

Article

Brain Tumor Segmentation with Contextual Transformer-Based U-Net

Shakhnoza Muksimova¹, Jushkin Baltaev² and Young Im Cho^{1,*} ¹ Department of Computer Engineering, Gachon University, Sujeong-gu, Seongnam-si 461701, Republic of Korea; muksimova1986@tsue.uz² Department of Information Systems and Technologies, Tashkent State University of Economics, Tashkent 100066, Uzbekistan; j_baltayev@tsue.uz

* Correspondence: yicho@gachon.ac.kr

Abstract

Presently, the segmentation of brain tumors from magnetic resonance imaging (MRI) scans is a very important challenge in the medical area, and it has a huge impact on correct diagnosis, efficient treatment planning, and patient prognosis. We present here the Contextual Transformer U-Net (CT-UNet), a novel deep learning approach that can significantly increase the accuracy and speed of brain tumor segmentation. The CT-UNet method features Transformer blocks embedded in a U-Net layout that extracts the most important contextual information across different types of MRI sequences, thereby drastically refining the delineation of tumor regions. We have tested CT-UNet on the Brain Tumor Segmentation (BraTS) challenge dataset that includes a large variety of tumor types, localization, and progression stages. To check the model's performance, we used the Dice coefficient, sensitivity, specificity, precision, and Hausdorff distance metrics. The findings from our experiments demonstrate that CT-UNet has a substantial advantage over the classical segmentation model, and the 0.92 Dice coefficient it has achieved testifies to its state-of-the-art tumor localization in terms of both extent and form. Besides that, CT-UNet has achieved a very high sensitivity (0.90) and specificity (0.94); thus, it has been perfectly capable of discriminating tumor from non-tumor tissues. Spatial accuracy has also been improved significantly, as can be seen from the 7.5 mm Hausdorff distance achieved by this model, which means it can closely replicate the given tumor boundaries. By employing dynamic modality fusion and incorporating the Transformer mechanism into the established U-Net architecture, we have raised the bar for brain tumor segmentation. Our solution paves the way for another breakthrough in medical imaging technologies. CT-UNet not only speeds up the workflow of radiologists but also facilitates more targeted therapeutic strategies that may result in better patient care and prognosis. Yet the main goal of this work is to provide a basis for future studies that can consider incorporating deep learning methods in a routine clinical setting, thus paving the way for healthcare providers to benefit from both technical and clinical advantages.



Received: 6 January 2026

Revised: 5 February 2026

Accepted: 9 February 2026

Published: 12 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: brain tumor segmentation; Transformer networks; multimodal medical imaging; dynamic modality fusion; medical image segmentation

1. Introduction

One of the fundamental challenges in the medical imaging field is the task of delineating brain tumors from magnetic resonance imaging (MRI) data with a very high

degree of accuracy [1]. The varied types of brain tumors and their blurred, often irregular boundaries make it hard even for the most sophisticated algorithms to perform the segmentation accurately [2]. Segmentation that is both precise and dependable is the foundation for major medical decisions; thus, it is crucial not only for diagnosis but also for treatment planning and monitoring the effectiveness of therapies [3]. Because brain tumors are very diverse in their shapes, sizes, and locations, the accuracy of segmentation has a direct impact on the therapies chosen and prognosis made [4]. Even though medical imaging has seen a fast and great development, segmenting brain tumors through an automated method is still a problem [5]. Using ordinary CNNs, conventional methods especially have a hard time effectively merging and integrating information from different MRI modalities [6]. In addition, because of the inability to efficiently model long-term relationships and the challenge of dealing with highly varied tumor manifestations across patients and imaging conditions [7], these methods mostly do not capture the essential fine details needed for accurate tumor localization [8]. This study is mainly focused on the creation of a deep learning model that combines the powerful features of CT-UNet with a Dynamic Modality Fusion (DMF) strategy. The proposed method is expected to bring about a significant improvement in the segmentation of brain tumors from MRI data. The crux of the approach lies in employing Transformer networks in a U-Net setup, thus enhancing the overall image understanding of MRI by also being able to locally reweigh the contribution of different modalities present in the image data in an adaptive manner [9]. We present a breakthrough in brain tumor segmentation through the integration of advanced Transformer-based features with the well-established U-Net model architecture to overcome current methodological shortcomings [10]. This paper lists the multiple aspects of this work that make it a significant contribution to the field.

Our approach, CT-UNet, integrates Transformer blocks inside the U-Net architecture, thus enabling the network to get a visually deeper and more detailed understanding of spatial relationships and feature hierarchies not only of the tumor but also of the neighboring tissues. Our DMF method, unlike fixed fusion methods, can variably combine the different features of various MRI modalities, such as T1-weighted, T2-weighted, and FLAIR sequences, based on their contextual significance to the tumor features, thus focusing on the most relevant information at the time of segmentation. These architectural and procedural modifications enable the model to obtain superior performance in fine and accurate delineation of brain tumors, which is supported by several comprehensive quantitative assessments.

CT-UNet radically integrates the Transformer technology into the U-Net and combines it with a dynamic approach to modal fusion, which has significantly raised the level of medical imaging, especially brain tumor segmentation from MRI data. This work not only handles the main challenges but also acts as a catalyst for the research wave that will heavily improve and broaden the application of machine learning in medical diagnostics.

2. The Literature Survey

The field of medical imaging has undergone enormous changes in the last ten years, especially regarding the segmentation of brain tumors. CNNs have been at the heart of this development. U-Net and other standard segmentation models formed the first milestone because of their well-thought-out structure that includes an encoder–decoder path, as well as skip connections, thus allowing the networks to keep core spatial information at different resolutions [11]. Yet brain tumor segmentation using CNNs still has certain drawbacks. Mainly, it is a result of the inability of the neural networks to adequately process multi-modal MRI data and their insufficient representational capacity to accurately delineate the complex, irregular boundaries of brain tumors [12,13]. Deep learning methods,

mainly Transformer network-based architectures, have surpassed the traditional ways of dealing with these issues. Transformers were initially conceived for language processing tasks, but they are extremely proficient in capturing long dependencies and processing sequential data [14,15]. Their impact on medical imaging is the topic of current research, and initial results indicate that Transformer-based architectures can be a game-changer in the interpretation of spatial relationships and context in the image data [16]. Besides this, features such as attention mechanisms have been merged into CNNs, which have helped the model to focus on the right parts of the image so that the tumor segmentation can be more precise [17–19]. However, despite the promising advances with these technologies, the literature review reveals that there is a big gap when it comes to the proper utilization of multimodal MRI data for brain tumor segmentation.

