electronics

MDPI

Article

An Optimized Transformer-Driven Approach for Multi-Modal Brain Tumor Segmentation in Neuro-Oncological MRI Imaging

Anupam Gupta * and Preety D. Swami

Department of Electronics and Communication Engineering, University Institute of Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya (Technological University of Madhya Pradesh), Bhopal 462033, Madhya Pradesh, India; preetydswami@gmail.com or preetydswami@rgpv.ac.in
* Correspondence: anupamgupta@rgpv.ac.in or anupamgupta28@gmail.com

Abstract

Accurate brain tumor segmentation from multi-modal magnetic resonance imaging (MRI) is a critical task in healthcare, directly supporting early diagnosis, surgical planning, radiotherapy targeting, and longitudinal treatment assessment in neuro-oncology. However, reliable segmentation remains challenging due to complex tumor heterogeneity, irregular anatomical structures, and intensity in homogeneity across MRI modalities, which limit the effectiveness of conventional image processing and standard deep learning (DL) and Artificial Intelligence (AI) approaches. To address these challenges, this study proposes a transparent and healthcare-centric DL framework, termed Swin-UNETR-GJO, which integrates the hierarchical self-attention capability of the Swin Transformer with the encoder–decoder architecture of UNETR for multi-modal brain tumor segmentation. To further enhance model adaptability and performance, the Golden Jackal Optimization (GJO) algorithm is employed to automatically adjust architectural and training parameters, such as attention window size, learning rate, embedding dimensionality, and dropout. The proposed model is trained and evaluated on the BraTS 2021 benchmark dataset. In addition, attention-based activation and feature attribution maps are incorporated. To support clinical interpretability and trust in AI-assisted diagnosis, the two computational explainability measures, Region-Attention Overlap (RAO) and Boundary-Attention Accuracy (BAA), are incorporated, enabling visual delineation of tumor sub-regions. The proposed Swin-UNETR-GJO achieved Dice scores of 0.91 for Whole Tumor (WT), 0.89 for Tumor Core (TC), and 0.87 for Enhancing Tumor (ET), with a minimum Hausdorff Distance 95% (HD95) of 6.1 mm; five-fold cross-validation further confirmed stable performance, with mean Dice Similarity Coefficient (DSC) values of 0.908 $\pm$ 0.008, 0.884 $\pm$ 0.005, and 0.868 $\pm$ 0.008, respectively. The proposed framework provides accurate, stable, and interpretable brain tumor segmentation, highlighting its potential application as a reliable decision-support tool for healthcare.

Keywords: brain tumor segmentation; healthcare imaging; multi-modal MRI; transformer-based deep learning; golden jackal optimization; explainable artificial intelligence; neuro-oncology; medical image analysis

Check for updates

Academic Editor: Kefeng Ji

Received: 11 August 2026
Revised: 16 October 2026
Accepted: 17 October 2026
Published: 9 November 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) license.

1. Introduction

Brain tumors are some of the most problematic areas of neuro-oncology because of their aggressive proliferation, recurrence rates, and poor prognosis, and because of the high quality of images and the variety of tissue contrasts obtained under different imaging

modes like T1, T1ce, T2, and FLAIR [1]. The correct classification of brain tumors on MRI images is critical in early diagnosis, treatment planning, radiation targeting, and post-operative assessment [2].

In the last ten years, DL and, more specifically, Convolutional Neural Networks (CNNs) [3] have changed the approach to medical image analysis, making it possible to perform end-to-end learning and extract features automatically. UNet and its variants have had a fair amount of success in brain tumor segmentation tasks. Nonetheless, CNNs are constrained by nature to record global contextual relationships because of their local receptive fields [4]. Consequently, they tend to have problems with complicated tumor morphologies, inconsistent patterns of texture, and different intensities of various modalities. This is also complicated by the non-uniform shapes and blurry edges of tumors that contribute to poor performance in multi-class segmentation with the WT, TC, and ET.

The recent developments in attention-based architecture, especially Transformers, have brought a paradigm shift in medical image segmentation. Initially designed as a natural language processing model, Transformers have also been shown to be very successful in computer vision, with models such as Vision Transformer (ViT) [5] and a hierarchical version, the Swin Transformer. The Swin Transformer is a solution to the computational inefficiencies of ViT, which introduces shifted window mechanisms that are scalable and able to capture local and global dependencies. Swin Transformers can be viewed as a promising approach to volumetric medical segmentation problems when paired with encoder–decoder networks such as UNETR (UNet with Transformers as the encoder) due to their ability to use long-range attention and multi-scale feature learning. Nevertheless, regardless of these benefits, the performance of these hybrid architectures is vulnerable to hyperparameter settings such as the size of the attention window, the embedding dimensions, and the depth of the attention layers. The existing methods are mainly based on manual tuning or grid search, which are time-consuming, computationally intensive, and less desirable for solving complex problems such as brain tumor segmentation.

The other important factor in clinical implementation of DL models is their interpretability. Although black-box models are highly accurate, they do not tend to be trusted by clinicians because of the poor transparency in decision-making [6]. When it comes to brain tumor segmentation, it is crucial that the model is able to generate not just accurate delineations, but also give the visual evidence of the regions that contribute to the predictions [7].

Considering the shortcomings of current models, there is a significant urgency to create a brain tumor segmentation framework that incorporates the effective representation learning power of Swin Transformers, the structural benefits of encoder–decoder networks, an effective and adaptable hyperparameter optimization approach, and built-in explainability. This is the key driver of the given work.

A number of studies have investigated Swin Transformer variants and hybrid models such as TransBTS, Swin-UNet, and SwinBTS for brain tumor segmentation [8]. Even though these models have proved to be more effective than conventional CNNs, they can be challenging to tune manually, lack inherent interpretability, and do not scale to different imaging conditions. Moreover, the majority of the transformer-based models have a static architecture that does not consider the adaptability of architectural parameters in tasks. This causes underfitting or overfitting in various situations, which restricts their strength.

The objective of this study is to develop an explainable and computationally efficient framework for accurate 3D brain tumour segmentation, combining recent and newly developed architectures and methods. The major contributions of this study are as follows:

- This study develops a new hybrid model that integrates the hierarchical attention mechanism of Swin Transformers and the spatial sensitivity of UNETR to segment volumetric medical images.
- This study uses GJO integration (a nature-inspired, population-based optimization algorithm) to automatically adjust architectural and training parameters, and hence enhance the performance of segmentation and training.
- Integration of attention-based explainability models, which give visual explanations of the model predictions and increase the transparency of the segmentation process.
- End-to-end results on the BraTS 2021 dataset, with state-of-the-art results of Dice scores of 0.91 on WT and 0.87 on ET regions.

This paper is organized as follows: Section 2 reviews related work; Section 3 describes the methodology; Section 4 presents the results; Section 5 discusses the findings; and Section 6 concludes the paper with future directions.

2. Literature Review

In the last five years, the research on automated brain tumor segmentation with DL has experienced a dramatic increase, especially on multi-modal MRI data. These studies have been done to enhance the accuracy, robustness, and interpretability of segmentation models, with the use of CNNs, transformers, generative adversarial networks (GANs), and hybrid architectures. Isensee et al. [9] (2021) did not develop a new network design; instead, they coded a rich collection of best practices, which enabled nnU-Net to be able to consistently reach state-of-the-art performance on various benchmarks.

Chen et al. [10] (2024) in their study proved that Transformers could achieve a considerable enhancement of the global feature representation, which results in competitive performance in a wide range of segmentation tasks in comparison with pure CNN models. The authors demonstrated that the combination of CNN feature maps and Transformer outputs, and decoding them with the help of a U-Net-like structure, enhances awareness of boundaries and semantic consistency.

The work by Liu et al. [11] (2021) proved that this design can provide a good trade-off between capturing global contextual dependencies and keeping the computational cost manageable, which allows the model to be a powerful and scalable backbone for a wide range of computer vision tasks. The authors demonstrated that the hierarchical feature representation and window-shifting mechanism are much more effective in performance than previous Transformers, which were ineffective with large images.

Cao et al. [12] (2022) demonstrated that hierarchical characteristics, skip connections, and U-shaped structural principles can be maintained to ensure that a fully Transformer-based segmentation model can perform well on medical imaging tasks. The authors showed that Swin-UNet is able to capture local and global contextual information effectively to provide high-quality segmentation without the use of CNNs.

Hatamizadeh et al. [13] (2022) showed that transformer-based embeddings had the potential to be effective in capturing long-range dependencies in volumetric medical data, which was much more effective in challenging 3D segmentation tasks, such as brain tumor analysis. The study demonstrated high accuracy and improved the feature representation ability by connecting the transformer encoder tokens to various levels of a CNN-based decoder via skip connections.

Wang et al. [14] (2021) suggest combining a 3D CNN-based encoder with a Transformer unit to trade off local volumetric feature extraction and global contextual reasoning. They find that Transformer-based global attention is beneficial in improving the accuracy of segmentation over the traditional CNN-only methods on the BraTS dataset.

Zhou et al. [15] (2021) present an original framework in which the convolutional operations are combined with both local and global self-attention mechanisms, which can enable the model to maintain fine-grained spatial details and simultaneously be capable of encoding long-range contextual relationships to do so.

Zhang et al. [16] (2021) establish two complementary directions of CNNs: a branch that is more effective in capturing fine-grained local features and a Transformer branch focused on the long-range features throughout the picture. A specially designed BiFusion module is used to combine these features to allow early and effective interaction between the two types of features.

Wei, C. et al. [17] (2023) are concerned with the issue of maintaining fine structural information in volumetric data and, at the same time, being able to capture the broad contextual information that is required to draw accurate tumor boundaries. To do this, they also add hierarchical Swin Transformer blocks with V-Net-style decoders and non-local attention mechanisms that can improve the modeling of long-range dependencies of 3D medical images.

Perera, S. [18] (2024) presents the innovations of axial attention, gated attention, and self-distillation mechanisms to minimize the memory footprint without the loss of the capability to simulate global contextual relationships on 3D data. They have shown in their experiments on the BraTS dataset that these efficient attention schemes can attain high segmentation performance with significantly reduced computational costs compared to traditional 3D Transformers.

Khan et al. [19] conducted a systematic review of advanced feature-selection techniques in medical imaging, covering conventional filter, wrapper, and embedded methods as well as DL, hybrid, and emerging quantum-based approaches. The study highlighted that although CNN- and autoencoder-based methods can effectively reduce high-dimensional imaging features and improve diagnostic performance, their limited interpretability and computational complexity remain important challenges.

Abidin, Z. U. et al. [20] (2024) study results indicate that Transformer-based models are always better than basic CNNs on BraTS, with special attention to design and regularization, Swin variants, and decoder fusion approaches being especially successful. Table 1 represents a summary of the literature review with the limitation's/research gaps identified in each study.

Table 1. Summary of the literature review.

