Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

MAAF-Net: A Modality-Adaptive Attention Fusion Network for Multimodal 3D MRI Brain Tumor Segmentation

Sandeep Kaur[1], Usha Mittal[2], Ankita Wadhawan[3]

Department of Computer Science and Engineering, Lovely Professional University, Phagwara, India.

E-mail: [1]sandeeprajput3065@gmail.com, [2]usha.20339@lpu.co.in, [3]ankita.23891@lpu.co.in

Abstract

Multimodal MRI offers a comprehensive, non-invasive assessment of structure, which improves the diagnostic accuracy of brain tumor segmentation (BTS). BTS faces major challenges due to tumor heterogeneity, data quality, modality-specific information and algorithm complexity. Many existing methods do not utilize the complementary information available in multimodal data, as they depend on early fusion strategies and neglect modality-specific features. To overcome these issues, a novel architecture termed MAAF-Net: Modality-Adaptive Attention Fusion Network for 3D BTS has been proposed in this study. The MAAF-Net model preserves the semantic information of each MRI modality through modality-specific encoders (MSEnc). The extracted features are integrated using a Modality Attention Module (MAM). The MAM learns the context-dependent importance of each modality and adaptively reweights modality features during fusion. This fusion technique enables the model to focus on clinically relevant information while discarding redundant or less informative features. In addition, multi-scale supervision is incorporated to improve gradient flow and training stability. The MAAF-Net model is trained and validated on the BraTS 2024 benchmark dataset using five-fold cross-validation. The MAAF-Net achieves Dice scores of 0.92 for Tumor Core (TC) and 0.91 for Enhancing Tumor (ET) with HD95 values of 3.7 mm and 3.5 mm, respectively. Additionally, compared with early and attention-based fusion methods, MAAF-Net improves Dice scores by up to 8% for ET. Experimental findings from the ablation study further validate the effectiveness of the proposed model.

Keywords: Attention-Guided Feature Fusion, Multimodal MRI, Modality-Specific Encoders, Deep Learning, 3D U-Net, BraTS 2024.

1. Introduction

A brain tumor is a mass of abnormal brain or skull cell growth and an aggressive neurological disease. It impacts emotionally, psychologically, and physically on the human body [1]. MRI scans are essential tools for detection and providing detailed images due to their soft-tissue contrast [2,3]. The main MRI imaging modalities typically consist of T1-weighted, contrast-enhanced T1 (T1c), T2-weighted, and FLAIR. The choice of a particular modality depends upon the location, type, and size of the tumor. However, a single modality cannot provide acceptable results due to tumor heterogeneity, irregular boundaries, and overlapping intensity patterns [4-6].

DOI: https://doi.org/10.36548/jiip.2026.2.004
Received: 22.02.2026, received in revised form: 21.03.2026, accepted: 07.04.2026, published: 17.04.2026

Although deep learning (DL) approaches have improved segmentation performance, there are many limitations when multimodal MRI data is used [7]. To extract the features from multiple sequences of MRI, either early fusion strategies have been used or some studies used a shared encoder, which leads to the loss of modality-specific information [8,9]. These methods consider equal importance across modalities and often mix information prematurely. As a result, key features unique to each MRI modality may be lost. Another limitation is the lack of adaptive fusion mechanisms. Most networks apply uniform fusion strategies and fail to capture spatially varying modality relevance, such as the importance of T1c for enhancing tumor regions and the importance of FLAIR for edema.

Recent transformer-based methods have gained high popularity in the area due to their self-attention mechanism. Transformers capture global features from images rather than relying on local features only [10]. However, these introduce high computational costs and limited interpretability. Effective and reliable tumor segmentation is possible by combining features from multiple modalities and exploiting complementary information. To overcome this issue, MAAF-Net, a Modality-Adaptive Attention Fusion Network designed for multimodal 3D BTS has been proposed. The following are the key contributions of the proposed study:

- A modality-adaptive attention fusion to learn spatially varying weights for each MRI modality, enabling the model to integrate context-aware multimodal features.
- Modality-specific encoders preserve the semantic information of each MRI modality prior to feature fusion.
- A multi-scale supervision strategy helps to improve gradient propagation.

The remainder of the paper is structured as follows: A detailed literature review is discussed in Section 2. Section 3 presents the proposed methodology. A discussion on the experimental findings is given in Section 4. Finally, conclusion and future work are provided in Section 5.

2. Literature Review

Multimodal MRI plays a crucial role in BTS. Feature fusion techniques integrate features from multiple modalities into a unified and discriminative representation. DL, especially CNN-based architectures, have demonstrated strong capability in multimodal MRI segmentation. Several recent studies have therefore focused on enhancing segmentation performance.

He et al. (2021) proposed MAFF-ResUNet, which integrates spatial and semantic features across different resolutions. The proposed architecture incorporates residual and skip connections within a U-Net model to improve feature representation [11]. Ranjbarzadeh et al. (2021) proposed a Distance-Wise Attention (DWA) module within a cascaded CNN to emphasize tumor-centered regions while integrating local and global contextual information [12]. Huang et al. (2021) implemented a lightweight NN with a hybrid loss function. The model integrates spatial and semantic information across multiple scales using a Feature Extraction Network (FEN) and a Multi-scale Feature Fusing Network (MSFFN) [13].

Zhang et al. (2022) introduced a multimodal framework, which includes CMFF and CMFT modules, where CycleGAN is used to train the features of modality-specific

representations and an attention mechanism to combine the complementary features [14]. Zhao et al. (2022) extended MM-UNet by employing shared encoders for each modality and refining the fused features using a Hybrid Attention Block combined with dilated convolutions to enhance contextual representation [15].

Zhu et al. (2023) implemented a Swin-Transformer in which edge features were integrated through graph-based multi-feature inference to enhance contextual reasoning [16]. Zhang et al. (2023) designed a four-branch encoder architecture with residual concatenation fusion to preserve discriminative modality information [17]. Yang et al. (2023) proposed F2Net with two modules, CFM for noise-suppressed alignment and attention-based fusion, and MCM for guiding the decoder with enriched multimodal features [18]. Zheng et al. (2023) developed CMMFNet which integrates self-attention and cross-attention fusion along with wide-focus modules to capture intra- and inter-modality dependencies across multiple scales [23]. Zhou et al. (2023) developed feature-level attention fusion with multi-scale context aggregation and latent feature learning to preserve inter-modality consistency. The developed model also supports missing modality reconstruction [19]. Nizamani et al. (2023) used image enhancement techniques combined with a hybrid U-Net–Transformer to preserve tumor boundaries [20]. Khan et al. (2023) extracted features from original and contrast-enhanced MRI and fused them using multiset CCA, reducing redundancy via a hybrid meta-heuristic strategy [22].