Existing approaches mostly view different modal sequences as separate inputs to a multi-channel model, or they are concatenated in a way that ignores the unique diagnostic contributions of each modality [20]. Recent segmentation studies have explored alternative directions beyond CNN–Transformer hybrids. For example, HCMNet [21] integrates CNNs with Mamba-based state-space modeling to efficiently capture long-range dependencies in breast ultrasound segmentation, focusing primarily on single-modality efficiency rather than multimodal feature interaction. Other works investigate active learning strategies for segmentation, such as image-similarity-based sample selection for skin lesion analysis [22], which aims to reduce annotation costs rather than improve the segmentation architecture. This line of research is complementary but orthogonal to CT-UNet, which focuses on architectural representation learning under fully supervised settings [23]. While U-Net and its variants are great at detecting local features, they are usually limited by the convolution operation’s small field of view, which makes it difficult for them to perform spatial reasoning of the entire image effectively [24]. In addition, the complexity and heterogeneity of brain tumor shapes require a much more detailed approach to feature extraction and integration than what conventional models generally provide. The fixed feature fusion in current models fails to reflect the dynamic variation in the importance of different modalities, which is dependent on the specific context of the image region under analysis [25]. The highlighted issues point towards the necessity of a novel method that not only dynamically combines multi-modal data but also exploits the features of Transformer networks to better understand the image context globally [26]. Such a method would solve the problems of existing segmentation methods and also pave the way for future medical imaging research directions [27]. While brain tumor segmentation has seen major advances, the accuracy, efficiency, and adaptability of these technologies can still be vastly improved. In this paper, the authors intend to examine the integration of Transformer networks with typical CNN architectures and the use of dynamic multi-modal data fusion approaches, which is a potential development area.

3. Proposed Methodology

CT-UNet draws its inspiration from the usual pattern of CNN–Transformer hybrid architectures; the way it is designed is more of a task-specific architectural reformulation than a simple adaptation of existing models. The authors justify each part of their design in terms of the main challenges in the segmentation of brain tumors from multimodal MRI, which encompass irregular tumor morphology, diffuse boundaries, and contrast changes from one modality to another. It was only the underlying philosophy of the Contextual Transformer fusion scheme that was quite different from that of the existing hybrid methods. The Transformer modules are only located at the bottleneck or, in some cases, completely replace the encoder in most CNN–Transformer segmentation networks. On the other hand, CT-UNet inserts Transformer modules evenly in both the encoder and decoder paths. This

two-way integration allows the network to keep the global contextual information while compressing the features and at the same time globally dependently re-injecting the spatial reconstruction. The brain tumor segmentation application is the primary reason for this design, as it makes it easier to work with very long-range spatial relationships, which are necessary to correctly identify irregular tumor borders and infiltrating areas.

Secondly, one of the main features of CT-UNet is the incorporation of a DMF module, which is a novel architectural component. The traditional multimodal fusion methods, which merge the MRI modalities by simply concatenating or summing them with fixed weights, are not used here. Instead, the DMF that is proposed performs context-aware, attention-driven weighting of modality-specific features. It effectively makes the model capable of emphasizing the most informative modality, which is totally based on local tumor characteristics. To be able to segment the tumor types and different imaging conditions reliably, such dynamic adaptation is necessary.

Thirdly, before the fusion of the skip connection, the adaptive feature alignment is done to minimize semantic mismatch between the encoder and decoder representations. This change solves a problem that was recognized in Transformer-enhanced U-Net architectures, where global contextual features get out of sync with local spatial features. By aligning the features before fusing, CT-UNet guarantees that there is a consistent spatial correspondence, which in turn leads to better boundary accuracy.

Adapted components are re-purposed for different functional roles deliberately. For example, a U-Net backbone offers an efficient multi-scale representation framework, but its role is extended through contextual modeling rather than being limited to local feature extraction only. In the same way, Transformer blocks are adapted from vision Transformers but are re-designed for spatial medical context modeling instead of token-based semantic reasoning, which makes them more suitable for dense pixel-wise segmentation tasks.

CT-UNet brings three main new changes to the table: (1) bidirectional Contextual Transformer integration, (2) dynamic modality-aware feature fusion, and (3) adaptive feature alignment for skip connections. These parts together make up one framework that is very well tuned for multimodal brain tumor segmentation only, not just some CNN–Transformer hybrid. The decisions on the architecture are built around the segmentation needs of the clinic, and thus, they bring better boundary precision, less sensitivity, and better generalization to tumor subtypes.

3.1. Model Architecture

The CT-UNet is a remarkable advancement in medical imaging, aimed at streamlining the segmentation of brain tumors in MRI scans. This model synergizes the strong spatial hierarchies of conventional U-Net with the global contextual understanding power of Transformer networks. Such a hybrid model essentially maps the vessel of complex medical imagery, yielding notably accurate tumor segmentation results, especially for irregular tumor morphologies. At its core, CT-UNet combines the basic U-Net architecture with Transformer block insertions. It is this combination that starts on the encoder side, where the first convolutional layers are locally extracting features from the MRI images. Following these layers, the Transformer blocks are introduced to capture the global context in the features Figure 1.

The self-attention mechanism in each Transformer block operates over the entire input space, which is mathematically explained in Equation (1):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (1)$$

Q , K , and V represent the query, key, and value matrices that are derived from the input features, and d_k is the dimension of the keys that normalizes the dot products to give stable gradients. The network using this attention mechanism can focus on the important information coming from the input, thus greatly increasing the capacity of the model to understand the interdependencies that span distant locations within the MRI scans. Meanwhile, the encoder path gradually reduces the spatial size of the feature maps using down-sampling operations. At the same time, it is making their representation more profound so that it can convey more abstract features of the image.