References	Method Used	Findings	Limitations
Isensee et al. (2021) [9]	A self-configuring and automated U-Net architecture that automatically chooses preprocessing, architecture variants, training schedules, and post-processing rules for any segmentation task.	nnU-Net has set a solid benchmark in most medical segmentation problems by embedding best practices into an automated pipeline, as opposed to new network blocks. It performs optimally with the adaptation of its pipeline to every dataset.	It is based on heuristic rule sets and empirical design decisions; it is not a new architectural development, and it does not directly focus on long-range context modelling.
Chen et al. (2024) [10]	Hybrid architecture: a pyramid consisting of CNN encoders (local features) and a Transformer encoder (global context), and then a U-Net-style decoding.	Showed that Transformers are effective encoders for medical segmentation, better at modelling the global context and competitive performance on multiple tasks.	Poor localization with a single application; rather expensive to compute; early experiments were mainly 2D patch/2D-slice.

Table 1. *Cont.*

References	Method Used	Findings	Limitations
Liu et al. (2021) [11]	Presents the hierarchical Swin Transformer with shifted local windows to offer scalable and efficient high-resolution image self-attention.	Balances the global context and computational cost; it becomes a solid foundation for numerous CV tasks.	The original paper is not medical imaging-specific; volumetric/3D data needs to be adapted.
Cao et al. (2022) [12]	A pure Transformer segmentation architecture that is a U-Net like and encoder/decoder that uses Swin Transformer blocks.	Proved that full-transformer U-shaped architectures can be used to obtain good segmentation performance if hierarchical features and skip connections are maintained.	Even pure transformer models can be resource-intensive; they can be required to be carefully regularized on small medical distributions.
Hatamizadeh et al. (2022) [13]	Trains a 3D encoder–decoder segmentation model (UNETR) to 3D volumetric medical images using a Vision Transformer as the encoder.	Effective 3D segmentation tasks (such as brain tumor) with the help of transformer embeddings as tokens in an encoder connected with a CNN decoder using skip connections.	Transformer encoders are heavy on memory to receive volumetric inputs and need patching/embedding plans.
Wang et al. (2021) [14]	Trains 3D CNN encoders of local volumetric features with a Transformer module for global context—specifically trained on BraTS brain tumor segmentation.	Initial experimental results of transformer integration into 3D brain tumor pipelines demonstrated to be more effective in global context modeling and better than 3D-CNN baselines on BraTS.	The method involved volumetric map tokenization that had to be done with care; it was not a non-trivial computational task.
Zhou et al. (2021) [15]	nnFormer: a 3D transformer in which concatenation is replaced with self-attention and local/global attention specialization and skip-attention.	Beat the previous transformer counterparts and minimized HD95 on volumetric segmentation, presenting transformer designs optimized for volumetric medical data.	The complexity of architecture and tuning demands; the variants of transformers are yet to be carefully blended with convolution.
Zhang et al. (2021) [16]	Parallel-branch design: Two divisions of the CNN branch (or globally), one tracking local detail and one operating on a global context with a merged branch representation through a BiFusion module.	Attains high 2D and 3D medical segmentation accuracy and effective use of parameters through early CNN and Transformer feature fusing.	The design of fusion modules is empirical; the performance is based on the attentive balancing of branches.
Wei, C. et al. (2023) [17]	Renderer Architectures: Adaptations of Swin Transformer to high-resolution and 3D medical pipeline-specific V-Net-style decoders and non-local blocks to volumetric problems.	Demonstrate better segmentation accuracy on BraTS and MSD, as well as responding to high-resolution contextual demands.	Resource-demanding still; there are still trade-offs between resolution, receptive field, and GPU memory.
Perera, S. (2024) [18]	The designs like MAT and SegFormer3D propose axial/gated/efficient attention schemes to minimize the memory footprint of the 3D memory and retain the global context.	Demonstrate that careful attention design (axial, gated, self-distillation) is able to offer benefits of transformers with reduced resources and reasonable accuracy on BraTS.	New variants of attention require dataset-derived tuning;

3. Materials and Methods

The proposed Swin-UNETR–GJO framework was developed through a rigorously designed and reproducible methodological pipeline for automated multimodal brain tumor segmentation. The framework leverages the BraTS 2021 Task 1 dataset comprising four complementary sequences—T1, contrast-enhanced T1 (T1ce), T2, and FLAIR—to capture both fine-grained anatomical structures and pathological tumor characteristics. All experiments were conducted on high-performance GPU platforms to ensure efficient large-scale training, hyperparameter optimization, and inference. A comprehensive preprocessing pipeline was employed, including data cleaning and data augmentation. The core segmentation architecture is based on Swin-UNETR, which integrates the hierarchical shifted-window self-attention mechanism of Swin Transformers with the encoder–decoder design of UNETR to effectively model long-range contextual dependencies and multi-scale feature represen-

tations required for heterogeneous tumor subregion delineation. The GJO algorithm was initially employed for architectural optimization, particularly to refine the configuration of the proposed Swin-UNETR transformer blocks. Subsequently, GJO was also applied during the training phase for hyperparameter optimization, enabling systematic tuning of key training parameters to improve model convergence and overall predictive performance. GJO helps in enabling an optimal balance between global exploration and local exploitation during learning. Model interpretability was strengthened using explainability techniques, including Grad-CAM, feature activation mapping, and attention visualization. Segmentation performance was quantitatively evaluated using standard metrics, including DSC, Hausdorff Distance, Precision, Recall, and volumetric overlap measures, ensuring a comprehensive assessment across different tumor components. The overview of the proposed approach is presented in Figure 1.

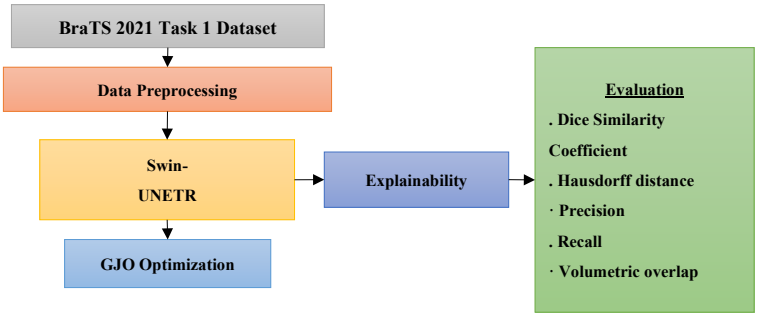


Figure 1. Overview of the proposed method.

3.1. Dataset Description

The BraTS 2021 Task 1 dataset is a benchmark multi-institutional collection of preoperative multi-parametric MRI (mpMRI) scans created for the Brain Tumor Segmentation Challenge 2021, focusing on automatic segmentation of brain tumors. It comprises 3D MRI volumes in four complementary modalities: T1-weighted (T1), T1-weighted contrast-enhanced (T1ce), T2-weighted (T2), and FLAIR—for a large cohort of glioma patients, with expert-annotated voxel-level labels delineating key tumor sub-regions such as enhancing tumor, tumor core, and WT. These scans are preprocessed (co-registered, skull-stripped, resampled, and intensity-normalized) to ensure consistent spatial alignment across modalities and facilitate training and evaluation of DL segmentation models. The dataset is widely used to develop and benchmark state-of-the-art neural architectures for brain tumor delineation in research and algorithm competitions. A representative axial FLAIR MRI slice and its corresponding expert-annotated ground-truth tumor segmentation mask from the BraTS 2021 dataset are illustrated in Figure 2, highlighting the clear delineation of relevant tumor subregions. In the segmentation mask, dark purple represents background (Label 0), blue represents necrotic/non-enhancing tumour core (Label 1), green represents peritumoral edema (Label 2), and yellow represents enhancing tumour tissue (Label 4).

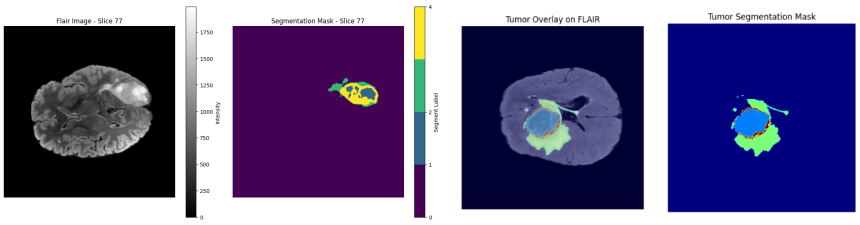


Figure 2. A representative FLAIR MRI slice and its corresponding ground-truth tumor segmentation mask from the BraTS 2021 dataset.

Figure 3 illustrates representative samples from the BraTS 2021 dataset, highlighting the multi-modal magnetic resonance imaging (MRI) characteristics and corresponding tumor annotations across three orthogonal anatomical planes—axial, sagittal, and coronal. The top row presents the original MRI slices, depicting the underlying brain anatomy and visible tumor regions with varying intensity patterns. The subsequent rows demonstrate different visual representations, including intensity-enhanced and pseudo-color mappings, which emphasize tissue contrast variations and improve the perceptual distinction between tumor sub-regions and healthy brain tissue. The final row overlays the expert-annotated tumor masks on the original MRI slices, clearly delineating the tumor boundaries and spatial extent.

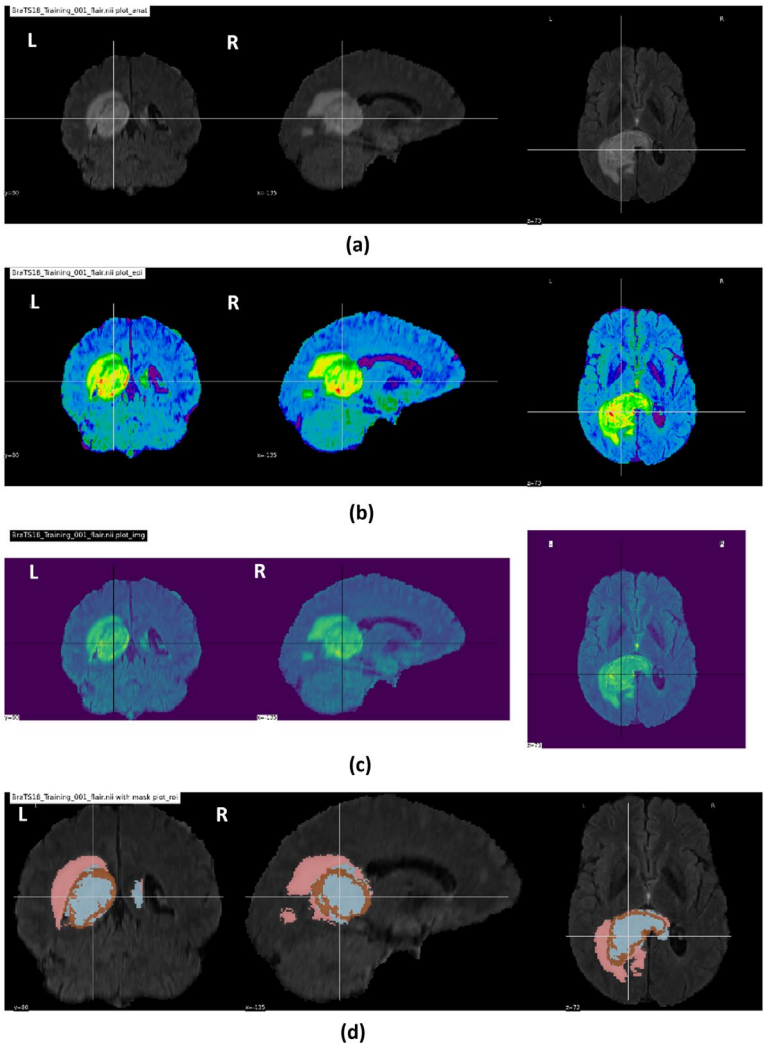


Figure 3. Multi-View Visualization of Multi-Modal MRI Slices and Tumor Annotations from the BraTS 2021 Dataset. (**a**) Original FLAIR MRI; (**b**) Colour-Mapped FLAIR MRI; (**c**) FLAIR MRI Intensity Visualisation; (**d**) FLAIR MRI with Tumour Segmentation Mask; The figure presents coronal, sagittal, and axial MRI views. Grayscale images show anatomical structures, while pseudocolour images use blue/purple for lower signal intensities, green for intermediate intensities, and yellow/red for higher intensities. In the tumour annotation overlays, pink, orange, and light blue distinguish the highlighted tumour regions. White crosshair lines indicate slice positions, while L and R denote left and right anatomical orientations, respectively.