Ullah et al. (2024) integrated handcrafted features with deep representations through a dual-stream CNN to improve tumor detection [21]. Zhu et al. (2024) proposed SDV-TUNet, to improve boundary delineation through convolution operations [24]. Zhou et al. (2024) utilized modality-specific features using self-attention. The method also enforced feature disentanglement and spatial consistency through DRL and RCL modules [25].

Wang et al. (2025) presented MSegNet, which employs cross-modal attention and tensor fusion for refining modality-specific representations [26]. Pathak et al. (2025) utilized dual-dimension attention with Swin Transformers to capture both global and local modality relationships for improved structural segmentation quality [27]. Hu et al. (2025) adopted a dual-domain attention mechanism with deformable alignment to fuse features from MRI, PET, and SPECT, enabling stronger multimodal dependency modeling [28].

The transformer-based architectures improve feature representation but often rely on global fusion mechanisms that do not assign spatially adaptive importance to each MRI modality. Consequently, their ability to accurately segment heterogeneous tumor regions may be limited. Table 1 summarizes key studies published between 2021 and 2025 that explore different feature fusion strategies, ranging from handcrafted to DL models and transformer-based models for BTS.

Table 1. Literature Review (2018-2025) on BTS Using Multimodal Fusion Techniques

Author & Year	Dataset	Fusion Strategy	Architecture	Findings
He et al. [11], 2021	BraTS 2019	Multi-scale attention feature fusion	MAFF-ResUNet	Improved TC and ET segmentation accuracy.
Ranjbarzadeh et al. [12], 2021	BraTS 2018	Distance-wise attention feature fusion	Cascaded CNN with DWA	Improved tumor-region localization.
Huang et al. [13], 2021	BraTS 2015	Multi-scale feature fusion	FEN with MSFFN	Improved semantic and boundary representation
Zhao et al. [14],	BraTS 2020	Hybrid attention	MM-UNet	Enhanced

2022		fusion with DCB		multimodal feature fusion
Zhu et al. [16], 2023	BraTS 2018–2020	Graph-based semantic-edge fusion	Swin Transformer integrated with Edge-aware Semantic Attention and Multi-level Feature Integration Block	Dice 88.2%, HD95 3.92
Zhang et al. [17], 2023	BraTS 2021	Residual multimodal fusion	Multi-branch U-Net	Dice: 83.3% (ET), 89.1% (TC), 91.4% (WT)
Zhou et al. [19], 2023	BraTS 2018	Multi-task modality-aware fusion	Multi-task fusion network consisting of CMFM, MSFM and SC-LFLM	Dice 84.1%; robust to missing modalities
Nizamani et al. [20], 2023	BraTS 2020, MSD	Hybrid U-Net–Transformer fusion	FE-UT variants	Dice 99.6%, Accuracy 99.7%
Zheng et al. [23], 2023	BraTS 2020	Intra- and inter-modality attention fusion	CMMFNet	Dice: 91.1% (WT), 86.5% (TC), 81.3% (ET)
Ullah et al. [21], 2024	BraTS	Handcrafted–CNN feature fusion	GCNN an ensemble of CSPCNN and MRIPCNN	Dice: 0.87 (WT), 0.79 (TC), 0.75 (ET)
Zhu et al.[24], 2024	BraTS 2020, 2021	Sparse Dynamic Attention with Edge Feature Fusion	SDV-TUNet	Dice: 93.1% (WT), 91.0% (TC), 87.6% (ET)
Zhou et al. [25], 2024	BraTS 2018, 2019	Multimodal attention with DRL and RCL	Multi-module fusion network	Dice 84.2%, HD95 4.1 mm

DL methods have significantly improved multimodal BTS. Early fusion models combine MRI modalities at the input stage, which often leads to the loss of modality-specific semantic information. Attention-based and transformer-enhanced fusion architectures improve feature representation but typically rely on global or channel-level weighting and lack spatially adaptive modality importance. Several approaches also employ shared encoders for all modalities, which limit modality-specific learning and reduce the ability to distinguish tumor subregions. In addition, transformer-based frameworks often introduce high computational complexity and reduce interpretability. Consequently, many existing approaches do not adequately preserve modality-specific information while adapting fusion to localized tumor contexts. Therefore, there is a need for an efficient architecture that enables voxel-wise modality-adaptive fusion while improving boundary precision and segmentation accuracy in heterogeneous tumor regions.

3. Proposed Methodology

This study proposes MAAF-Net, a 3D CNN architecture designed for multimodal BTS. The proposed model addresses the challenges of uniform fusion, in which all modalities are considered equally, irrespective of their different clinical relevancies. In MAAF-Net, all MRI modalities are processed independently through dedicated modality specific encoders to preserve their unique features. The extracted multi-scale features are then passed to the MAM, which learns the relative importance of each modality based on global contextual information. The MAM allocates adaptive weights to modality-specific features and enables the proposed framework to focus clinically relevant modalities. The weighted features are fused and passed to a shared decoder, where multi-scale supervision is applied to improve segmentation

accuracy. Figure 1 shows the MAAF-Net architecture, including the four modality-specific encoder branches, the attention-guided fusion through MAM, and the decoder responsible for generating the final BTS maps.