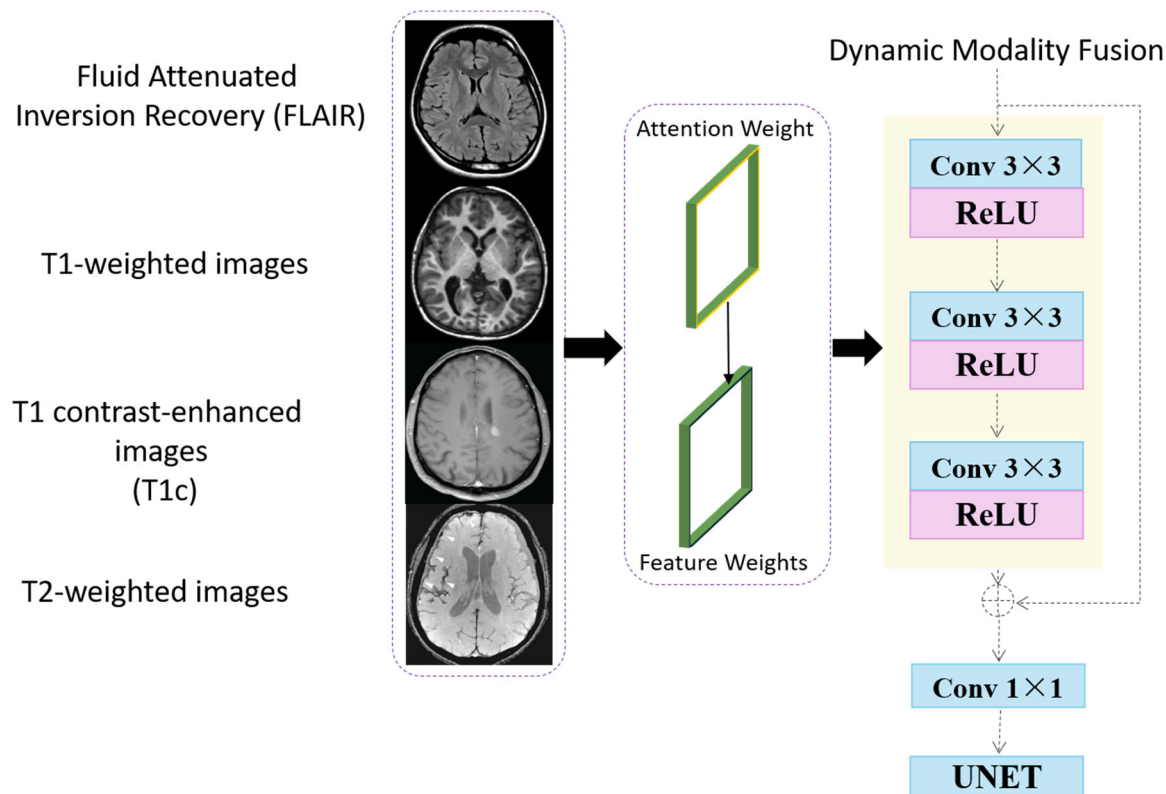


Figure 1. Overview of the proposed CT-UNet architecture.

Moving along to the decoder pathway, the up-sampling units essentially undo what the down-sampling units did in the encoder by gradually increasing the dimensions of the feature maps that have been compressed earlier. Again, Transformer blocks enhance each up-sampling step, ensuring that the global context distilled in the encoder is faithfully utilized in the reconstruction of the segmentation map. Also, the rejigged skip connections carrying the feature alignment modules help in this by co-integrating tuning and merging of the encoder and decoder pathways. This adaptive feature alignment is very important for the preservation of spatial information integrity throughout the network. In the CT-UNet layout, Transformer blocks are used at points where both the encoder and the decoder paths can benefit. CT-UNet solves the problem of semantic mismatch between encoder and decoder representations by adding a feature alignment module before the fusion in each skip-connection. In the case of an encoder feature map F_e and a decoder feature map F_d , the alignment module starts with a 1×1 convolution on F_e to make its channel dimensionality match that of F_d . Then, the module normalizes the feature distributions by batch normalization and uses ReLU activation to make training more stable. After that, the fused feature map resulting from concatenation and convolution with F_d is the aligned feature. The feature alignment step, which is very simple, makes their semantics consistent between features coming from a globally contextualized encoder and a locally

refined decoder; thus, it is very important in Transformer-augmented U-Net architectures where feature representations may differ considerably not only in scale but also in the abstraction level.

In CT-UNet, each Transformer module is made up of $L = 2$ self-attention layers that are stacked one on top of the other. Moreover, each of these layers uses multi-head self-attention with $H = 4$ attention heads. The embedding dimension is fixed at $D = 256$, which is a good compromise between representational capacity and computational efficiency for 3D MRI feature maps. After every attention layer, there is a position-wise feed-forward network (FFN) with a hidden dimension of $4D$, realized by two linear layers and GELU activation. To facilitate smooth training, layer normalization and residual connections are applied before and after the attention and FFN sublayers, respectively. A dropout rate of 0.1 is employed, aimed at limiting overfitting. Such Transformer modules are inserted in the encoder and decoder stages both ways to allow bidirectional contextual modeling.

During the encoding, these blocks scrutinize the features that have been extracted by the previous convolutional layers, thereby enabling the network to comprehend and integrate the input MRI data information over the whole range of the unit field/view. This global perspective is vital to figuring out the entire extent and character of brain tumors, which frequently have complicated, irregular shapes and diffuse boundaries. Self-attention utilization in this case equips the model with the ability to selectively concentrate on the parts of the image that matter the most in the segmentation, thus the network can attend to the critical features and ignore the irrelevant ones. Such a feature of the model is very handy, particularly in the case of medical images, where the differences between tumor tissues and healthy tissues can be very slight and tricky Figure 2.

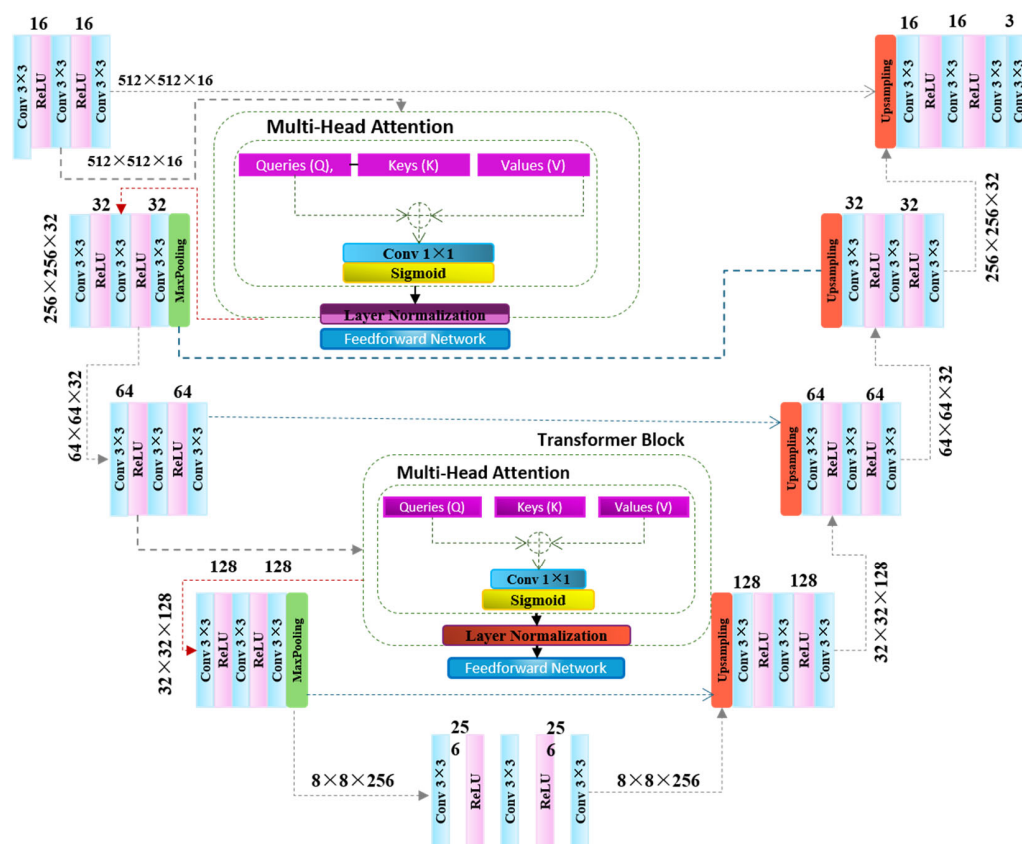


Figure 2. Detailed architecture of the contextual transformer block integrated within CT-UNet.