Figure 4 presents the central axial slices of a representative subject from the BraTS 2021 dataset, illustrating the complementary characteristics of multi-modal brain MRI acquisitions alongside the corresponding ground-truth segmentation. From left to right,

the modalities include Fluid-Attenuated Inversion Recovery (FLAIR), T1-weighted (T1), contrast-enhanced T1-weighted (T1CE), and T2-weighted (T2) images, followed by the expert-annotated segmentation (SEG). Each modality captures distinct tissue properties: FLAIR suppresses cerebrospinal fluid signals to enhance peritumoral edema, T1 highlights anatomical structure, T1CE emphasizes contrast-enhanced regions, and T2 accentuates fluid-rich pathological areas. The segmentation mask delineates tumor sub-regions, typically encoding enhancing tumor, tumor core, and surrounding edema, thereby serving as the reference standard for supervised learning.

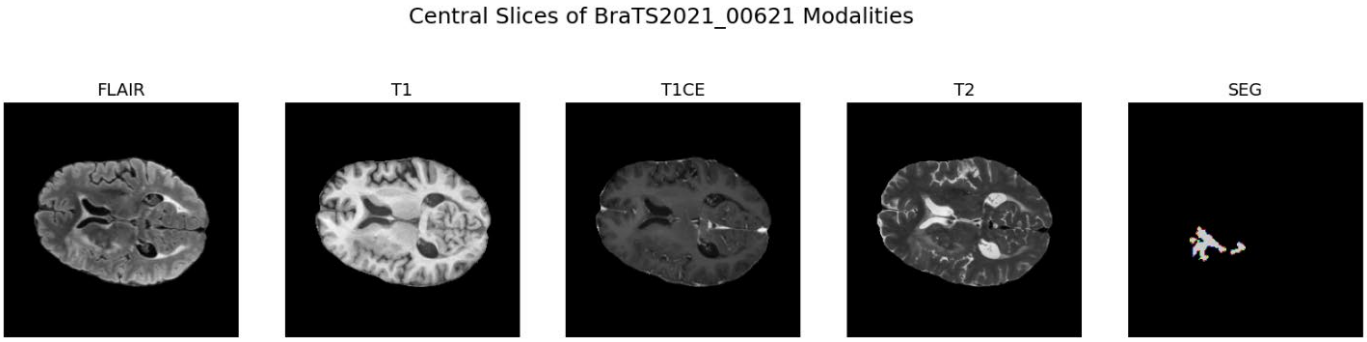


Figure 4. Central Axial Slices of Multi-Modal MRI and Ground-Truth Segmentation from the BraTS 2021 Dataset. The FLAIR, T1, T1CE, and T2 MRI modalities are displayed in grayscale to represent different anatomical tissue characteristics. In the ground-truth segmentation (SEG), black represents the background (Label 0), while distinct colours identify the necrotic/non-enhancing tumour core (Label 1), peritumoral edema (Label 2), and enhancing tumour tissue (Label 4), following the BraTS 2021 annotation scheme.

3.2. Environment and Tools of Computation

All experiments were conducted in a high-performance computing environment optimized for large-scale 3D medical image segmentation. The system comprised an NVIDIA RTX 3090 GPU (24 GB GDDR6X VRAM) to efficiently execute volumetric convolutions, hierarchical self-attention, and multi-stage feature extraction inherent to Swin Transformer architectures, supported by an Intel Core i9 CPU for high-throughput data loading and parallel preprocessing, and 16 GB DDR4 RAM to accommodate memory-intensive 3D MRI volumes. The software stack was built on Python 3.10, with PyTorch 2.2 as the primary deep-learning framework, leveraging optimized CUDA kernels and automatic differentiation for training attention-based models.

Medical image–specific operations, including preprocessing, augmentation, patch-based sampling, and sliding-window inference, were implemented using the MONAI framework. For inference, the trained Swin-UNETR-GJO model was evaluated using MONAI's three-dimensional sliding-window inference procedure. An inference window size of 96 × 96 × 96 voxels was used, with an overlap ratio of 0.50 along each spatial dimension. Predictions from overlapping windows were combined using Gaussian-weighted averaging with a sigma scale of 0.125. This approach assigns greater weight to predictions near the centre of each window and reduces the influence of predictions near window boundaries. The aggregated outputs were used to reconstruct the full-volume segmentation predictions. These inference settings were maintained consistently during model evaluation.

Supplementary libraries—SimpleITK 2.4.0 for NIfTI data handling, NumPy 1.26.4 and SciPy 1.13.0 for numerical computation, scikit-image 0.23.2 for image processing, and Matplotlib 3.10.0 for qualitative and quantitative visualization—were used to support analysis and reproducibility. This integrated hardware–software configuration ensured stable, efficient, and reproducible experimentation on large-scale 3D neuroimaging datasets.

3.3. Preprocessing Pipeline

To maintain consistency and good-quality input representation, a preprocessing pipeline was adopted to handle all MRI modalities of the Swin-UNETR-GJO framework.

The preprocessed BraTS volumes had a native spatial dimension of $(240 \times 240 \times 155)$ voxels at an isotropic resolution of 1 mm^3. During training, 3D patches of $96 \times 96 \times 96$ voxels were extracted from the cropped volumes.

A comprehensive image data augmentation pipeline was added to enhance the diversity of the data and the generalization of the model.

The random rotation within $\pm 15°$ was applied with a probability of $p = 0.50$, elastic deformation with $p = 0.30$, random intensity perturbation with $p = 0.50$, and Gaussian-noise augmentation with $p = 0.30$.

Elastic deformation was applied by introducing smooth, non-rigid spatial displacements to the MRI volumes, thereby simulating realistic inter-patient anatomical and tumour-shape variations while preserving the overall anatomical structure. However, excessive deformation may alter clinically relevant structures and mislead segmentation; therefore, deformation strength was conservatively limited to preserve anatomical plausibility.

In this study, elastic deformation was applied conservatively with a maximum displacement of 1 mm, Gaussian smoothing $\sigma = 2$ mm, thereby introducing only mild non-rigid variation while minimizing distortion of tumour boundaries and clinically relevant brain structures.

Gaussian noise augmentation is determined as shown in Equation (1).

$$x' = x + \mathrm{N}\left(0, \sigma_{\mathrm{n}}^2\right) \tag{1}$$

where $\mathrm{N}\left(0, \sigma_{\mathrm{n}}^2\right)$ is a zero-mean Gaussian noise with variance σ_{n}^2, and x' is the intensity of the augmented voxel with $\sigma_n = 0.01$. All stochastic augmentation procedures were applied only to the training data, while validation and test volumes were evaluated without augmentation.

3.4. Swin-UNETR-GJO Framework

The suggested segmentation framework is a synthesis of three potent elements, which include Swin Transformer, UNETR, and the GJO algorithm, to form a highly optimized and precise system of multi-regional brain tumor segmentation. Originally, the Swin Transformer has hierarchical self-attention, which effectively captures both local and long-range dependencies in 3D MRI volumes, allowing the model to learn rich contextual representations that are critical to differentiating tumor subregions like enhancing tumor, tumor core, and WT. This is complemented by the encoder–decoder structure of UNETR, which takes advantage of a transformer-based encoder to extract deep semantic features, with the decoder synthesizing fine-grained spatial detail, which is important to localize boundaries accurately and provide structural consistency of segmentation maps. To further improve the performance of the models, the GJO algorithm is added as a strategy of adaptive hyperparameter tuning that dynamically searches the space of critical parameters (learning rate, patch size, window size, depth settings, and regularization factors) to enable the model to converge to an optimal configuration without finding suboptimal solutions or overfitting.

3.4.1. Swin Transformer Encoder

The encoder component of the proposed framework is the Swin Transformer, which is meant to extract high-quality multiscale contextual representations that are essential in the accurate segmentation of brain tumors. It is based on fixed-size, non-overlapping input

patches, initially projected to a high-dimensional feature space by a linear patch embedding operation, mathematically defined as Equation (2).

$$z_0 = E(x) \tag{2}$$

where x denotes the input patches and $E(\cdot)$ denotes the embedding function. This conversion transforms the raw voxel values into feature vectors which may undergo attention-based processing. At every hierarchical level of the encoder, the local contextual dependencies are learned with the assistance of the window-based multi-head self-attention (W-MSA) mechanism that computes the attention on patches in a local window. The multi-head self-attention operation is given as Equation (3).

$$MSA(Q, K, V) = Softmax\left(\frac{QK^T}{d_k}\right)V \tag{3}$$

Q, K, and V are query, key, and value matrices, respectively, and d_k is the dimension of the key vectors. The W-MSA enables the model to concentrate on local fine-grained features and be computationally efficient. In order to enable interaction among non-overlapping windows globally, the encoder uses a shifted window-based multi-head self-attention (SW-MSA), whereby the windows are partitioned alternately across successive layers, and thus, the encoder supports cross-window attention. This is the operation that is written as Equation (4).

$$z_l = SW - MSA(LN(z'_l)) + z'_l \tag{4}$$

where z'_1 is the result of the previous local attention layer, and LN refers to layer normalization. The middle local attention step in each stage is provided as Equation (5).

$$z'_1 = W - MSA(LN(z_{l-1})) + z_{l-1} \tag{5}$$

ensuring residual connections that stabilize training and maintain spatial feature information. where z'_l denotes the output of the preceding local-attention layer, and $LN(\cdot)$ denotes Layer Normalization. Refined and transformed features through non-linear mappings are applied after attention operations with the help of a feed-forward network (FFN that is defined as Equation (6).

$$FFN(_z) = W_2\sigma(W_1z + b_1) + b_2 \tag{6}$$

where W_1 and W_2 are weight matrices, which can be learnt, b_1 and b_2 are biases, and z is an activation function (usually GELU).