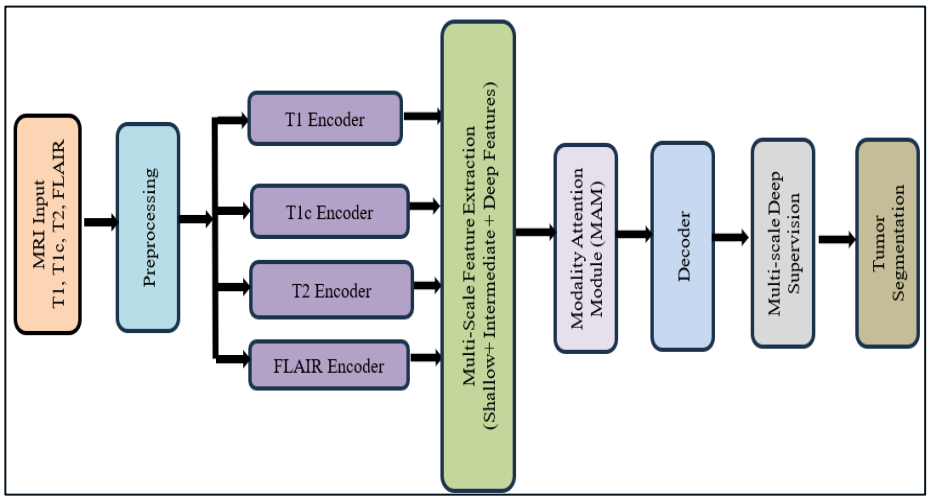

Figure 1. Flow Diagram of the Proposed MAAF-Net

3.1 BraTS 2024 Dataset

The study uses the BraTS 2024 benchmark dataset, which is publicly available for the MICCAI 2024 BraTS Challenge [29]. It contains multi-institutional and multi-parametric MRI images of patients diagnosed with gliomas, including both High-Grade Glioma (HGG) and Low-Grade Glioma (LGG). All subjects include four co-registered MRI modalities as shown in Figure 2. The images are processed using skull-stripping, co-registration to a common anatomical template, and resampling to an isotropic resolution of 1 mm^3. The dataset contains approximately 1,080 patient cases in the training set. In this study, the dataset is divided using a five-fold cross-validation protocol. For each fold, 864 subjects (80%) were used for training and 216 subjects (20%) were used for validation.

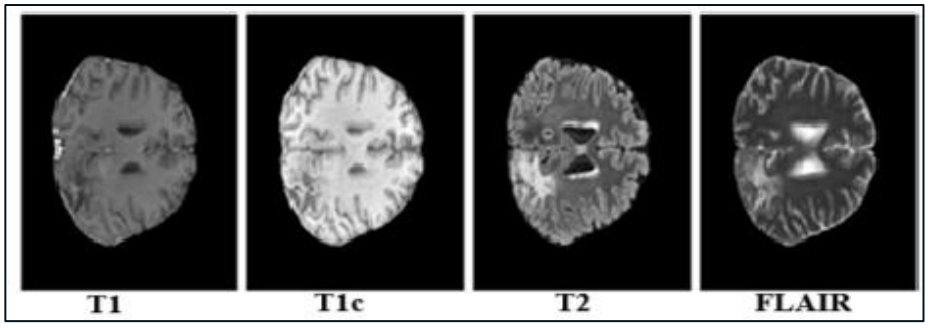

Figure 2. Complementary MRI Modalities for Brain Tumor Analysis

3.2 Preprocessing

Preprocessing is essential for reducing inter-subject variability and increasing the robustness of DL models. All input images from the BraTS 2024 dataset underwent standardized preprocessing to ensure spatial and intensity consistency across cases and modalities. The preprocessing pipeline consists of label relabeling, image resizing, intensity normalization, and data augmentation. The preprocessing pipeline used in this work is illustrated in Figure 3.

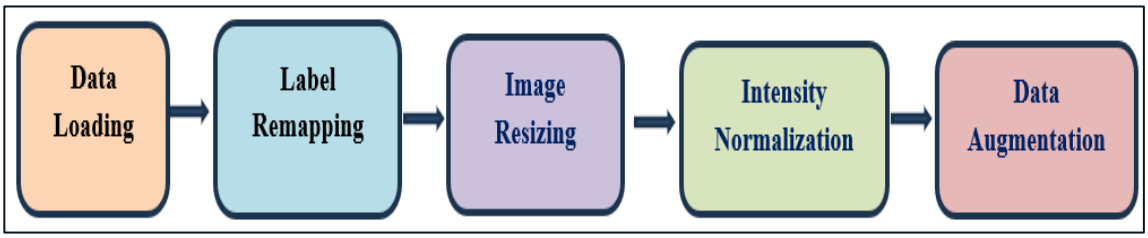

Figure 3. Preprocessing Pipeline Used in Proposed Methodology

3.2.1 Label Reindexing

The original BraTS 2024 dataset contains non-contiguous label values for different tumor subregions. The Whole Tumor (WT) region is not included in the proposed work, as it does not represent biologically active tumor tissue. Additionally, WT occupies a large spatial area, shows stronger contrast, and makes it easier to segment. As a result, the inclusion of WT may inflate Dice scores without accurately reflecting the model's capability to segment ET and TC.

Thus, labels are reindexed into a contiguous class for optimization and compatibility. Label 0 represents background, label 2 is reassigned to 1 to represent TC, and label 4 is reassigned to 2 for ET. Therefore, the study aims to focus on the TC and ET regions, as these are clinically relevant and challenging.

3.2.2 Resizing

The original MRI images (240 × 240 × 155) are resized to 1283 for reducing memory consumption. It also helps in efficient training on standard GPUs. Training with full-resolution volumes requires substantially higher memory because the computational cost of 3D CNNs increases with the volumetric size of the input. The resized resolution preserves sufficient anatomical information for accurate segmentation of TC and ET while maintaining the overall spatial context of the tumor. In addition, resizing standardizes the input dimensions across all subjects, allowing consistent batch-wise training and stable model convergence. Patch-based training is not adopted because it may remove important global spatial context which is necessary for capturing tumor structure and boundary relationships within the whole brain volume.

3.2.3 Intensity Normalization (Z-Score Normalization)

Due to inter-scanner and inter-subject variability, raw MRI intensities are not inherently standardized. To address this issue, z-score normalization is applied on a per-modality, per-subject basis. It ensures zero mean and unit variance for each modality and improves training while reducing intensity variations across scanners.

3.2.4 Data Augmentation

A set of online data augmentation techniques is applied during training with a probability of 80% for model generalization and reducing overfitting. These include spatial transformations such as random flipping along the axial, coronal, and sagittal planes, as well as random 3D rotations within ±10o. To simulate anatomical variability, elastic deformation is applied with parameters $\alpha = 90$ and $\sigma = 9$. In addition, image brightness is randomly adjusted within a small range of ±0.1, and Gaussian noise is added with values ranging between 0 and

0.05 to increase robustness to intensity variations. These augmentations improve robustness to anatomical variability and scanner-induced intensity differences.