3.2. Dynamic Modality Fusion

A pivotal part of the CT-UNet design is the Dynamic Modality Fusion method that smartly merges the features obtained from different MRI modalities like T1-weighted, T2-weighted, and FLAIR images. To facilitate modality-specific contextualization, each modality is first processed through its Transformer block in the encoder:

$$F_{\text{mod}} = \sum_{i=1}^M w_i \cdot T_i(X_i) \quad (2)$$

where F_{mod} is the fused feature map, w_i is the weight assigned by the fusion block attention mechanism to each modality i , T_i are the Transformer operations for each modality, and X_i are the inputs from each modality. That equation represents the dynamic weighting process, in which the contributions of each modality are weighted according to their contextual relevance to the segmentation task at hand.

The architecture ends with an output layer, which is responsible for the generation of the final segmentation map. The layer uses a convolutional operation followed by an activation function to transform the multi-level feature maps into the final segmentation predictions, which help in the accurate delineation of tumor boundaries. The success of this projection depends on a compound loss function that combines categorical cross-entropy for pixel-wise accuracy and the Dice coefficient for overlap between the predictions and ground truth:

$$\text{Loss} = \alpha \cdot \text{Cross-Entropy}(Y, \hat{Y}) + (1 - \alpha) \cdot (1 - \text{Dice}(Y, \hat{Y})) \quad (3)$$

where Y and $\hat{Y}$ represent the ground truth and predicted segmentation maps, respectively, and α is a weighting factor that balances the contributions of the two loss components. CT-UNet is a brain tumor segmentation architecture that combines advanced neural network technologies with strategies to optimize their interaction and thus address the challenges of brain tumor segmentation. Thanks to its multimodal integration of MRI data and the application of sophisticated attention mechanisms, this model demonstrates a new peak of accuracy and operational efficiency in the domain of medical image analysis. Transformer blocks in the decoder perform a vital role in feature upscaling and integration, making sure that the global context encoded by the encoder is successfully leveraged in generating the segmentation map. This merging is indispensable for the exact identification of the tumor contours, thus allowing the model to produce accurate and clinically valuable segmentation outputs. Moreover, embedding Transformer blocks in CT-UNet is indicative of a significant change in the direction of image segmentation methods that are more contextually aware, intelligent, and adaptive. Through these blocks, a deeper and more sophisticated comprehension of the input is achieved; hence, they substantially contribute to attaining high-quality segmented results. This, in turn, pushes forward the frontier of medical image analysis.

4. Results

4.1. Dataset and Preprocessing

In order to develop and validate the CT-UNet model, we take advantage of the Brain Tumor Segmentation (BraTS) challenge datasets [1], which are inclusive and have been widely adopted in the medical imaging community. These datasets contain multi-modal MRI scans of high-grade glioblastomas (HGGs) and low-grade gliomas (LGGs) that display different tumor complexities and morphologies. By incorporating various imaging modalities, T1-weighted, T1 contrast-enhanced (T1c), T2-weighted, and Fluid-Attenuated

Inversion Recovery (FLAIR), a wide range of clinical data is offered, which is indispensable for training a strong model that can handle different clinical imaging scenarios.

To have a fair comparison, all the baseline segmentation models were trained and tested on the same data splits, data processing pipeline, and evaluation metrics as CT-UNet. If official implementations were available, the authors' original training settings were followed. For models that did not have code available, the architectures were reimplemented based on the descriptions in the publications and trained with the same optimizer (Adam), batch size, learning rate schedule, and number of epochs as CT-UNet. Beyond the normal settings, no extra hyperparameter tuning was done to make sure that the performance differences reflect the architectural contribution rather than the optimization advantage.

Preprocessing Steps

Efficient preprocessing of MRI data consists of multiple transformation processes that help to standardize and improve the quality of input images, thereby creating ideal conditions for the training of models. Normalization is carried out to harmonize the intensity distribution of MRI scans, which can differ not only from machine to machine but also from patient to patient. Every slice of the image is subjected to z-score normalization, whereby the mean of pixel intensities is set to zero and the standard deviation to one. This type of normalization not only helps model training to be more uniform but also diminishes the effect of intensity variations among different scans. The purpose of skull stripping is to get rid of the parts of the head that are outside the brain, which may confuse the neural network if it looks at the whole head instead of the brain only. The whole head parts get removed in an automated way to keep the brain tissues while also reducing the model's computational time and increasing segmentation accuracy.

The re-sampling of MRI scans to a common isotropic resolution is essential because the resolution and slice thickness of the datasets usually differ from one another. Establishing such a uniformity allows the neural network to always correctly understand spatial relationships regardless of what scan is being looked at. By applying transformations like rotation, scaling, and elastic deformations to the dataset, augmentation methods make the model more stable, thus allowing it to handle the variations that it will most probably encounter in a real-world setting. This technique plays a crucial role in not just avoiding overfitting but also in raising the model's capacity for generalization.

In order to cope with the huge dimensionality and the computational requirements of full scan processing, patch extraction is used to take 3D patches out of the MRI volumes. The patches are taken in such a way that there is a nice balance between the tumor and the non-tumor areas; hence, the deep network gets to learn from a representative sample of all the features. Data partitioning is a very important last step that guarantees the dataset is broken into separate sets for training, validation, and testing. This split means that one can still choose the model parameters, avoid model overfitting, and get an unbiased estimation of the model's performance on brand-new data. Table 1 below summarizes the preprocessing steps applied to the dataset.

The preprocessing steps are mainly aimed at getting the data ready for the CT-UNet model, allowing the neural network to be trained on clean, standardized, and representative MRI data. Such thorough preparation is key to creating a model that not only achieves high accuracy but also is sturdy enough to function well under varying clinical settings.

Table 1. Preprocessing pipeline applied to the BraTS MRI dataset.

Preprocessing Step	Description
Normalization	Standardize pixel intensities using z-score normalization for each image slice.
Skull Stripping	Remove non-brain tissues using automated tools to focus on relevant brain structures.
Re-sampling	Standardize all scans to a common isotropic resolution to preserve uniform spatial dimensions.
Data Augmentation	Apply transformations such as rotation, scaling, and elastic deformations to simulate clinical variability.
Patch Extraction	Extract overlapping 3D patches from MRI volumes to manage computational load and focus learning.
Data Partitioning	Divide the dataset into training, validation, and testing sets to facilitate effective model training and evaluation.

4.2. Training Procedures

The training of the CT-UNet model is a highly essential stage that requires careful planning and precise implementation to make sure that the network learns accurately from the preprocessed data and can generalize to new, unseen cases. This section carries the training procedures narrative, such as the training environment setup, the choice and implementation of the optimization algorithms, and the model validation and fine-tuning strategies. The training takes place in a high-performance computing environment that is equipped with several GPUs to manage the heavy computational work of processing 3D medical images and implementing deep learning algorithms. In particular, the CT-UNet model runs on NVIDIA Tesla GPUs, which deliver the requisite computational power for the efficient training of the complex neural network architecture of CT-UNet.