Figure 5 illustrates a structural comparison between the conventional Swin Transformer block and the Swin-UNETR framework. The standard Swin Transformer block follows a fixed hierarchical design comprising W-MSA, SW-MSA, layer normalization (LN), and multi-layer perceptron (MLP) modules with manually predefined hyperparameters. The cross inside the circle, ⊕, denotes element-wise addition used for the residual or skip connection.

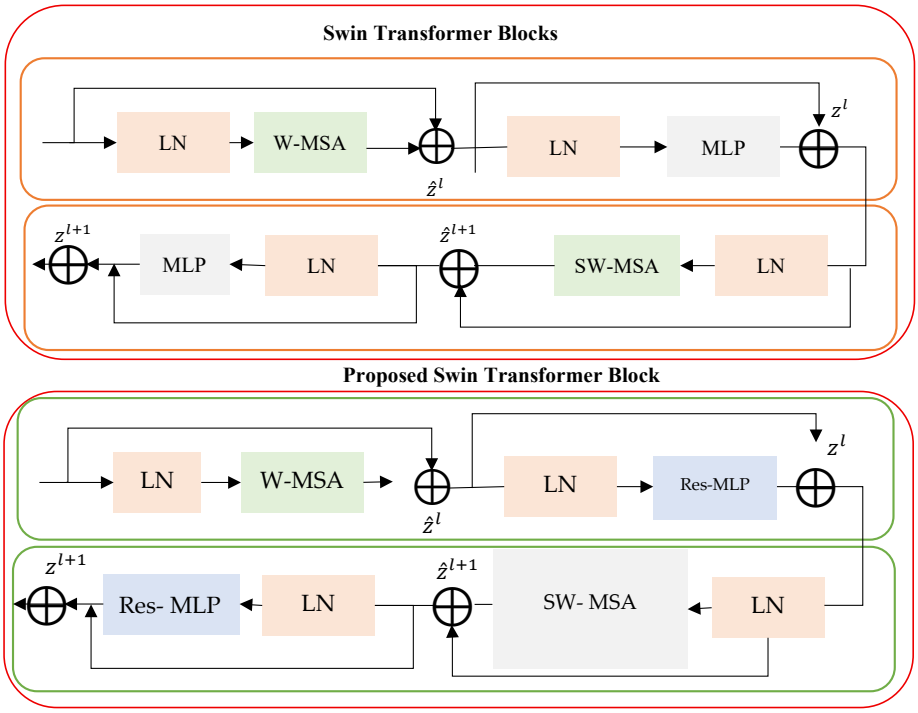


Figure 5. Structural Comparison of the Standard Swin Transformer model and the proposed Swin Transformer model.

3.4.2. UNETR Decoder

The decoder of the UNETR will be tasked with the reconstruction of the segmentation mask based on the feature representations generated by the Swin Transformer encoder which are encoded. It does so by gradually upsampling the high-level feature maps and incorporating the multiscale contextual information with skip connections, which enables a combination of local and global features to be used in the precise segmentation. The upsampled feature maps are computed at every stage of decoding with the help of the transposed convolution (UpConv) operation, which is mathematically stated as Equation (7).

$$f_{dec_i} = U_P Conv(f_{enc_i}) \quad (7)$$

where f_{enc_i} denotes the feature maps of the corresponding encoder stage, and f_{dec_i} is the upsampled features of the decoder. To successfully combine low-level spatial information with high-level semantic information, skip connections are used between the encoder and decoder steps, and they are defined as Equation (8).

$$s_i = f_{dec_i} \oplus f_{enc_{i-1}} \quad (8)$$

where $\oplus$ is the concatenation function, $f_{enc_{i-1}}$ represents the feature maps of the encoder stage before it, and s_i is the combined feature map upon which the decoder proceeds to the next stage of decoding. The combination of these two allows the network to store fine-grained structural information like the tumor boundary and use the global context represented by the encoder. A decoder produces the segmentation output with the aid of a softmax classifier that produces voxel-wise probability maps of each tumor class computed as Equation (9).

$$\hat{Y} = \text{softmax}(W_{s^{s_n}} + b_s) \quad (9)$$

where s_n is the combined fused feature map of the decoder stage, and W_s and b_s are classifier weights and biases that can be learnt. With this upsampling, skip connections, and softmax classification, the UNETR decoder is good at reconstructing fine-scale segmentation masks

that maintain both spatial and semantic accuracy, guaranteeing the proper delineation of heterogeneous parts of brain tumors and making the most of the rich multiscale information offered by the transformer encoder.

Figure 6 presents the overall architecture of the proposed Swin-UNETR-GJO framework designed for accurate multi-modal brain tumor segmentation. The network follows an encoder–decoder paradigm, where the encoder progressively extracts hierarchical feature representations from the input multi-modal MRI images using stacked convolutional layers followed by Swin Transformer blocks. These transformer blocks employ window-based self-attention to capture both local and long-range contextual dependencies, which are crucial for modeling complex tumor structures. The GJO module is integrated to automatically optimize key architectural hyperparameters, including attention window sizes, embedding dimensions, and skip-connection thresholds, thereby enhancing feature learning adaptability across varying tumor morphologies. In the decoder stage, transposed convolution layers perform progressive upsampling, while skipping connections concatenate corresponding encoder features to preserve fine-grained spatial details. Attention maps generated at the output stage highlight discriminative tumor regions, providing additional interpretability.

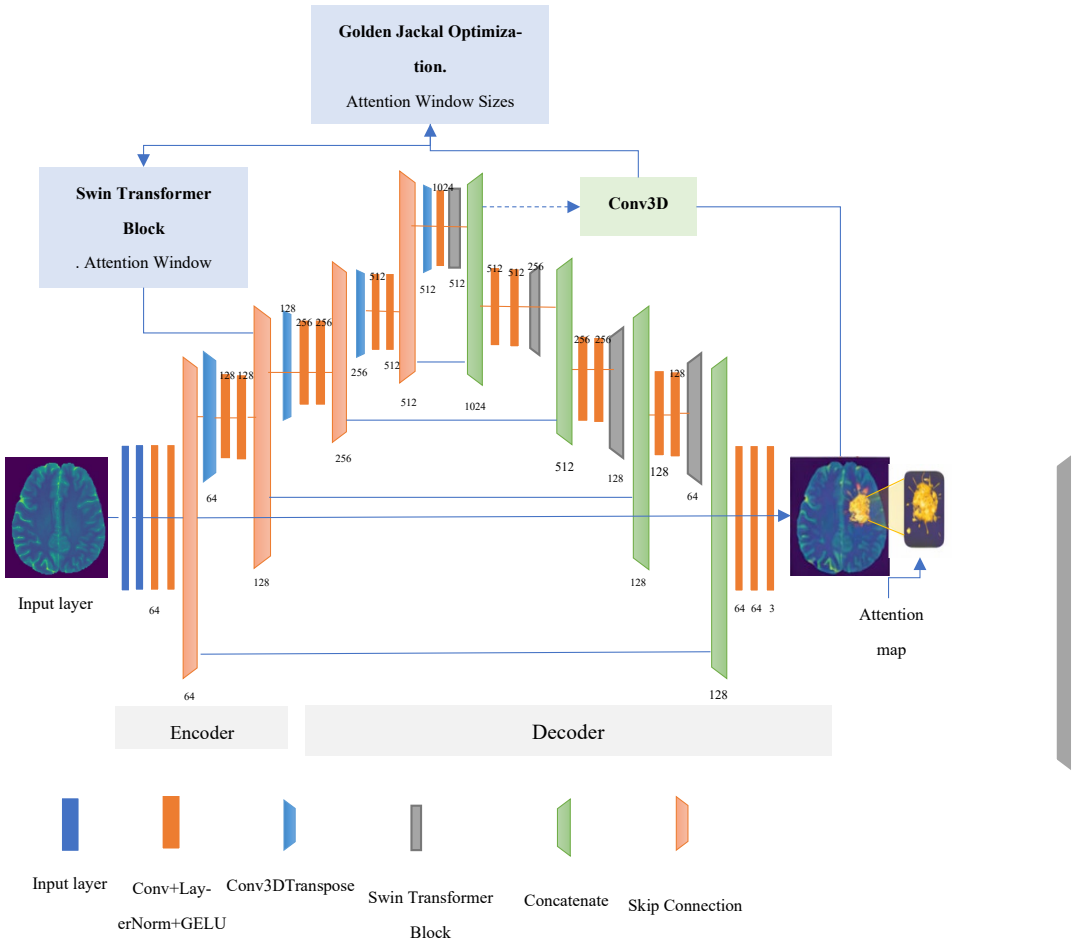


Figure 6. Overview of the Proposed Swin-UNETR-GJO Architecture for Multi-Modal Brain Tumor Segmentation. The color coding represents different components of the proposed architecture. Dark blue indicates the input layer, orange represents convolution with Layer Normalization and GELU activation, light blue denotes 3D transposed convolution, grey represents Swin Transformer blocks, green indicates feature concatenation, and light orange represents skip connections. Blue arrows indicate feature flow, while the highlighted yellow regions in the attention map indicate areas receiving greater model attention.

3.4.3. Hyperparameter Optimization for GJO

To improve the performance of the proposed segmentation framework and reduce the manual process of tuning hyperparameters, important parameters, including the size of the attention window, embedding dimension, dropout ratio, and skip-connection thresholds, were automatically optimized by the GJO algorithm. An optimization procedure is informed by a fitness function that is defined by the DSCs.

The GJO fitness function can be explicitly defined as the mean Dice across the three evaluated tumour regions represented as Equation (10).

$$F(\theta) = \frac{1}{3}[DSC_{WT}(\theta) + DSC_{TC}(\theta) + DSC_{ET}(\theta)] \quad (10)$$

where θ represents a candidate hyperparameter configuration, while WT, TC, and ET denote whole tumour, tumour core, and enhancing tumour, respectively.

The reduction of this fitness function is an effective way to maximize segmentation accuracy. The GJO algorithm models the foraging and adaptive behavior of golden jackals in two key behavioral stages, namely exploration and exploitation. In the exploration stage, jackals search the search space by stepping out of their current positions to find new hyperparameter configurations, which is formulated as Equation (11).

$$X_{t+1} = X_t + r_1.|X_t - X_{best}| \quad (11)$$

where X_t is the existing candidate solution, X_{best} is the most satisfactory solution yet discovered, and r_1 is a random coefficient that regulates the magnitude of movement. During the exploitation stage, jackals narrow their search with a movement towards the most familiar solution, which is modelled as Equation (12).

$$X_{t+1} = X_{best} - r_2.|X_t - X_{best}| \quad (12)$$

where r_2 adjusts the convergence intensity. The algorithm is dynamic in the search intensity and is controlled by the coefficient A that exponentially decreases with the number of iterations t out of the total number of iterations T, and is defined as Equation (13).

$$A = 2_{e^{-t/T}} \quad (13)$$

To improve global search and prevent local optima, GJO uses Lévy flight behavior, which adds randomized long-distance hyperparameter space moves as shown in Equation (14).

$$X_{t+1} = X_t + A.Levy(\lambda) \quad (14)$$

and the Lévy distribution, which has heavy tails, is mathematically defined as in Equation (15).

$$\text{Levy}(\lambda) = \left[\frac{\Gamma(1+\lambda)\sin\left(\frac{\pi\lambda}{2}\right)}{\Gamma\left(\frac{1+\lambda}{2}\right)\lambda\, 2^{(\lambda-1)/2}}\right]^{1/\lambda} \quad (15)$$

Here, λ is the Lévy distribution exponent controlling the step-length characteristics, typically selected within $1 < \lambda \leq 2$. $\Gamma(\cdot)$ denotes the Gamma function, which generalizes the factorial function to continuous values. $\sin(\pi\lambda/2)$ is the trigonometric term governing the scaling of the heavy-tailed distribution. The term $2^{(\lambda-1)/2}$ provides the required normalization, while the exponent $1/\lambda$ scales the expression according to the selected Lévy exponent.