3.3 Proposed Model

A standard 3D U-Net model for volumetric medical image segmentation is utilized as the baseline. The model follows an encoder–decoder architecture incorporating symmetric skip connections to recover spatial features lost in downsampling. It takes multimodal MRI inputs and produces a voxel-wise segmentation map. The architecture of the baseline 3D U-Net model is given in Figure 4.

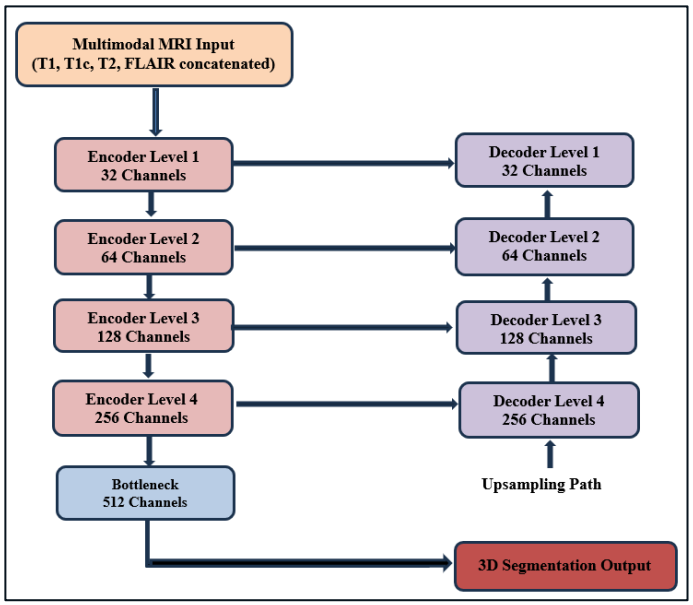

Figure 4. 3D U-Net Model Architecture

3.3.1 Proposed MAAF-Net

The proposed MAAF-Net consists of three stages:

1. Modality specific feature extraction using modality specific encoders to extract features from each modality.
2. Global context modeling and modality adaptive fusion to assign dynamic weights to each modality based on attention weights.
3. Segmentation reconstruction to generate voxel-wise tumor segmentation maps.

A. Modality-Specific Feature Extraction

The proposed MAAF-Net employs four independent MSEnc. This design preserves the distinct structural and intensity characteristics of each modality before feature fusion. Each encoder follows a five-stage hierarchical structure. At each resolution level, two consecutive 3D convolutional layers with a kernel size of 3×3×3, stride of 1, and padding of 1 are applied, along with instance normalization and ReLU activation. Downsampling between successive levels is performed using 2^3 max pooling with a stride of 2. The complete configuration of the modality specific encoder is given in Table 2.

Table 2. Architecture of Modality-Specific Encoder Used in MAAF-Net

Stage	Channels	#Conv Layers	Kernel Size	Stride	Padding	Parameters
Level 1	1 → 32	2	3x3x3	1	1	28,576
Max Pool	32→32	--	2x2x2	2	0	0
Level 2	32→64	2	3x3x3	1	1	166016
Max Pool	64→64	--	2x2x2	2	0	0
Level 3	64→128	2	3x3x3	1	1	663808
Max Pool	128→128	--	2x2x2	2	0	0
Level 4	128→256	2	3x3x3	1	1	2654720
Max Pool	256→256	--	2x2x2	2	0	0
Bottleneck	256→512	2	3x3x3	1	1	10617856
Total (Per Encoder)	--	10 Conv3D	--	--	--	14130976

Since the model contains four independent MSEnc, the sum of learnable parameters of all encoders is approximately 56.52 million. Each modality is passed independently through its dedicated encoder consisting of stacked convolutional blocks. The features at layer I are extracted as given in eq. (1):

$$F_{M_i}^{(\iota)} = \mathrm{ReLU}(\mathrm{IN}\left(W^{(\iota)} * F_{M_i}^{(\iota-1)} + b^{(\iota)}\right)) \quad (1)$$

Where:

- $F_{M_i}^{(\iota)}$: Feature map for modality M_i at layer ι
- $W^{(\iota)}$: Learnable convolution weights
- ReLU: Non-linearity
- ι: Layer index

Each encoder produces modality-specific feature map as shown in eq. 2:

$$F_{T1},\ F_{T1C},\ F_{T2}\ ,\ F_{FLAIR} \in \mathbb{R}^{C\prime \times D\prime \times H\prime \times W\prime} \quad (2)$$

Where:

- C': Number of output channels
- D', H', W: Down sampled spatial dimensions

Each modality-specific encoder is designed to extract multi-scale features across several resolution levels. The encoder progressively extracts hierarchical feature representations as spatial resolution decreases while channel depth increases. At each scale *ι*th feature map has the shape given in eq. 3:

$$F_{M_i}^{(\iota)} \in \mathbb{R}^{C_\iota \times D_\iota \times H_\iota \times W_\iota} \quad (3)$$

Where:

- C_ι: feature depth increases with depth
- Dι, Hι, Wι: spatial dimensions decrease progressively.

These multi-scale modality-specific features are later aggregated using a MAM. It generates a unified representation that captures both complementary and discriminative information across modalities.

B. Global Context Extraction and Modality Attention Module (MAM)

Global average pooling is applied to each modality-specific feature map to extract the global context of all modalities. These global features are then given to a shared multi-layer perceptron (MLP) to produce modality-specific importance scores. The scores are normalized by a softmax function. Then feature maps are scaled with the help of normalized weights before fusion. This process enables the model to dynamically prioritize modalities like T1c for ET regions and FLAIR for edema.

The MAM performs adaptive fusion of modality-specific features. In multimodal MRI, different sequences contribute unequally to tumor identification. Clinically, T1c images are particularly useful for identifying ET, while FLAIR emphasizes edema. To capture this variability, the MAM computes attention weights to represent the relative importance of each modality. The flow diagram of the MAM architecture is given in Figure 5 and all the components are explained as follows:

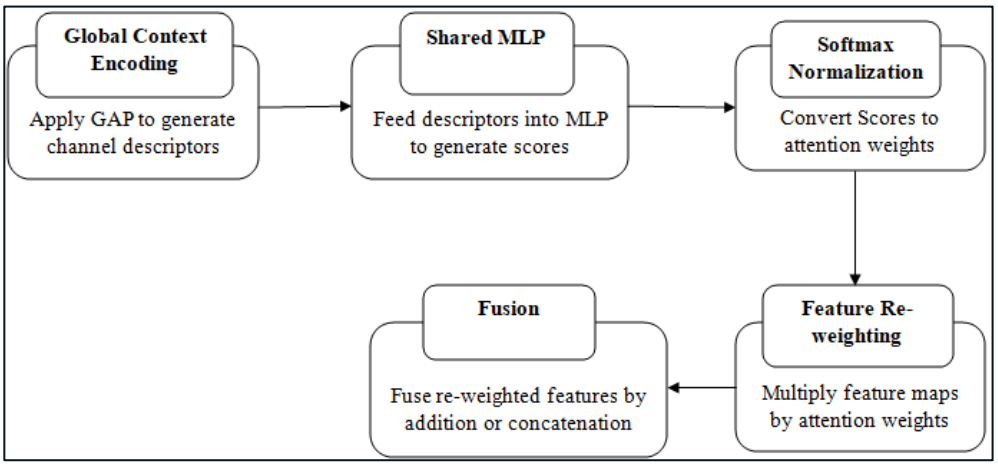

Figure 5. Flow Diagram of Modality Attention Module

- **Global Context Encoding:** To capture the global importance of each modality, Global Average Pooling (GAP) is applied over the spatial dimensions of each feature map. $\mathbb{g}_i$ represent channel-wise global descriptor.
- **Attention Weight Computation:** Each global descriptor is passed through a shared multi-layer perceptron comprising two fully connected layers (FCL) with ReLU activation to generate a modality importance score S_ii. The MLP uses a channel reduction ratio $r = 16$, where the first FCL decreases the channel dimension from C to C/r, followed by ReLU activation and a second FCL that returns the scalar attention score.
- **Softmax Normalization:** Softmax function is used to normalize the modality scores Si to produce attention weights α_{i}. The attention scores are generated within the Modality Attention Module using global descriptors extracted from modality-specific encoder features. These scores are then applied to reweight the corresponding modality feature maps before the fusion stage.

- **Feature Re-weighting and Fusion:** Each modality feature map is then reweighted by its attention score as shown in eq. 4:

$$\tilde{F} = \alpha_{\mathrm{i}} \,.\, \mathrm{F}_i \tag{4}$$

The reweighted features are fused using either element-wise summation or channel-wise concatenation as given in eq. 5. Addition preserves dimensionality and reduces computational cost. Concatenation retains modality-specific separation for downstream processing.

$$F_{fused} = \begin{cases} \sum_{i=1}^{4} \tilde{F}_i & (Addition) \\ Concat\left(\tilde{F}_1, \tilde{F}_2, \tilde{F}_3, \tilde{F}_4\right) & (Concatenation) \end{cases} \tag{5}$$

The proposed MAM differs from the SE-block attention mechanism in the level at which attention is computed. While the SE-block performs channel-wise recalibration within a single feature map using sigmoid activation, the proposed MAM computes attention across modalities using softmax normalization. This allows the network to dynamically assign different importance to MRI modalities during feature fusion.

C. Segmentation Reconstruction

The decoder reconstructs a full-resolution segmentation map from fused features. Multi-resolution skip connections send encoder features to the decoder, preserving spatial details during upsampling as given in eq. 6:

$$F_{skip}^{(\iota)} = \sum_{\mathrm{i}=1}^{4} \alpha_{\mathrm{i}}^{(\iota)} . \mathrm{F}_{\mathrm{i}}^{(\iota)} \tag{6}$$

Where $\mathrm{F}_{\mathrm{i}}^{(\iota)}$is the encoder output from modality i at level ι, and $\alpha_{\mathrm{i}}^{(\iota)}$is the learned attention weight for that modality at the same level.

D. Multi-Scale Supervision

Multi-scale supervision is introduced to improve gradient flow during training. Auxiliary segmentation outputs are generated at intermediate decoder stages and supervised using an auxiliary loss.

E. Loss Function

The proposed model utilizes a composite loss function, in which Dice loss and Cross-Entropy loss are combined. Dice loss mitigates class imbalance, while Cross-Entropy improves voxel-level classification stability. Let $\mathrm{P_c(x)}$ denote the predicted probability for class $c \in \{0,1,2\}$ at voxel x and $\mathrm{G_c(x)}$ denote the corresponding ground truth label.

Dice:

$$\mathcal{L}_{Dice}^{(c)} = 1 - \frac{2\sum_x P_c(x) G_c(x) + \epsilon}{\sum_x P_c(x)^2 + \sum_x G_c(x)^2 + \epsilon} \tag{7}$$

Where $\epsilon = 10^{-5}$ to prevent division error. The total Dice Loss across all tumor classes is:

$$\mathcal{L}_{Dice} = \frac{1}{C}\sum_{c=1}^{C}\mathcal{L}_{Dice}^{(c)} \tag{8}$$

Cross-Entropy Loss:

$$\mathcal{L}_{CE} = -\sum_{x}\sum_{c=1}^{C} G_c(x)\log\left(P_c(x)\right) \tag{9}$$

Total Loss:

$$\mathcal{L}_{main} = \lambda_1\mathcal{L}_{Dice} + \lambda_2\mathcal{L}_{CE} \tag{10}$$

Where $\lambda_1 = 0.7$ and $\lambda_2 = 0.3$

More weight is given to dice loss due to ET class imbalance.

Auxiliary Loss: To support better learning at different resolution levels, auxiliary losses are applied to intermediate decoder outputs. This improves gradient propagation, stabilizes the training process, and helps to capture meaningful features at different scales.

Upsampled intermediate prediction:

$$\widehat{S^k} = Upsample(s^{(k)}) \tag{11}$$

Auxiliary Loss at Scale k:

$$\mathcal{L}_{aux}^{(k)} = \lambda_1\mathcal{L}_{Dice}^{(k)} + \lambda_2\mathcal{L}_{CE}^{(k)} \tag{12}$$

Multi-Scale Supervision:

$$\mathcal{L}_{MS} = \sum_{k=1}^{K}\alpha_k\,\mathcal{L}_{aux}^{(k)} \tag{13}$$

Final total loss:

$$\mathcal{L}_{total} = \mathcal{L}_{main} + \mathcal{L}_{MS} \tag{14}$$

4. Results and Observations

4.1 Experimental Setup

The MAAF-Net is trained and validated on the BraTS 2024 benchmark dataset. All models are trained and validated with the same configuration. The model performance is evaluated using the Dice Score, HD95, sensitivity, and specificity. The hyperparameters used for implementation are given in Table 3.