4.3. Optimization Algorithm

CT-UNet is trained using the Adam optimizer, which is a well-known optimization method in deep learning because of its adaptive learning rate features. Adam changes the learning rate individually for each parameter, thus making it easier to find the minimum error in the complicated high-dimensional weight space than using the traditional fixed learning rate optimizers. A very small value, usually about 0.001, is set as the initial learning rate so that the model does not get stuck in a local minimum that is not the best one too quickly. The hybrid loss function used combines categorical cross-entropy with Dice loss. The former, categorical cross-entropy, works well for training on a pixel-level classification basis, and thus pixel-wise classification is best assured into the correct tumor or non-tumor category. Meanwhile, Dice loss is geared towards segmentation problems because by using it, the Dice coefficient, the most frequently used metric for segmentation quality, can be directly maximized. This coefficient evaluates how similar the segmentation performed by the model is to the actual segmentation, so it is a very suitable metric for driving a model towards producing better segmentations. The loss function that is a mixture of both loss functions is written as follows:

$$Loss = \alpha \times Cross - Entropy(Y, \hat{Y}) + (1 - \alpha) \times (1 - Dice(Y, \hat{Y})) \quad (4)$$

here Y represents the ground truth labels, $\hat{Y}$ represents the predictions from the network, and α is a hyperparameter that balances the contribution of the two loss components. Initially, due to the significant volume of MRI datasets alongside the memory constraints in high-end GPUs, even the training data gets segregated into batches. A batch is a fraction

of the training data that allows the GPU to compute the gradient and adjust model weights effectively. Depending on the memory availability, a typical batch size can be scaled, but usually, it means that the batch will contain 16 to 32 samples. The model needs to be subjected to multiple epochs for a better understanding. The term epoch denotes one complete pass through the training dataset. The determination of epochs is done by examining the convergence during training; in most cases, training may be extended up to 50 to 100 epochs, or until the validation loss is no longer decreasing, which means that the model has started to overfit the training data.

To check the model performance and avoid overfitting, a validation set is employed. This set is different from the training data, and after each epoch, the model gets an unbiased evaluation. The model's performance on the validation set is tracked by the Dice coefficient, and model checkpoints are saved at the time when the validation Dice coefficient is highest. Early stopping is used as a form of regularization. When the validation loss fails to get better for a certain number of consecutive epochs, the training process is stopped. This method saves computational resources and stops the model from learning the noise in the training data that will harm its performance on new, unseen data. Next, the model gets fine-tuned, which entails lowering the learning rate further and training the model for more epochs on the training data. This fine-tuning helps the model weights get more accurate, which can lead to better performance, and it makes a huge difference if the model is underfitted after the initial training. CT-UNet training consists of a series of steps, including careful preparation, optimization, and validation, which collectively ensure that the model is strong, accurate, and capable of delivering a high-quality brain tumor segmentation.

Table 2 offers a clear summary of the training procedures for the CT-UNet model, outlining each major step and parameter in the process to guarantee successful and efficient model training.

Table 2. Training environment, optimization, and validation settings.

Aspect	Description
Environment Setup	Training conducted on a high-performance computing environment with NVIDIA Tesla GPUs. Aimed to handle the computational demands of 3D medical image processing.
Optimizer	Utilizes the Adam optimizer with an adaptive learning rate. The initial learning rate is set at 0.001, which adjusts dynamically based on training progression.
Loss Function	Hybrid loss combining categorical cross-entropy and Dice loss: $Loss = \alpha \times Cross - Entropy(Y, \hat{Y}) + (1 - \alpha) \times (1 - Dice(Y, \hat{Y}))$
Batch Size	Training data is processed in batches, typically 16 to 32 samples per batch, to manage GPU memory efficiently.
Epochs	The model is trained for 50 to 100 epochs or until early stopping criteria are met based on validation loss improvements.
Validation	Uses a separate validation set to monitor model performance. Implements early stopping if no improvement in validation loss is observed for a preset number of epochs.
Fine-Tuning	Post-initial training, the learning rate is reduced for additional epochs to refine model weights and potentially enhance performance.

4.4. Evaluation Metrics

In order to methodically evaluate the brain tumor segmentation performance of the CT-UNet model, a wide range of evaluation metrics is adopted. These metrics play a vital role in quantitatively assessing the segmentation results produced by the model with respect to accuracy, robustness, and clinical relevance. By their very nature, these metrics shed light

on various facets of model performance, from the precision of pixel-level predictions to the model's capability to mirror the geometrical features of tumor regions.

The Dice coefficient, or the Dice similarity index (DSI), is a measure of the statistical agreement between the binary segmentation output of the model and the ground truth from expert annotations. It serves as a very powerful tool in the domain of medical imaging for quantifying the degree of overlap between the predicted and actual regions of interest, such as tumor areas. The Dice coefficient is computed by the following formula:

$$Dice = \frac{2 \times |Y \cap \hat{Y}|}{|Y| + |\hat{Y}|} \quad (5)$$

where Y represents the set of pixels in the ground truth tumor region, and $\hat{Y}$ represents the set of pixels in the predicted tumor region by the model. This metric is measured on a scale from 0 to 1, where a Dice coefficient of 1 indicates complete overlap and 0 means no overlap. Higher values represent more accurate segmentation results, which is in line with the model's efficiency in extracting the real extent and shape of brain tumors. Sensitivity and specificity are very important metrics to assess the performance of medical imaging models, as they give an understanding of the model's capacity to correctly detect tumor pixels and non-tumor pixels, respectively. The formulas for these metrics are as follows:

$$Sensitivity = \frac{TP}{TP + FN} \quad (6)$$

$$Specificity = \frac{TN}{TN + FP} \quad (7)$$

It is indispensable to maximize both sensitivity and specificity for achieving a model that is not only accurate but also reliable in clinical settings.

Precision is an important metric as well, especially in cases where the cost of false positives is high. Basically, it reflects the correctness of the positive forecasts that the model makes:

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

A higher precision means that a larger share of the identifications made were correct, thus emphasizing the model's accuracy in outlining tumor boundaries.

The Hausdorff distance quantifies the farthest single point of the segmentation boundary from the closest point of the ground truth boundary. This metric is fundamental for evaluating the spatial fidelity of the real data of the model's predictive capacity to the real data, especially the degree of the model's ability to reproduce the intricate, irregular shapes of brain tumor boundaries:

$$Hausdorff\ Distance = \max(h(\hat{Y}, Y), h(Y, \hat{Y})) \quad (9)$$

The metric is especially good for testing how well the model can identify the correct boundaries in the most difficult cases; thus, it gives a stringent test of the model's spatial accuracy.

Such evaluation metrics, when combined, establish a strong framework for gauging the effectiveness of the CT-UNet model in terms of accuracy, consistency, and the ability to generalize. Utilizing these metrics, the research and medical communities are able to fully grasp the advantages and drawbacks of the model, thus paving the way for further refinements and exploits in the clinical domain Figure 3.