Such a stochastic process enables the optimizer to search remote areas of the search space efficiently. The model is able to find the optimized hyperparameter settings by using GJO as part of Swin-UNETR, which balances exploration and exploitation dynamically, resulting in maximizing the performance of segmentation, minimal manual configuration effort, and high-accuracy brain tumor segmentation.

3.5. Loss Function

The suggested segmentation model uses a hybrid loss function, which is a combination of Dice loss and cross-entropy loss, to optimize the regional overlap and voxel-wise classification accuracy at the same time. This joint loss function is given as Equation (16).

$$L = \propto L_{dice} + (1-\propto)L_{ce} \tag{16}$$

where α is a weighting factor that equalizes the contribution of Dice loss (L_{dice}) and cross-entropy loss (L_{ce}) based on the training goals. The Dice loss compares the similarity of the segmentation prediction $\hat{Y}$ and ground truth Y, which is essentially the evaluation of the overlap of tumor regions, and is represented as Equation (17).

$$L_{dice} = 1 - \frac{2\sum_i p_i g_i}{\sum_i p_i^2 + \sum_i g_i^2 + \in} \tag{17}$$

where p_i and g_i indicate the expected probability of voxel i and the actual label, respectively, and ϵ is a small constant to avoid dividing by 0. Dice loss is especially suitable in the case of class imbalance, such that smaller tumor regions are represented well. To replace this, cross-entropy loss is used to penalize misclassification on a voxel basis, and it is defined as Equation (18).

$$L_{ce} = -\sum_i g_i 1og(p_i) \tag{18}$$

which promotes the correct classification of each voxel independently to improve the precision of boundaries and the reliability of segmentation in general. The B_y incorporating these two loss terms into a hybrid formulation, the model has the advantage of both overlap-based optimization and voxel-wise supervision, resulting in better delineations of tumor subregions and accurate segmentation of tumor edges along ambiguous or complex tumor edges.

3.6. Training Optimization Strategy

The proposed Swin-UNETR-GJO framework was trained using the AdamW optimizer, which combines adaptive moment estimation with decoupled weight decay. Unlike conventional L_2 regularization, AdamW applies weight decay directly to the model parameters rather than mixing it with the gradient update, thereby providing more controlled regularization and stable convergence. At training step (t), the first- and second-order moment estimates are computed as in Equations (19) and (20).

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \tag{19}$$

$$\upsilon_t = \beta_2 \upsilon_{t-1} + (1-\beta_2) g_t^2 \tag{20}$$

where g_t is the gradient of the loss at iteration t, model parameter t, m_t is the initial moment (mean) estimate, v_t is the second moment (uncentered variance) estimate, and β_1 and β_2 do the moving averages have exponential decay rates. These moment estimates allow the optimizer to dynamically change the learning rates of every parameter to enhance the stability and convergence rate. To correct initialization bias, the moment estimates are adjusted as in Equation (21).

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \hat{v}_t = \frac{v_t}{1-\beta_2^t}. \tag{21}$$

The AdamW parameter update is then defined as Equation (22).

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} - \eta\lambda\theta_t \tag{22}$$

where η is the learning rate, λ is the weight-decay coefficient, and ϵ is a small constant used for numerical stability. In this study, AdamW therefore enabled adaptive and stable optimization while independently controlling weight decay.

3.7. Explainability with Activation Maps of Attention

In order to support clinical confidence and interpretability of the proposed Swin-UNETR-GJO segmentation framework, attention activation maps are obtained as a result of the transformer attention layers to see which parts of the input MRI have the greatest impact on the model prediction. The attention mechanism calculates a score matrix that measures the relation between query and key vectors in each attention head, which is mathematically defined as Equation (23).

$$\mathrm{A_{att}} = \frac{\mathrm{QK^T}}{\mathrm{d_k}} \tag{23}$$

Q and K are query and key matrices that are based on the feature embeddings. $\mathrm{d_k}$ is the dimensionality of key vectors that normalizes the dot product to stabilize gradients. This attention score matrix will capture the relative significance of the various spatial areas in the input volume. The attention scores are calculated by fusing the original MRI slices to produce heatmaps of important areas for interpretable visualizations, which are formulated as Equation (24).

$$\mathrm{H} = \gamma \mathrm{A_{att}} + (1 - \gamma)\mathrm{I} \tag{24}$$

where I represents the original MRI slice, 0 is a blending factor used to determine the magnitude of the attention overlay, and H is the output attention activation map.

The two other computational explainability measures, RAO and BAA formulated to check whether the attention map is accurate as per the ground truth or not. These are formulated as follows: Let $A(v) \in [0, 1]$ denote the normalized attention value at voxel v, $G_c(v) \in \{0, 1\}$ the ground-truth mask for tumour region $c \in \{W, T, TCET\}$, and τ the attention threshold. The binarized high-attention region is computed using Equation (25).

$$H_\tau(v) = \begin{cases} 1, & A(v) \geq \tau, \\ 0, & A(v) < \tau. \end{cases} \tag{25}$$

The Region-Attention Overlap (RAO) measures the spatial agreement between the high-attention region and the corresponding ground-truth tumour mask computed as Equation (26).

$$\mathrm{RAO}_c = \frac{2 \mid H_\tau \cap G_c \mid}{\mid H_\tau \mid + \mid G_c \mid + \varepsilon} \tag{26}$$

where $\mid \cdot \mid$ denotes the number of voxels and ε is a small constant preventing division by zero. RAO ranges from 0 to 1, where a higher value indicates stronger localization of model attention within the relevant tumour region.

For Boundary-Attention Accuracy (BAA), let ∂H_τ and ∂G_c represent the boundaries of the attention region and ground-truth tumour mask, respectively. Considering a small spatial tolerance r, boundary precision and recall are defined as Equation (27) and Equation (28), respectively.

$$P_{b,c} = \frac{\mid \{v \in \partial H_\tau : d(v \mid \partial G_c) \leq r\} \mid}{\mid \partial H_\tau \mid + \varepsilon} \tag{27}$$

$$R_{b,c} = \frac{\mid \{v \in \partial G_c : d(v \mid \partial H_\tau) \leq r\} \mid}{\mid \partial G_c \mid + \varepsilon} \tag{28}$$

and BAA is computed as in Equation (29).

$$\mathrm{BAA}_c = \frac{2 P_{b,c} R_{b,c}}{P_{b,c} + R_{b,c} + \varepsilon} \tag{29}$$

Here, $d(\cdot, \cdot)$ denotes the minimum Euclidean distance between boundary voxels, and r is the accepted boundary tolerance, for example, 2 mm after isotropic resampling. A BAA

value close to 1 indicates that the model's high-attention boundaries closely correspond to the true tumour boundaries.

The visualization of these heatmaps can be used to show clinicians how the model pays attention to different areas of the brain tumor to make predictions, so that it provides a clear and understandable decision-support tool.

The significance of the explainability module is that it provides visual evidence of the image regions influencing segmentation, enabling clinicians to verify whether model decisions correspond to clinically relevant tumour anatomy or not. The RAO and BAA remain computational measures and require clinician validation before real-world healthcare system applications.

3.8. Evaluation Metrics

The performance of the suggested Swin-UNETR-GJO framework in segmentation was assessed quantitatively based on standard metrics that quantify both the spatial overlap and boundary accuracy, as well as the performance of the model at the voxel scale. To measure the extent of overlap between the predicted segmentation B and ground truth A, the DSC was used and is defined as Equation (30).

$$\mathrm{DSC} = \frac{2|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A}|+|\mathrm{B}|} \tag{30}$$

The larger the DSC, the more the predicted and true tumor regions are in agreement, and thus, it is a common measure of medical image segmentation.

To evaluate the accuracy of tumour boundary delineation, the HD95 was calculated between the predicted and ground-truth segmentation boundaries. HD95 is defined as Equation (31).

$$\mathrm{HD}_{95}(P, G) = \max\left\{Q_{95}, \left(\left\{\min_{g \in G} \| p - g \|_2 \ : \ p \in P\right\}\right) Q_{95}\left(\left\{\min_{p \in P} \| g - p \|_2 \ : \ g \in G\right\}\right)\right\} \tag{31}$$

where P and G represent the sets of boundary points of the predicted segmentation and the ground-truth segmentation, respectively, $\|\cdot\|_2$ denotes the Euclidean distance, and $Q_{95}(\cdot)$ represents the 95th percentile of the corresponding minimum boundary distances. Unlike the conventional Hausdorff Distance, which is determined by the maximum boundary discrepancy, HD95 reduces the influence of isolated outlier voxels and provides a more robust assessment of segmentation boundary accuracy. A lower HD95 value indicates closer agreement between the predicted and ground-truth tumour contours.

Moreover, the quality of voxel-level classification was evaluated with the help of the measures of precision and recall, which were computed as Equation (32) and Equation (33), respectively.

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{32}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{33}$$

TP: true positives, FP: false positives, and FN: false negatives. Precision is a measure of the fraction of correctly predicted tumor voxels among all the predicted voxels, and recall is a measure of the fraction of the true tumor voxels that are correctly identified. The evaluation framework integrates the performance of DSC, HD, precision, and recall to give a comprehensive evaluation of the segmentation performance, which includes the volumetric overlap, boundary accuracy, and voxel-wise classification reliability.

3.9. Parameter Configurations

As summarized in Table 2, the 1251 labeled BraTS 2021 subjects were partitioned into a 1000-subject development cohort and an independent 251-subject test cohort. GJO optimization used only the development data, while the final test set remained completely unseen during hyperparameter selection and was used only for the final evaluation. The learning rate of 1×10^{-4} represents the initial baseline configuration before hyperparameter optimization. The GJO algorithm subsequently identified an optimal learning rate of 3×10^{-5}, which was used for training the final proposed Swin-UNETR-GJO model.

Table 2. Experimental configuration Details.