Table 3. Training Hyperparameters Used for Model Implementation

Parameter	Value
Optimizer	Adam
Initial Learning Rate	$1x10^{-4}$
Batch Size	6
Epochs	50
Input Size	128x128x128

Loss Function	Dice, Cross Entropy (CE)
Dice Weight	0.7
CE Weight	0.3
Framework	PyTorch (2.6.0)
GPU	NVIDIA A100

4.2 Performance Comparison on BraTS 2024

The quantitative comparison of the proposed MAAF-Net with baseline fusion strategies is presented in Figure 6. It has been observed that the proposed model attains the best performance for both TC and ET regions. For TC segmentation, Dice scores have been improved by approximately 5–6% compared to early and mid-fusion strategies. Additionally, the proposed model shows a reduction in HD95 by 40% demonstrating more accurate boundary delineation. Similarly, for ET, MAAF-Net improves Dice scores by to 8% for early fusion and 3% for attention-based fusion. Experimental results further demonstrate improvements in sensitivity and specificity indicating that the MAAF-Net effectively balances tumor detection and false positive suppression. The performance gain in ET segmentation is due to the strong contrast of ET regions in T1c images, captured by the MAM to assign higher weights to this modality during feature fusion. In contrast, TC segmentation depends on features extracted from multiple modalities and receives smaller benefits from MAM.

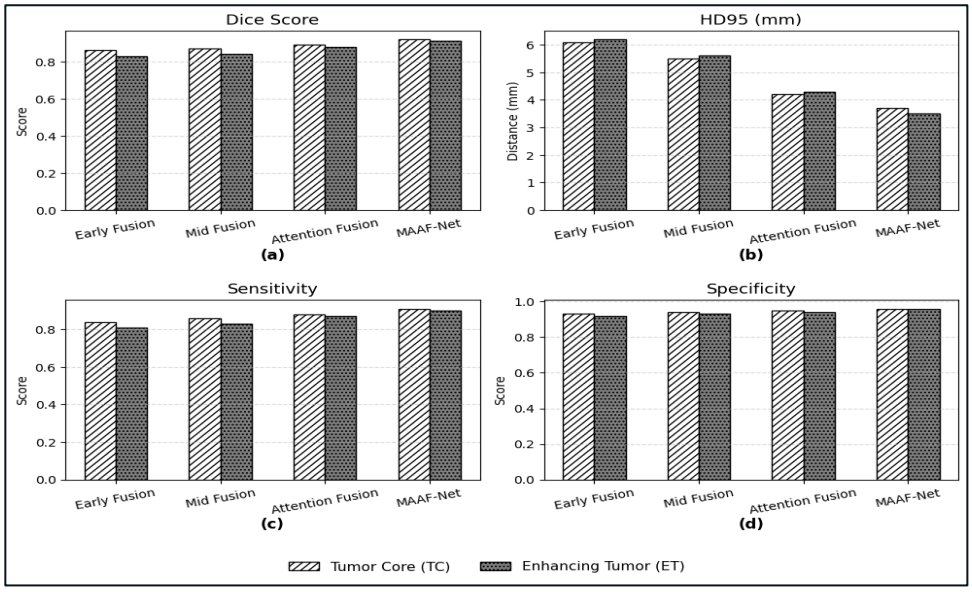

Figure 6. (a–d) Evaluation of Fusion Strategies for TC and ET, (a) Dice Score, (b) HD95, (c) Sensitivity, and (d) Specificity

4.3 Qualitative Segmentation Analysis

The qualitative assessment of different fusion strategies is presented in Figure 7. It has been observed that early and mid-level fusion models produce over-segmentation. Attention-based fusion enhances localization but fails to analyze the full tumor structure. In contrast, the proposed model visualizes results that closely match the ground truth, with improved boundaries and segmentation of ET and TC regions across axial, sagittal, and coronal views.