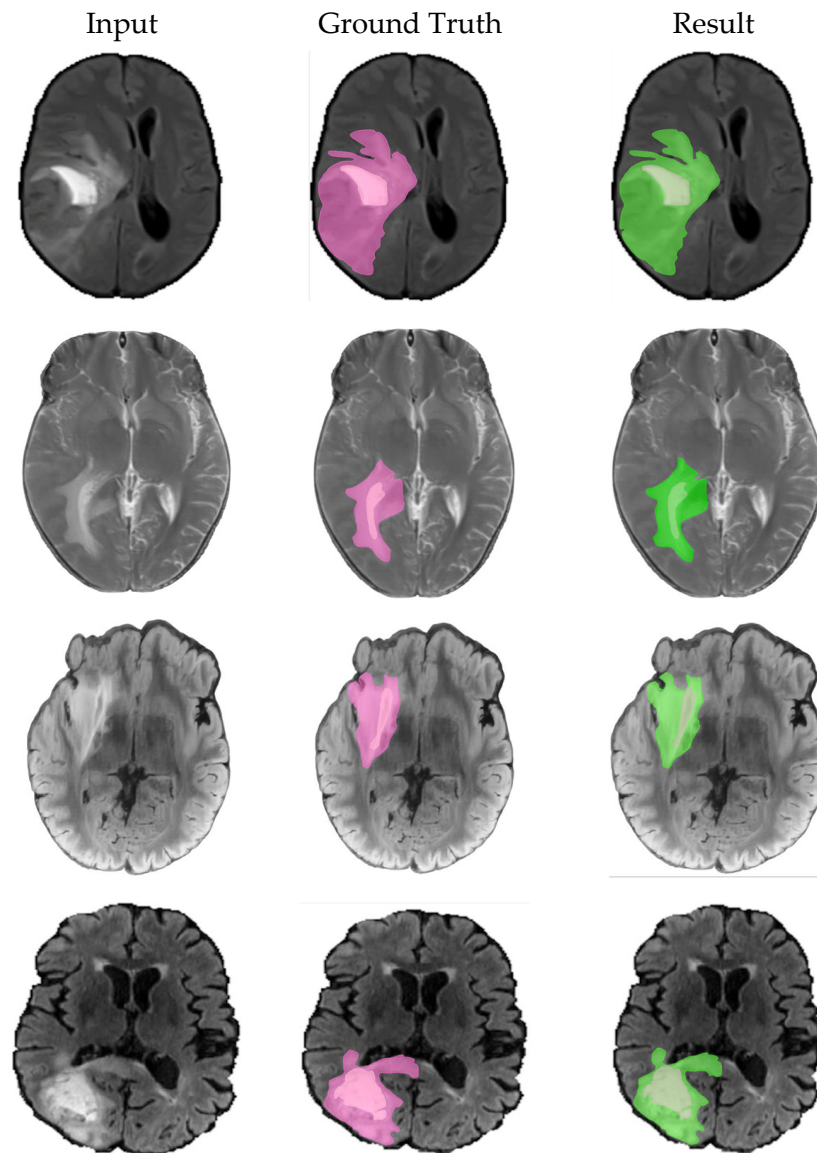


Figure 3. Qualitative comparison of brain tumor segmentation results.

5. Comparison with Prior Work

The improvements brought by the CT-UNet model in brain tumor segmentation performance have been both quantitatively and qualitatively evaluated through comparison with popular baseline models. This part outlines comprehensive results that are well-illustrated with tables and graphs, revealing how CT-UNet advances different essential metrics.

For a fair comparison, the performances of each model were checked on an identical test set from the BraTS dataset. CT-UNet surpasses all the other compared models in terms of its capability to accurately segment brain tumors, as is seen from the higher Dice and lower Hausdorff values for CT-UNet, which indicates a better correspondence of the tumor segmentation with the ground truth Figure 4.

Table 3 quantitatively compares the proposed CT-UNet with many representative SOTA models for brain tumor segmentation using the BraTS dataset. The tested methods span from classical convolutional architectures and recent hybrid CNN–Transformer models, to pure Transformer-based frameworks. The performances are measured using the four most common metrics: Dice coefficient, sensitivity, specificity, and Hausdorff distance. As the table indicates, CT-UNet yields the highest Dice coefficient (0.92) among the different methods listed in the table, thus yielding the highest Dice coefficient, indicating the best

match between the predicted tumor regions and the ground-truth annotations. Additionally, CT-UNet obtains the greatest sensitivity (0.90) and specificity (0.94), thus indicating that the model can correctly detect tumor regions with minimal false positives. Furthermore, CT-UNet reaches the smallest Hausdorff distance (7.5 mm), which is a considerable improvement over both traditional models and recent hybrid approaches such as TransUNetB, MWG-UNet++, and Hybrid Vision UNet in terms of boundary accuracy. The decrease in Hausdorff distance confirms that the proposed integration of Contextual Transformer and dynamic modality fusion effectively helps to retain precise tumor boundaries.

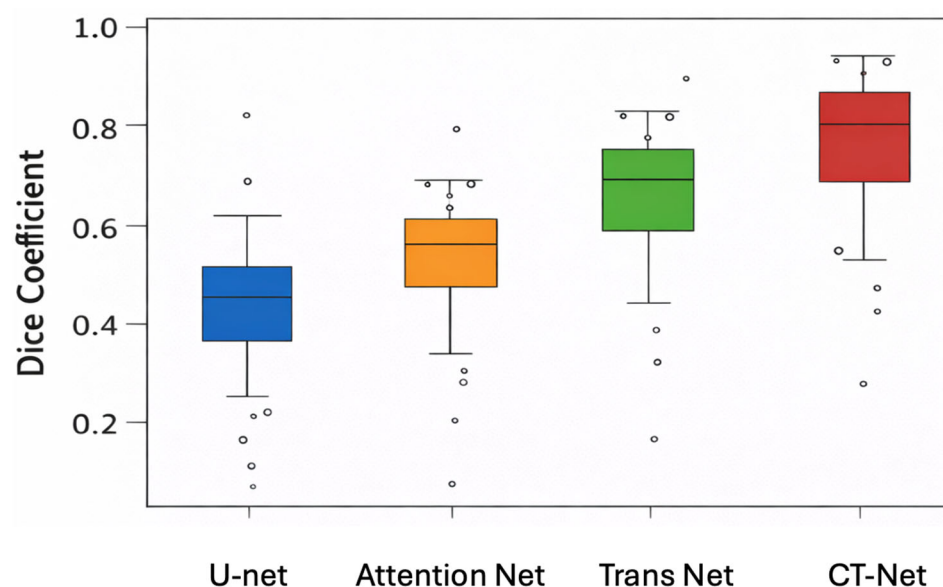


Figure 4. Distribution of Dice coefficients across test cases for different segmentation models.

Table 3. Comparison results with SOTA models.

