ting, H., Liu, M.: Multimodal transformer of incomplete mri data for brain tumor segmentation. *IEEE J Biomed Health Inform* **28**(1), 89–99 (2024)
25. Vimbi, V., Shaffi, N., Mahmud, M.: Interpreting artificial intelligence models: a systematic review on the application of lime and shap in alzheimer's disease detection. *Brain Informatics* **11**(1) (2024)
26. Viswan, V., Shaffi, N., Mahmud, M., Subramanian, K., Hajamohideen, F.: Explainable artificial intelligence in alzheimer's disease classification: A systematic review. *Cognitive Computation* **16**(1), 1–44 (2024)
27. Viswan, V., et al.: Multimodal fusion and explainability of artificial intelligence models in alzheimer's disease detection. *Brain Informatics* **13**(1) (2026)
28. Wang, W., Chen, C., Ding, M., Li, J., Yu, H., Zha, S.: Transbts: Multimodal brain tumor segmentation using transformer (2021)
29. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging* **39**(6), 1856–1867 (2020)
30. Zhuang, Y., Liu, H., Song, E., Hung, C.C.: A 3d cross-modality feature interaction network with volumetric feature alignment for brain tumor and tissue segmentation. *IEEE Journal of Biomedical and Health Informatics* **27**(1), 75–86 (2023)