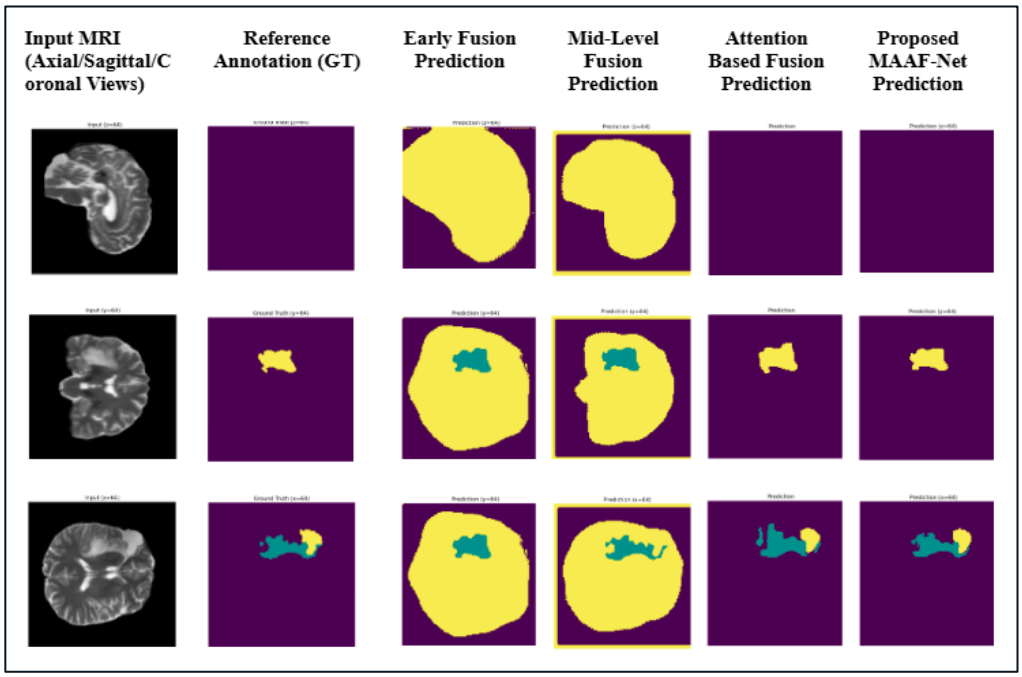

Figure 7. Qualitative Comparison of Different Fusion Strategies

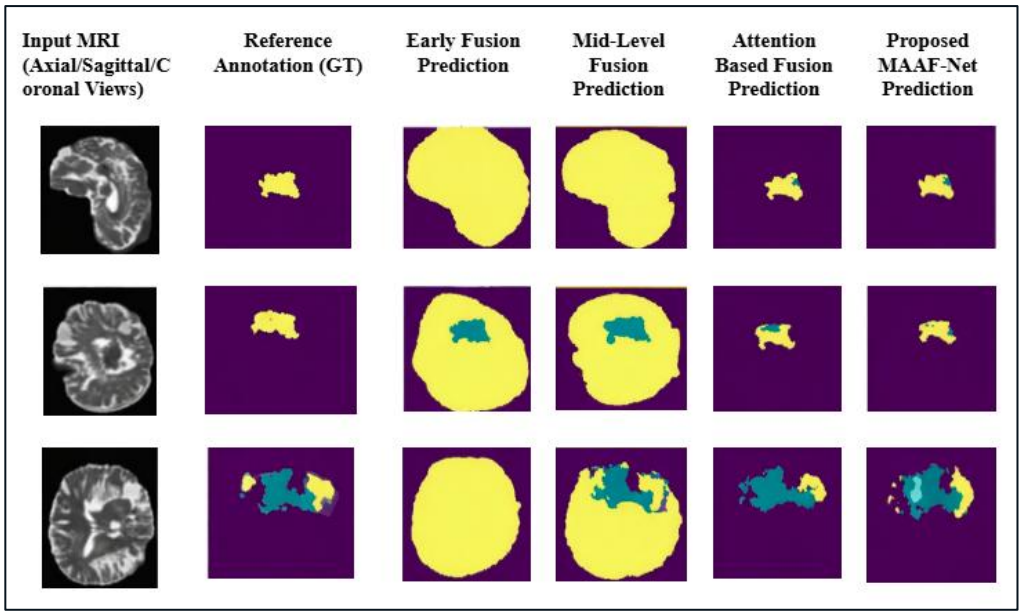

Figure 8. Failure Case Analysis

A failure case is presented in Figure 8 which illustrates the limitations of different fusion strategies, producing damaged tumor boundaries and over-segmentation, causing a decrease in overall Dice Score and HD95.

4.4 Ablation Study

The ablation study is performed by gradually removing the MAM and the MSEnc as demonstrated in Table 4. It has been analyzed from the table that in the first variant, the attention mechanism has been removed. The extracted features from different MRI modalities are fused together using concatenation. This configuration attains Dice scores of 0.86 and 0.80 for TC and ET, respectively. In the second variant, the MSEnc is replaced with a shared

encoder. This variant attains Dice scores of 0.87 for TC and 0.82 for ET, demonstrating that independent encoders preserve modality-specific information in a better way.

Overall, the exclusion of the attention module degraded performance, particularly for ET segmentation. The complete MAAF-Net produces the best results and validates the proposed model architecture for both modality-specific feature extraction and adaptive attention fusion.

Table 4. Impact of MSEnc and Attention Fusion on TC and ET Segmentation

Model Variant	Dice (TC)	Dice (ET)	HD95 (TC)	HD95 (ET)	Sensitivity (TC)	Sensitivity (ET)	Specificity (TC)	Specificity (ET)
Without Modality Attention (Concat only)	0.86	0.80	5.6	6.9	0.84	0.78	0.91	0.92
Without MSEnc	0.87	0.82	4.8	5.4	0.86	0.81	0.93	0.93
Full MAAF-Net (Proposed)	0.92	0.91	3.7	3.5	0.91	0.90	0.96	0.96

4.5 Computational Efficiency

MAAF-Net contains approximately 56.52 million trainable parameters due to the inclusion of MSEnc and attention-based fusion modules. Despite this architectural complexity, the model maintains efficient inference, requiring an average of 1.9 seconds per 3D volume of size 128 × 128 × 128. Peak GPU memory usage is limited to approximately 3.8 GB.

5. Conclusion

This study proposes MAAF-Net, a modality-adaptive attention fusion framework for multimodal BTS. The proposed model combines MSEnc with a MAM within a 3D U-Net model. The MAAF-Net effectively combines information from multiple MRI modalities and performs segmentation of the TC and ET region. Experimental findings on the BraTS 2024 dataset show improved segmentation, with Dice scores of 0.92 for TC and 0.91 for ET, and HD95 values of 3.7 mm and 3.5 mm, respectively. The findings demonstrate that modality-specific feature extraction and adaptive attention-based fusion improve tumor boundary delineation. Although the model provides promising results, further evaluation on multi-institutional datasets and missing modalities is still needed. Thus, future work focuses on extending the framework to handle missing or incomplete modalities to enhance its clinical applicability. Additionally, efforts will be made to optimize its computational complexity by using a lightweight architecture or a parameter-sharing mechanism.

References

[1] Louis, David N., Arie Perry, Guido Reifenberger, Andreas Von Deimling, Dominique Figarella-Branger, Webster K. Cavenee, Hiroko Ohgaki, Otmar D. Wiestler, Paul Kleihues, and David W. Ellison. "The 2016 World Health Organization Classification of Tumors of the Central Nervous System: A Summary." Acta neuropathologica 131, no. 6 (2016): 803-820.

[2] Menze, Bjoern H., Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren et al. "The Multimodal Brain Tumor Image

Segmentation Benchmark (BRATS)." IEEE transactions on medical imaging 34, no. 10 (2014): 1993-2024.

[3] Litjens, Geert, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I. Sánchez. "A Survey on Deep Learning in Medical Image Analysis." Medical image analysis 42 (2017): 60-88.

[4] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." In International Conference on Medical image computing and computer-assisted intervention, Cham: Springer international publishing, 2015, 234-241.

[5] Bakas, Spyridon, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. "Advancing the Cancer Genome Atlas Glioma MRI Collections with Expert Segmentation Labels and Radiomic Features." Scientific data 4, no. 1 (2017): 170117.

[6] Zhou, Zongwei, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. "Unet++: A Nested U-Net Architecture for Medical Image Segmentation." In International workshop on deep learning in medical image analysis, Cham: Springer International Publishing, 2018, 3-11

[7] Havaei, Mohammad, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. "Brain Tumor Segmentation with Deep Neural Networks." Medical image analysis 35 (2017): 18-31.

[8] Pereira, Sérgio, Adriano Pinto, Victor Alves, and Carlos A. Silva. "Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images." IEEE transactions on medical imaging 35, no. 5 (2016): 1240-1251.