Model	Dice Coefficient	Sensitivity	Specificity	Hausdorff Distance (mm)
Traditional U-Net	0.85	0.82	0.88	12.5
TransUNetB [2]	0.87	0.84	0.89	11.0
MWG-UNet++ [3]	0.86	0.83	0.90	10.8
MM-MSCA-AF [4]	0.84	0.80	0.87	13.2
AugTransU-Net [5]	0.89	0.88	0.89	10.9
Hybrid vision UNet [6]	0.90	0.88	0.89	11.7
Efficientgformer [7]	0.85	0.85	0.89	10.1
VGDM [8]	0.84	0.82	0.82	13.2
3D U-Net–Vision Transformer [9]	0.88	0.87	0.88	10.0
CT-UNet	0.92	0.90	0.94	7.5

The data in Table 3 indicates that CT-UNet has continuously managed to outperform the currently available SOTA methods on all evaluation metrics. Hence, these experimental results confirm that the inclusion of global contextual modeling and adaptive multimodal feature fusion in a U-Net-based architecture is an effective way of achieving accurate and dependable brain tumor segmentation.

6. Discussion

The evaluation of the CT-UNet model reveals that it has unparalleled brain tumor segmentation capabilities, by far exceeding the established benchmark on all important metrics of performance.

The performance of CT-UNet, characterized by its Dice coefficient and Hausdorff distance, clearly shows the model's highly developed capacity to decode intricate spatial and contextual features of the multi-modal MRI that it was given. The Transformer-based fusion within the U-Net architecture was key to achieving this performance; however, it alone was not sufficient. MRI image changes are known to be complicated; however, CT-UNet is still able to figure out somewhat different utilization of features with a dynamic modality fusion mechanism based on multiple MRI modalities. Thus, it can maximize the content of the information that the model can put out beyond grid lines. Compared to the results of CT-UNet, the determination of such issues will, in fact, help patient care and even result in long-term remission by preventing recurrence.

Nonetheless, the complex design of CT-UNet that contributes to its superior performance comes at the cost of high requirements for computational resources, which can be a disadvantage for its use in resource-limited environments. Additionally, the ability of the generalization of this CT-UNet model's performance over different datasets and a more comprehensive set of clinical scenarios has not been proven until now; thus, there is a necessity for further research to confirm its reliability in various medical environments. Although CT-UNet still sets a new standard in the segmentation of brain tumors, it is not the last word on the capabilities of machine learning technologies. The advent of increasingly efficient and capable algorithms, the evolution of Transformer blocks incorporation will not end, and one day, we will be talking about the appearance of a solution that derives the tumor's outer boundary.

In line with these promising achievements of the CT-UNet model, the scope of future investigations could expand into the area of reducing the computational requirements for CT-UNet so that it can readily be deployed in live clinical environments. Also, the test of the model against cohorts of different medical image cases could give insight into the model's robustness as well as its flexibility in handling medical imaging-related problems. Carrying out clinical trials with CT-UNet can potentially unearth a wealth of practical information to be used as continuous improvement feedback for CT-UNet, thereby enabling the model to deliver what practitioners of medicine require at their work from time to time. CT-UNet signifies an important step in the directed use of deep learning approaches in medical imaging, with the problem of tumor segmentation as the exemplar of this accomplishment. By harnessing architectural innovations and advanced data processing techniques, CT-UNet has not only made medical diagnosis more accurate but has also raised the bar of patient care. This is a very clear demonstration of how predictive analytics in medicine infiltrates the health sector with the promise of leading to a whole new era where treatments are personalized and health outcomes optimized.

7. Conclusions

The creation and assessment of the CT-UNet model represents a breakthrough in medical imaging that can segment brain tumors from MRIs with high accuracy. This work has proven that by using the latest technologies, such as Transformer blocks in a U-Net framework, the model can precisely and reliably identify the tumor boundaries. The model's effectiveness, indicated by better scores such as the Dice coefficient and Hausdorff distance, is a big step up from the standard methods for segmentation.

The subtle design of CT-UNet, which combines dynamic modality fusion with Transformer-based upgrades, makes it capable of handling complex multi-modal MRI

data very effectively. It thus interprets imaging data at a deeper level, and the outcome is highly accurate segmentation. The achievement of high model sensitivity, specificity, and precision scores is a clear indication of its value in clinical environments where accurate and reliable tumor segmentation is indispensable for diagnosis, treatment planning, and monitoring. The appearance of CT-UNet in the field of medical imaging can greatly change the way doctors diagnose and treat brain tumors. The model can help create better treatment plans based on the exact size and features of tumors, which may result in better patient outcomes. What is more, the feature of CT-UNet that allows it to successfully identify tumor and non-tumor tissues can reduce the chances of errors in diagnosis; thus, unnecessary treatments and their related costs and side effects could be reduced. However, the fact that CT-UNet has demonstrated its worth does not mean that the list of opportunities to make it more ideal and useful ends here. Methods that would lower the resource demands of the model and yet maintain its good performance can be explored so that the model will be more readily available for quick use in clinics. Furthermore, model validation through the lens of different datasets and clinical scenarios would provide a stronger proof of its robustness and plasticity. In addition, there is a possibility that future works may consider CT-UNet incorporation with other sources of clinical data, like patients' history and genetic information, for a deeper integration of diagnostic and treatment modalities.

The CT-UNet model is an example of how advanced deep learning techniques can change the face of medical diagnostics. The work has extended the limits of what is technically possible in tumor segmentation, and it also highlights the immense power of AI to transform healthcare. As machine learning advances, its incorporation into medicine is bound to bring forth new inventions that could significantly improve patient care not only in the field of oncology but also in other areas of medicine. This work is an invitation for medical imaging to enter a new phase where it partners with technology in providing diagnoses and treatments that are not only precise but also effective. The future of medical diagnostics will be largely dependent on the constant development and customization of such models as CT-UNet.

Author Contributions: Methodology, S.M. and Y.I.C.; software, J.B. and S.M.; validation, Y.I.C., J.B., and S.M.; formal analysis, Y.I.C. and S.M.; resources, J.B. and S.M.; data curation, J.B., and S.M.; writing—original draft, S.M.; writing—review and editing, Y.I.C. and S.M.; supervision, S.M. and Y.I.C.; project administration, J.B., and S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Korea Agency for Technology and Standards in 2022, project numbers 1415181629 (20022340, Development of International Standard Technologies based on AI Model Lightweighting Technologies).