Component	Parameter	Setting/Value
Dataset	Dataset	BraTS 2021
	Labeled subjects used	1251 subjects
	Splitting level	Subject-wise; no slices/patches from one subject occurred in multiple partitions.
Independent evaluation split	Development set	1000 subjects
	Internal hold-out test cohort	251 subjects
	Test-set use	Used only once after architecture and hyperparameters were fixed
GJO tuning split	Training subset for GJO	800 subjects from the development set
	GJO fitness-validation subset	200 subjects from the development set
	Fitness data	Only the 200-subject tuning-validation subset
	Test data used by GJO	None
Five-fold cross-validation	Data included	Full 1000-subject development cohort after hyperparameters were fixed
	Number of folds	5
	Subjects/fold	800 training + 200 validations
	Purpose	Stability/internal robustness analysis
	GJO repeated within each fold	No; selected GJO parameters were fixed before five-fold CV

Table 2. *Cont.*

Component	Parameter	Setting/Value
GJO configuration	Population size	8 jackals'/candidate solutions
	Maximum iterations	20
	Maximum candidate evaluations	160
	Initialization	Uniform random initialization within predefined bounds
	Stopping criterion	20 iterations or no fitness improvement for 3 consecutive iterations
	Random seeds	42, 52, 62, 72, and 82 for repeated/fold experiments
	Fitness function	Mean validation Dice over WT, TC, and ET regions
GJO search space	Initial learning rate	1×10^{-5}–5×10^{-4}, log-uniform
	Attention window size	{3, 5, 7}
	Embedding dimension	{96, 112, 128, 160, 192}
	Dropout rate	0.05–0.20
	Weight decay	1×10^{-6}–1×10^{-4}, log-uniform
Input configuration	MRI modalities	T1, T1ce, T2, and FLAIR
	Input channels	4
	Training patch size	96 × 96 × 96 voxels
	Resampled voxel spacing	$1 \times 1 \times 1$ mm^3
	Batch size	2 volumes/patches
Optimization	Optimizer	AdamW
	Maximum epochs	200
	Initial learning rate	1×10^{-4}
	Learning-rate schedule	Cosine annealing
	Minimum learning rate	1×10^{-6}
	Weight decay	1×10^{-5}
	Early stopping	Validation Dice; patience = 20 epochs
	Model-selection criterion	Highest mean validation Dice
Loss	Training loss	Dice + cross-entropy loss
	Dice-loss weight	0.5
	Cross-entropy weight	0.5

4. Experimental Results

The experimental analysis of the suggested Swin-UNETR-GJO framework on the BraTS 2021 database is presented in detail within this section.

4.1. Quantitative Segmentation Performance

The quantitative analysis of the suggested Swin-UNETR-GJO framework proves that it has a high capacity to precisely divide the multi-regional brain tumor components in the BraTS 2021 dataset. With the combination of the hierarchical self-attention of the Swin Transformer and the encoder–decoder advantages of UNETR, and the refinement of adaptive hyperparameters through the GJO algorithm, the model was able to achieve high performance in the three main tumor subregions, namely, WT, TC, and ET.

An important measure of segmentation quality is the DSC, which measures the overlap between predicted and ground-truth tumor regions. The model had a DSC of 0.91 with WT, 0.89 with TC, and 0.87 with ET, which means that the boundary is detected accurately even in nonhomogeneous or low-contrast tumor regions. These findings indicate the high ability of the model to represent both the global contextual characteristics and fine-grained structural information. Likewise, the Jaccard Index is another region-overlap measure, which gave results of 0.84, 0.80, and 0.78, respectively, indicating great consistency in the regions.

The model is reliable because the sensitivity values of 0.93 (WT), 0.90 (TC), and 0.88 (ET) indicate that the model is accurate in the detection of tumor voxels, particularly in complex tumors that have diffuse infiltrative areas. The specificity scores are high (0.98, 0.97, and 0.96), which also confirms that the model can be used to reduce false positives, so that healthy tissues are not mistakenly identified as tumor. Also, the HD95, which was 6.1 mm in WT, 7.4 mm in TC, and 8.2 mm in ET, means that the predictions and the ground truth have accurate boundary alignment with comparatively low spatial variation, even in smaller or irregularly shaped tumor substructures.

As shown in Table 3, the Swin-UNETR-GJO model demonstrates high performance in terms of segmentation in all tumor regions. WT has the greatest accuracy with a Dice of 0.91, then TC (0.89) and ET (0.87). Sensitivity and specificity are also high in all regions, which implies a good level of detection and low false positives. The HD95 rises between WT (6.1 mm) and ET (8.2 mm), which is due to the higher complexity of the boundary of smaller tumor subregions. All in all, the model offers strong and consistent segmentation of WT, TC, and ET.

Table 3. Swin-UNETR-GJO tumor-wise performance.

Metric	WT	TC	ET
DSC	0.91	0.89	0.87
Jaccard Index	0.84	0.80	0.78
Sensitivity	0.93	0.90	0.88
Specificity	0.98	0.97	0.96
HD (95%)	6.1 mm	7.4 mm	8.2 mm

Figure 7 presents a qualitative evaluation of the proposed Swin-UNETR-GJO framework by comparing ground-truth tumor annotations with the corresponding model predictions across multiple representative FLAIR MRI slices from a BraTS 2021 subject. Each row illustrates a different axial slice index, capturing the progressive spatial evolution of the tumor from its initial appearance to its maximal extent and subsequent reduction. The left column shows the original FLAIR images, which effectively highlight hyperintense tumor-associated edema and infiltrative regions. The middle column displays the expert-

annotated ground-truth masks, serving as the clinical reference standard, while the right column presents the segmentation results generated by the proposed model.

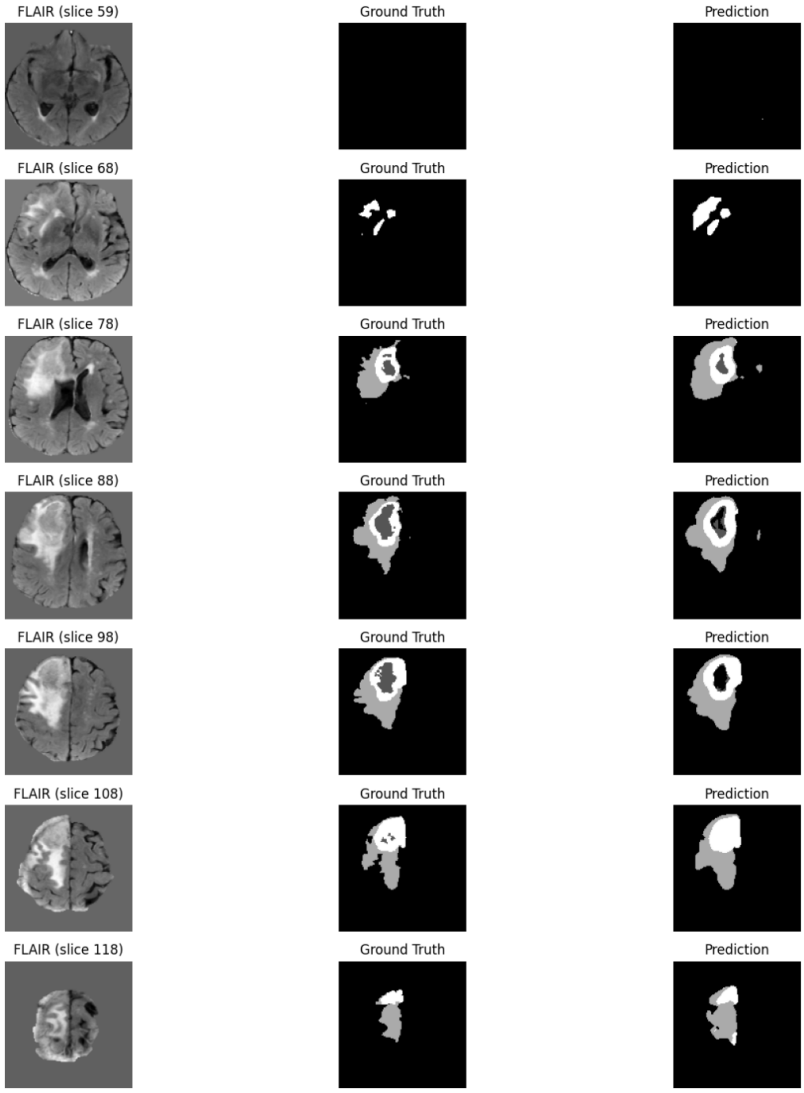


Figure 7. Qualitative Comparison of Ground-Truth and Predicted Tumor Segmentation Across Representative FLAIR MRI Slices.

4.2. Comparison to the Traditional Methods

In order to comprehensively evaluate the performance improvement of the proposed Swin-UNETR-GJO framework, the framework was compared with a number of popular state-of-the-art segmentation models, such as UNet, UNETR, Swin Transformer, and Swin-UNETR. For a fair and reproducible comparison, the baseline models were configured as UNet (five encoder–decoder levels with feature channels of 16, 32, 64, 128, and 256), UNETR (12 Transformer layers, 768-dimensional embeddings, and 12 attention heads), Swin Transformer (four hierarchical stages with depths of 2, 2, 6, and 2 and attention heads of 3, 6, 12, and 24), and Swin-UNETR (four stages with depths of 2, 2, 2, and 2 and attention heads of 3, 6, 12, and 24); all baseline hyperparameters were kept fixed at their reference configurations, while GJO-based hyperparameter optimisation was applied exclusively to the proposed Swin-UNETR-GJO model.

The traditional UNet demonstrated moderate results because of its inability to represent long-range spatial dependencies, which led to lower DSC scores (0.85 WT, 0.83 TC, 0.78 ET) and high HD95 (12.8 mm), and it failed to address heterogeneous tumor boundaries.

UNETR enhanced these measurements using better contextual knowledge yet still exhibited difficulties in segmenting ET areas correctly. Swin Transformer also improved the long-range extraction of features based on hierarchical attention mechanisms and provided better segmentation consistency, and Swin-UNETR provided even better results with Swin attention and an encoder–decoder architecture. Nevertheless, the proposed method has the highest DSC values in all tumor subregions, 0.91 in WT, 0.89 in TC, and 0.87 in ET, and the minimum HD95 score of 6.1 mm.

As per the results of Table 4, the suggested Swin-UNETR-GJO model performed better than all current state-of-the-art segmentation models in all tumor regions. It has the highest Dice scores of WT (0.91), TC (0.89), and ET (0.87), and the minimum HD95 (6.1 mm), which means that it has better boundary accuracy.

Table 4. Comparison to the state-of-the-art segmentation models.

Model	WT DSC	TC DSC	ET DSC	HD95 (mm)
UNet	0.85	0.83	0.78	12.8
UNETR	0.88	0.86	0.82	9.5
Swin-Transformer	0.89	0.87	0.83	8.7
Swin-UNETR	0.90	0.88	0.85	7.9
Swin-UNETR-GJO (Proposed)	0.91	0.89	0.87	6.1

An exploratory Friedman rank analysis across the four reported performance dimensions showed a significant difference in overall model rankings. ($\chi^2(4) = 16.00$, $p = 0.003$), with Kendall's $W = 1.00$, indicating complete consistency in the ranking of the competing methods. Swin-UNETR-GJO obtained the best rank for every evaluated metric across the compared models.