[9] Guo, Lu, Ping Wang, Ranran Sun, Chengwen Yang, Ning Zhang, Yu Guo, and Yuanming Feng. "A Fuzzy Feature Fusion Method for Auto-Segmentation of Gliomas with Multi-Modality Diffusion and Perfusion Magnetic Resonance Images in Radiotherapy." Scientific reports 8, no. 1 (2018): 3231.

[10] Hagargi, Padmanjali A., and D. Shubhangi. "Brain Tumor MR Image Fusion Using Most Dominant Features Extraction from Wavelet and Curvelet Transforms." Brain 5, no. 05 (2018): 33-38.

[11] He, Xueqin, Wenjie Xu, Jane Yang, Jianyao Mao, Sifang Chen, and Zhanxiang Wang. "Deep Convolutional Neural Network with a Multi-Scale Attention Feature Fusion Module for Segmentation of Multimodal Brain Tumor." Frontiers in Neuroscience 15 (2021): 782968.

[12] Ranjbarzadeh, Ramin, Abbas Bagherian Kasgari, Saeid Jafarzadeh Ghoushchi, Shokofeh Anari, Maryam Naseri, and Malika Bendechache. "Brain Tumor Segmentation Based on Deep Learning and an Attention Mechanism Using MRI multi-Modalities Brain Images." Scientific reports 11, no. 1 (2021): 10930.

[13] Huang, Daobin, Minghui Wang, Ling Zhang, Haichun Li, Minquan Ye, and Ao Li. "Learning Rich Features with Hybrid Loss for Brain Tumor Segmentation." BMC Medical Informatics and Decision Making 21, no. Suppl 2 (2021): 63.

[14] Zhang, Dingwen, Guohai Huang, Qiang Zhang, Jungong Han, Junwei Han, and Yizhou Yu. "Cross-Modality Deep Feature Learning for Brain Tumor Segmentation." Pattern Recognition 110 (2021): 107562.

[15] Zhao, Liang, Jiajun Ma, Yu Shao, Chaoran Jia, Jingyuan Zhao, and Hong Yuan. "MM-UNet: A Multimodality Brain Tumor Segmentation Network in MRI Images." Frontiers in oncology 12 (2022): 950706.

[16] Zhu, Zhiqin, Xianyu He, Guanqiu Qi, Yuanyuan Li, Baisen Cong, and Yu Liu. "Brain Tumor Segmentation Based on the Fusion of Deep Semantics and Edge Information in Multimodal MRI." Information Fusion 91 (2023): 376-387.

[17] Zhang, Guying, Jia Zhou, Guanghua He, and Hancan Zhu. "Deep Fusion of Multi-Modal Features for Brain Tumor Image Segmentation." Heliyon 9, no. 8 (2023). e19266.

[18] Yang, Hengyi, Tao Zhou, Yi Zhou, Yizhe Zhang, and Huazhu Fu. "Flexible Fusion Network for Multi-Modal Brain Tumor Segmentation." IEEE Journal of Biomedical and Health Informatics 27, no. 7 (2023): 3349-3359.

[19] Zhou, Tongxue. "Feature Fusion and Latent Feature Learning Guided Brain Tumor Segmentation and Missing Modality Recovery Network." Pattern Recognition 141 (2023): 109665.

[20] Nizamani, Abdul Haseeb, Zhigang Chen, Ahsan Ahmed Nizamani, and Kashif Shaheed. "Feature-Enhanced Fusion of U-NET-Based Improved Brain Tumor Images Segmentation." Journal of Cloud Computing 12, no. 1 (2023): 170.

[21] Ullah, Faizan, Muhammad Nadeem, Mohammad Abrar, Muna Al-Razgan, Taha Alfakih, Farhan Amin, and Abdu Salam. "Brain Tumor Segmentation from MRI Images Using Handcrafted Convolutional Neural Network." Diagnostics 13, no. 16 (2023): 2650.

[22] Khan, Muhammad Attique, Reham R. Mostafa, Yu-Dong Zhang, Jamel Baili, Majed Alhaisoni, Usman Tariq, Junaid Ali Khan, Ye Jin Kim, and Jaehyuk Cha. "Deep-Net: Fine-Tuned Deep Neural Network Multi-Features Fusion for Brain Tumor Recognition." Computers, Materials and Continua 76, no. 3 (2023): 3029-3047.

[23] Zheng, Jiangpeng, Fan Shi, Meng Zhao, Chen Jia, and Congcong Wang. "Learning Intra-Inter-Modality Complementary for Brain Tumor Segmentation." Multimedia Systems 29, no. 6 (2023): 3771-3780.

[24] Zhu, Zhiqin, Mengwei Sun, Guanqiu Qi, Yuanyuan Li, Xinbo Gao, and Yu Liu. "Sparse Dynamic Volume TransUNet with multi-Level Edge Fusion for Brain Tumor Segmentation." Computers in Biology and Medicine 172 (2024): 108284.

[25] Zhou, Tongxue. "Multi-Modal Brain Tumor Segmentation via Disentangled Representation Learning and Region-Aware Contrastive Learning." Pattern Recognition 149 (2024): 110282.

[26] Wang, Yu, Juan Xu, Yucheng Guan, Faizan Ahmad, Tariq Mahmood, and Amjad Rehman. "MSegNet: A Multi-View Coupled Cross-Modal Attention Model for Enhanced MRI Brain Tumor Segmentation." International Journal of Computational Intelligence Systems 18, no. 1 (2025): 63.

[27] Pathak, Disha Mohini, and Tribhuwan Kumar Tewari. "M3IFT2. 0: Multi-Modal Medical Image Fusion Framework with Vision Transformer Features and Attention Integration." International Journal of Information Technology 17, no. 5 (2025): 2905-2918.

[28] Hu, Tianqing, Xiaofei Nan, Xiabing Zhou, Yu Shen, and Qinglei Zhou. "A Dual-Stream Feature Decomposition Network with Weight Transformation for Multi-Modality Image Fusion." Scientific Reports 15, no. 1 (2025): 7467.

[29] de Verdier, Maria Correia, Rachit Saluja, Louis Gagnon, Dominic LaBella, Ujjwall Baid, Nourel Hoda Tahon, Martha Foltyn-Dumitru et al. "The 2024 Brain Tumor Segmentation (Brats) Challenge: Glioma Segmentation on Post-Treatment MRI." arXiv preprint arXiv:2405.18368 (2024).