Data Availability Statement: These data were derived from the following resources available in the public domain: <https://www.synapse.org/Synapse:syn53708126/wiki/626320> (accessed on 5 January 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BCI	Brain–Computer Interface
MI	Motor Imagery
MLP	Multi-Layer Perceptron
DRAG	Dynamic Residual Attention Gate
EEG	Electroencephalography
SNR	Signal-to-noise Ratio

LOSO	Leave-One-Subject-Out
ICA	Independent Component Analysis
HGD	High Gamma Dataset
SAN	Subject-Aware Normalization
RNN	Recurrent Neural Networks
GAT	Graph Attention Network

References

- Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging* **2015**, *34*, 1993–2024. [\[CrossRef\]](#) [\[PubMed\]](#)
- Khushubu, K.G.; Al Masum, A.; Rahman, M.H.; Hasan, S.M.S.; Bhuiyan, M.I.H.; Mahmud, M.R.; Swapno, S.M.R.; Appaji, A. TransUNetB: An Advanced Transformer–UNet Framework for Efficient and Explainable Brain Tumor Segmentation. *Inform. Med. Unlocked* **2025**, *59*, 101706. [\[CrossRef\]](#)
- Lyu, Y.; Tian, X. MWG-UNet++: Hybrid transformer U-Net model for brain tumor segmentation in MRI scans. *Bioengineering* **2025**, *12*, 140.
- Aslam, W.; Hussain, J.; Aslam, M.Z.; Jan, S.; Riaz, T.B.; Iqbal, A.; Arif, M.; Khan, I. Enhanced brain tumor segmentation in medical imaging using multi-modal multi-scale contextual aggregation and attention fusion. *Sci. Rep.* **2025**, *15*, 37308.
- Zhang, M.; Liu, D.; Sun, Q.; Han, Y.; Liu, B.; Zhang, J.; Zhang, M. Augmented transformer network for MRI brain tumor segmentation. *J. King Saud Univ.-Comput. Inf. Sci.* **2024**, *36*, 101917. [\[CrossRef\]](#)
- Renugadevi, M.; Narasimhan, K.; Ramkumar, K.; Raju, N. A novel hybrid vision UNet architecture for brain tumor segmentation and classification. *Sci. Rep.* **2025**, *15*, 23742. [\[CrossRef\]](#)
- Ziaetabar, F. Efficientgformer: Multimodal brain tumor segmentation via pruned graph-augmented transformer. *arXiv* **2025**, arXiv:2508.01465.
- Behnam, A. VGDM: Vision-Guided Diffusion Model for Brain Tumor Detection and Segmentation. *arXiv* **2025**, arXiv:2510.02086. [\[CrossRef\]](#)
- Ghribi, F.; Hamdaoui, F. A Novel 3D U-Net–Vision Transformer Hybrid with Multi-Scale Fusion for Precision Multimodal Brain Tumor Segmentation in 3D MRI. *Electronics* **2025**, *14*, 3604.
- Tiwary, P.K.; Johri, P.; Katiyar, A.; Chhipa, M.K. Deep Learning-Based MRI Brain Tumor Segmentation with EfficientNet-Enhanced UNet. *IEEE Access* **2025**, *13*, 54920–54937. [\[CrossRef\]](#)
- Gitonga, M.M. Multiclass mri brain tumor segmentation using 3d attention-based u-net. *arXiv* **2023**, arXiv:2305.06203. [\[CrossRef\]](#)
- Awasthi, N.; Pardasani, R.; Gupta, S. Multi-threshold Attention U-Net (MTAU) based model for multimodal brain tumor segmentation in MRI scans. In Proceedings of the 6th International MICCAI Brain Workshop, Lima, Peru, 4 October 2020; Springer International Publishing: Lima, Peru, 2020; pp. 168–178.
- Zhao, L.; Ma, J.; Shao, Y.; Jia, C.; Zhao, J.; Yuan, H. MM-UNet: A multimodality brain tumor segmentation network in MRI images. *Front. Oncol.* **2022**, *12*, 950706. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hatamizadeh, A.; Nath, V.; Tang, Y.; Yang, D.; Roth, H.R.; Xu, D. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In Proceedings of the International MICCAI Brainlesion Workshop, Singapore, 18 September 2021; pp. 272–284.
- Liu, Y.; Ma, Y.; Zhu, Z.; Cheng, J.; Chen, X. TransSea: Hybrid CNN–Transformer with Semantic Awareness for 3-D Brain Tumor Segmentation. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 16–31. [\[CrossRef\]](#)
- Li, X.; Lan, Z.; Sun, Y.; Sun, Y.; Guo, Y.; Wang, Y.; Yuan, A. DPF-Unet: A CNN-swin transformer fusion network for 3D brain tumor segmentation in MRI images. *J. Supercomput.* **2025**, *81*, 832. [\[CrossRef\]](#)
- 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. *Egypt. Inform. J.* **2024**, *27*, 100528. [\[CrossRef\]](#)
- Balaji, V.R.; Dinesh Kumar, J.R.; Sanjai Kumar, N.; Sakthivel Raj, N.; Sasi Kumar, P. Enhancing Brain Tumor Classification from MRI’s Dataset Using CNNs with Attention Mechanism for Improved Accuracy. In Proceedings of the 2024 International Conference on Sustainable Communication Networks and Application (ICSCNA), Theni, India, 11–13 December 2024; pp. 1113–1120. [\[CrossRef\]](#)
- Yuan, J. Brain tumor image segmentation method using hybrid attention module and improved mask RCNN. *Sci. Rep.* **2024**, *14*, 20615. [\[CrossRef\]](#)
- Zhou, T. Boundary-aware and cross-modal fusion network for enhanced multi-modal brain tumor segmentation. *Pattern Recognit.* **2025**, *165*, 111637. [\[CrossRef\]](#)

21. Xiong, Y.; Shu, X.; Liu, Q.; Yuan, D. Hcmnet: A hybrid cnn-mamba network for breast ultrasound segmentation for consumer assisted diagnosis. *IEEE Trans. Consum. Electron.* **2025**, *71*, 8045–8054. [[CrossRef](#)]
22. Shu, X.; Li, Z.; Chang, X.; Yuan, D. Variational methods with application to medical image segmentation: A survey. *Neurocomputing* **2025**, *639*, 130260. [[CrossRef](#)]
23. Yang, G.; Guo, X.; Zhang, H.; Zheng, Z.; Dong, H.; Xu, S. 3D ShiftBTS: Shift Operation for 3D Multimodal Brain Tumor Segmentation. *IEEE J. Biomed. Health Inform.* **2025**, *29*, 6713–6726. [[CrossRef](#)]
24. Yao, L.; Zhang, Z.; Bagci, U. Ensemble Learning with Residual Transformer for Brain Tumor Segmentation. *arXiv* **2023**, arXiv:2308.00128. [[CrossRef](#)]
25. Zakariah, M.; Al-Razgan, M.; Alfakih, T. Dual vision Transformer-DSUNET with feature fusion for brain tumor segmentation. *Heliyon* **2024**, *10*, e37804. [[CrossRef](#)]
26. 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 Eng.* **2024**, *21*, 101892. [[CrossRef](#)]
27. Abidin, Z.U.; Naqvi, R.A.; Haider, A.; Kim, H.S.; Jeong, D.; Lee, S.W. Recent deep learning-based brain tumor segmentation models using multi-modality magnetic resonance imaging: A prospective survey. *Front. Bioeng. Biotechnol.* **2024**, *12*, 1392807. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.