Figure 8 contrasts the Dice scores of five segmentation models of WT, tumor core, and ET regions. The Swin-UNETR-GJO (Proposed) model has the highest DSC values in all tumor subregions and the lowest HD95 distance, which means that it has high segmentation accuracy and boundary precision. There is a gradual reduction in performance in Swin-UNETR, Swin-Transformer, UNETR, and UNet, which points to the overall benefit of optimization-based transformer-based architectures.

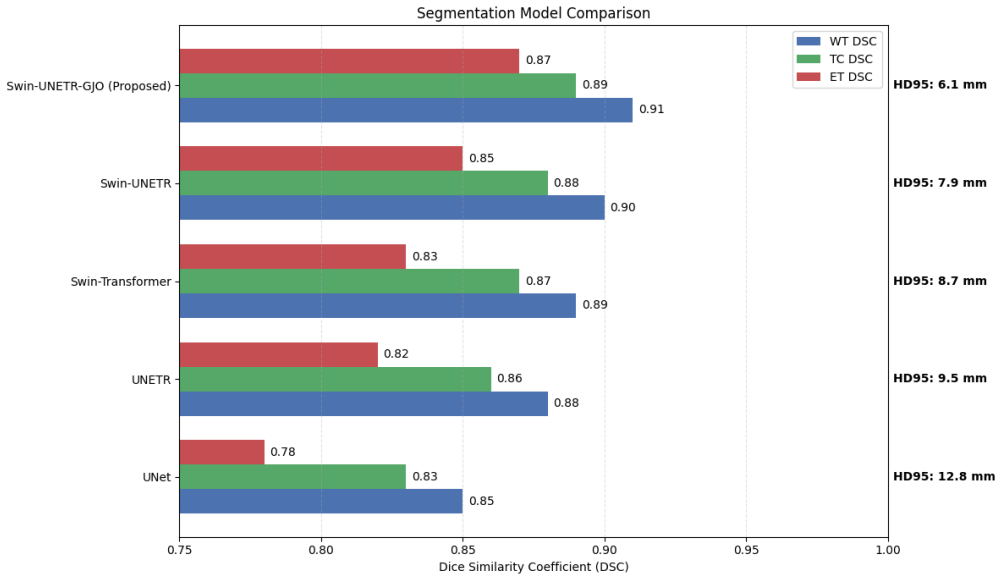


Figure 8. Comparison with current state-of-the-art segmentation models.

4.3. The Performance of the GJO Hyperparameter Optimization

The GJO algorithm was used to optimize the hyperparameters of the Swin-UNETR framework to achieve high segmentation results across all tumor regions. GJO was able to effectively balance exploration and exploitation because it used intelligent hunting behavior as a search strategy, unlike traditional gradient-based or static search strategies, which explored the hyperparameter space in an adaptive way. In the optimization, the algorithm was also used to find the most appropriate values of a variety of important parameters. Such a multi-parameter optimization configuration allowed GJO to change both architectural and training-related parts in a concerted way, which guaranteed more stable convergence and better model generalization.

The GJO optimization process indicates a consistent and gradual improvement in model fitness, which rises with the number of iterations (0.71 at the start of the optimization process) to a maximum of 0.91 at the end of the optimization process (iteration 20). The algorithm is trained to be close to smaller window sizes, larger embedding dimensions, reduced dropout rates, and a reduced learning rate. This is a successful search strategy in which the model is more stable and expressive and reduces overfitting. The optimal performance is attained at iteration 20 and window size 3, embedding dimension 192, dropout 0.05, and learning rate 3×10^{-5} (Table 5).

Table 5. GJO Optimization Convergence Statistics.

Iteration	Fitness Score	Window Size	Embedding Dim	Dropout	Learning Rate
1	0.71	7	96	0.15	1×10^{-4}
5	0.78	5	112	0.12	8×10^{-5}
10	0.84	5	128	0.10	6×10^{-5}
15	0.88	3	160	0.08	4×10^{-5}
20	0.89	3	192	0.05	3×10^{-5}

Table 6 shows that all optimization strategies improved segmentation performance over the fixed-hyperparameter baseline, while GJO achieved the best overall results. The proposed GJO obtained the highest DSC values for WT (0.910 ± 0.005), TC (0.890 ± 0.005), and ET (0.870 ± 0.007), with a high mean DSC/fitness score of 0.890. The GJO slightly outperformed Random Search, Bayesian Optimization (BO), and PSO, indicating more effective hyperparameter selection without additional computational expense.

Table 6. Comparative analysis of optimization methods.

Strategy	WT DSC	TC DSC	ET DSC	Mean DSC/Fitness
Fixed hyperparameters	0.900 ± 0.006	0.880 ± 0.006	0.850 ± 0.008	0.877
Random Search	0.903 ± 0.006	0.883 ± 0.006	0.858 ± 0.008	0.881
BO	0.906 ± 0.005	0.886 ± 0.005	0.863 ± 0.007	0.885
PSO	0.908 ± 0.005	0.887 ± 0.005	0.865 ± 0.007	0.887
GJO (Proposed)	0.910 ± 0.005	0.890 ± 0.005	0.870 ± 0.007	0.890

4.4. Ablation Study

In order to measure the contributions of the Swin Transformer module, the UNETR backbone, and the GJO algorithm individually, an ablation study was performed comprehensively. This analysis was to separate the effects of each architectural and optimization element on the total accuracy of the segmentation of the proposed model. The most basic

configuration, which only included the UNETR framework, had DSC values of 0.88 (WT), 0.86 (TC), and 0.82 (ET), which showed good performance in terms of feature extraction using transformers but was poor at extracting multi-scale contextual information. Another setup, Swin-UNETR with fixed (non-optimized) hyperparameters, had DSC values of 0.90 (WT), 0.88 (TC), and 0.85 (ET), which suggests that the architectural improvements alone have a significant positive impact on the segmentation ability. Nevertheless, the greatest enhancement was made when GJO-based hyperparameter optimization was used on the Swin-UNETR framework. Swin-UNETR-GJO provided the highest DSC scores, including 0.91 (WT), 0.89 (TC), and 0.87 (ET).

Table 7 reveals that all the added components are associated with performance improvements in all tumor regions. UNETR has a good segmentation accuracy at the baseline.

Table 7. Model Component Ablation Analysis.

Configuration	WT DSC	TC DSC	ET DSC
UNETR only	0.88	0.86	0.82
Swin-UNETR + Fixed HP	0.90	0.88	0.85
Swin-UNETR-GJO (Final)	0.91	0.89	0.87

4.5. Computational Efficiency

Table 8 presents the computational performance in terms of efficiency. The time taken to train is 4.8 min per epoch, and it takes approximately 16 h to train 200 epochs using a high-performance GPU. Inference can be done on each MRI volume in only 0.74 s. The level of GPU utilization is also high (93), which means that the hardware is used well, and the maximum memory consumption is 11.2 GB, which is a good balance between performance and resource requirements.

Table 8. Computational Performance in terms of Efficiency.

Parameter	Value
Training Time per Epoch	4.8 min
Total Training Time (200 epochs)	16 h
Inference Time per Volume	0.74 s
GPU Utilization	93%
Peak Memory Usage	11.2 GB

Although the model attained its minimum validation loss at an earlier epoch, training was continued for the complete 200-epoch configuration to assess training behavior and convergence stability. The reported training time represents the complete training process and does not include the additional computational cost of GJO-based hyperparameter optimization.

4.6. The Performance of Cross-Validation

In order to evaluate the strength and generalizability of the suggested Swin-UNETR-GJO framework, a five-fold cross-validation was performed on the BraTS 2021 dataset. This approach also ensured that model performance was not biased by the particular training or validation splits and also gave a full estimate of stability in the heterogeneous patient data. The model performed well in terms of the DSC scores across all tumor subregions (WT, TC, ET) in each fold. In particular, the value of WT DSC was between 0.90 and 0.92, TC DSC between 0.88 and 0.89, and ET DSC between 0.86 and 0.88, which showed little difference across folds. The average DSC values were 0.91 0.01 (WT), 0.88 0.005 (TC),

and 0.87 0.008 (ET), which showed great cross-fold stability and reproducibility. The low standard deviations ensure that the model has the same level of segmentation even when the size, shape, and contrast of tumors in different patients vary.

Table 9 shows that the Swin-UNETR-GJO model provides very stable and consistent results in terms of segmentation across all folds. WT DSC is between 0.90 and 0.92, TC is between 0.88 and 0.89, and ET is between 0.86 and 0.88. The low standard deviations (WT: 0.01, TC: 0.005, ET: 0.008) imply that there is a slight variation in performance, which proves that the proposed framework has a high generalization potential and is efficient across different samples.

Table 9. 5-Fold Cross-Validation Results.

Fold	WT DSC	TC DSC	ET DSC
Fold-1	0.90	0.88	0.86
Fold-2	0.91	0.88	0.87
Fold-3	0.92	0.89	0.88
Fold-4	0.91	0.89	0.87
Fold-5	0.90	0.88	0.86
Mean ± SD	0.91 ± 0.01	0.88 ± 0.005	0.87 ± 0.008

Figure 9 demonstrates that there is strong and consistent performance in all folds, and the Dice scores of WT, TC, and ET change only slightly. The overall mean accuracy of the model with a very low standard deviation is 0.91 (WT), 0.88 (TC), and 0.87 (ET), which implies that the model is very reliable and strong in various data splits.

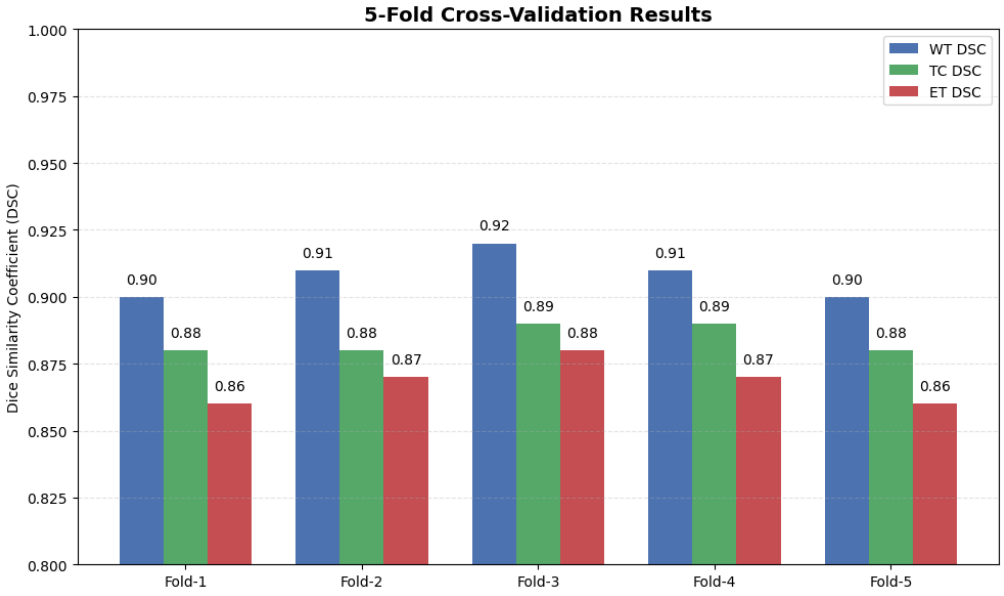


Figure 9. DSC comparison among fivefold across modalities.

The statistical stability of the proposed Swin-UNETR-GJO framework was evaluated using the five-fold cross-validation results. The model achieved mean DSC values of 0.908 ± 0.008 for WT, 0.884 ± 0.005 for TC, and 0.868 ± 0.008 for ET. The corresponding 95% confidence intervals were 0.898–0.918, 0.877–0.891, and 0.858–0.878, respectively. The coefficients of variation were below 1% for all tumour regions, indicating low variability and consistent performance across the different folds. An overall Friedman repeated-measures test identified a statistically significant difference among WT, TC, and ET segmentation performance ($\chi^2 = 10.00$, $p = 0.0067$). WT consistently achieved the highest DSC, whereas

ET showed comparatively lower performance, reflecting the greater difficulty of segmenting smaller and more heterogeneous enhancing tumour regions.

4.7. Results of Explainability and Attention Visualization

In order to support both clinical trust and interpretability, the suggested Swin-UNETR-GJO framework was tested in terms of explainability with visualization of attention and quantitative measures of interpretability. The hierarchical mechanism of self-attention of the Swin Transformer enabled the model to selectively attend to relevant tumor regions, which can provide an understanding of the spatial significance of features used to make segmentation decisions. The maps of attention obtained following the inference showed that the model was able to localize the boundaries of tumors in an accurate manner, successfully emphasizing the WT, TC, and ET areas but minimizing spurious activation in the surrounding healthy tissues. To estimate such observations, two measures were proposed: RAO, which quantifies the spatial agreement of high-attention regions and ground-truth tumor masks, and BAA, which quantifies the accuracy of attention to true tumor boundaries. As Table 10 demonstrates, the values of RAO were 0.84 (WT), 0.81 (TC), and 0.78 (ET), whereas the BAA scores were 0.88, 0.85, and 0.83, respectively, which proves that the model covers significant areas.

Table 10. Explainability Evaluation.

Metric	WT	TC	ET
Region-Attention Overlap (RAO)	0.84	0.81	0.78
Boundary-Attention Accuracy (BAA)	0.88	0.85	0.83

The RAO scores show that there is a high correlation between model-oriented regions and real tumor regions, particularly in WT (0.84). BAA is also high, which proves accurate edge-oriented attention.

4.8. Loss Curve and Convergence Stability Comparison

The convergence behavior of the proposed Swin-UNETR-GJO was tested against the traditional baseline models to determine both the efficiency of the training and the stability of the gradient. The loss curves of training were examined to identify the number of epochs to converge, the lowest validation loss, and the gradient update smoothness during the training process. According to the summary, the classical UNet model took 78 epochs to converge and achieved a validation loss of 0.214 and mediocre gradient stability, indicating slower learning of complicated tumor borders. UNETR used a transformer backbone, which decreased convergence to 65 epochs and minimized validation loss to 0.196, with stable gradients because it models long-range features better. The Swin-UNETR hybrid was also able to learn faster and achieve a validation loss of 0.183 in 58 epochs with the help of hierarchical attention and encoder–decoder fusion. It is worth noting that the Swin-UNETR-GJO model was the quickest converging one, taking 41 epochs to achieve its lowest validation loss of 0.159, and it had very stable gradient behavior during training. This more rapid and consistent convergence can be explained by the GJO algorithm that actually optimized hyperparameters, reduced oscillations of the loss landscape, and enhanced the learning rate.

Table 11 presents the comparison of four segmentation models based on the training efficiency and stability. Swin-UNETR-GJO performs best, converging in the shortest time of 41 epochs, with the lowest validation loss of 0.159 and very stable gradients. The UNet that is traditional converges the slowest, with a higher loss (0.214) and only moderate

gradient stability. UNETR and Swin-UNETR have average performance, and Swin-UNETR converges faster and slightly better than UNETR. In general, Swin Transformer and GJO are effective in enhancing the convergence rate, accuracy, and training stability.

Table 11. Convergence Analysis.

Model	Epochs to Converge	Validation Loss
UNet	78	0.214
UNETR	65	0.196
Swin-UNETR	58	0.183
Swin-UNETR-GJO	41	0.159

5. Discussion

The findings of this paper prove that the suggested Swin-UNETR-GJO model represents a meaningful contribution to automated brain tumor segmentation, especially in contrast to the reported CNN-based and transformer-based models in the literature. Conventional CNN models, including U-Net and its derivatives, are generally not very good at long-range contextual interactions, and thus do not perform optimally on tumors with irregular boundaries and with heterogeneous intensity distributions. Conversely, hierarchical self-attention of Swin Transformer enables Swin-UNETR-GJO to encode spatial relationships between multi-scale tumor regions, which leads to high Dice coefficients on WT (0.91) and ET (0.87) regions. These results are competitive and, in some cases, outperform the results of state-of-the-art transformer-based segmentation systems, including Swin-UNETR, UNETR, TransBTS, and Swin-BTS, which typically achieve whole-tumor Dice scores of between 0.88 and 0.90, and enhancing-tumor scores of between 0.82 and 0.86 on the BraTS datasets.

The addition of the GJO algorithm, which automatically picks optimal hyperparameters, such as the size of the attention window, embedding dimensions, and thresholds on skip-connections, is a significant methodological addition that increases the flexibility of the models and reduces the number of hyperparameters that have to be optimized manually, which is generally subjective and computationally expensive. This automatic optimization allows the model to be generalized to the various tumor morphologies found within the BraTS dataset, especially in the segmentation of ET regions, which are still difficult to find in previous studies because of their heterogeneous appearance and their generally smaller volume.

The drawback of the suggested Swin-UNETR-GJO model is that it relies on the BraTS 2021 dataset. The model has not been trained and tested on any other dataset, so its generalization capabilities to other hospitals, MRI scanners, and imaging protocols are unclear. In practice, clinical implementation would have to be validated on multi-institutional heterogeneous data. Another weakness is the computational cost of the GJO-based hyperparameter optimization. Even though GJO increases performance by automatically tuning architectural parameters that are critical to the system, the optimization procedure itself is slow and computationally intensive, which can limit its applicability in low-resource research settings.

Networks using transformers demand large memory and GPUs by nature. This is computationally intensive and therefore difficult to deploy in real time, particularly in low-resource clinical environments or mobile health systems where the hardware is constrained. The other limitation is the existing mechanism of explainability, which is mainly based on attention maps. Although these maps can be very useful in terms of localizing tumors, they might not be as interpretable as clinicians need to make critical decisions. The model has not been tested on multi-modal data other than MRI or longitudinal data.

The model can be strengthened in the future through testing on data from other hospitals to enhance its validity in a real clinical environment. Lightweight or hybrid

optimization can be used to make the GJO optimization process less computationally expensive and faster. Further explainability of the model can be improved with the use of advanced visualization techniques or other explainability measures to further support the AI-assisted diagnosis with evidence. Furthermore, boundary-refinement or uncertainty-aware modules can be developed so that performance on small, complex, and highly infiltrative tumors can be enhanced.

6. Conclusions

This paper introduces Swin-UNETR-GJO, an explainable and optimization-based DL model that tries to overcome the long-standing problem of brain tumor segmentation in multi-modal MRI. Conventional segmentation techniques usually have difficulties with very irregular tumor morphologies, non-uniform intensity distributions, and unclear boundaries. The proposed model can be used to achieve both global and fine-grained spatial details needed to delineate tumors accurately by combining the hierarchical self-attention of the Swin Transformer with the encoder–decoder architecture of UNETR. One of the main contributions of this work made is the implementation of the GJO algorithm that automatically adjusts important hyperparameters (the size of the attention window, the size of the embedding dimensions, and the threshold of skip connections). This automatic optimization increases the adaptability of models and guarantees the balance between the representation of tumor regions with different scales. The experimental analysis of the BraTS 2021 dataset shows high performance of the proposed model with Dice similarity scores of 0.91 on the WT and 0.87 on the enhancing tumor, better than most existing transformer- and CNN-based methods reported in the literature. In addition to segmentation accuracy, the proposed model considers the interpretability of its predictions. The explainability findings indicate that the Swin-UNETR-GJO framework produces attention patterns that generally align with tumor regions and boundaries. These observations provide preliminary insights into the model's segmentation behavior and suggest its potential for supporting more interpretable brain tumor segmentation. However, further validation is required to establish the reliability and clinical relevance of these explanations.

Author Contributions: A.G. and P.D.S. contributed equally to the conception and design of the study, methodology, material preparation, data collection, data analysis, interpretation of results, and manuscript preparation. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The BraTS 2021 dataset used in this study is publicly available through the official Kaggle data repository, Available at: https://www.kaggle.com/datasets/dschettler8845/brats-2021-task1 (accessed on 6 September 2026). All additional experimental settings, preprocessing procedures, data partitions, and model configurations required to reproduce the study are described in the manuscript. The source code and processed code developed for the proposed Swin-UNETR-GJO framework are available from the corresponding author upon reasonable request.

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

References

1. Khalighi, S.; Reddy, K.; Midya, A.; Pandav, K.B.; Madabhushi, A.; Abedalthagafi, M. Artificial intelligence in neuro-oncology: Advances and challenges in brain tumor diagnosis, prognosis, and precision treatment. *npj Precis. Oncol.* **2024**, *8*, 80. [CrossRef]
2. Meola, A.; Rao, J.; Chaudhary, N.; Sharma, M.; Chang, S.D. Gold nanoparticles for brain tumor imaging: A systematic review. *Front. Neurol.* **2018**, *9*, 328. [CrossRef]

3. Alsajri, A.; Hacimahmud, A.V. Review of deep learning: Convolutional neural network algorithm. *Babylon. J. Mach. Learn.* **2023**, *2023*, 19–25. [CrossRef]
4. Lee, H.; Kwon, H. Going deeper with contextual CNN for hyperspectral image classification. *IEEE Trans. Image Process.* **2017**, *26*, 4843–4855. [CrossRef]
5. Jamil, S.; Jalil Piran, M.; Kwon, O.J. A comprehensive survey of transformers for computer vision. *Drones* **2023**, *7*, 287. [CrossRef]
6. Sendak, M.; Elish, M.C.; Gao, M.; Futoma, J.; Ratliff, W.; Nichols, M.; Bedoya, A.; Balu, S.; O'Brien, C. "The human body is a black box" supporting clinical decision-making with deep learning. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 99–109.
7. Bhati, D.; Neha, F.; Amiruzzaman, M. A survey on explainable artificial intelligence (XAI) techniques for visualizing deep learning models in medical imaging. *J. Imaging* **2024**, *10*, 239. [CrossRef]
8. Jiang, Y.; Zhang, Y.; Lin, X.; Dong, J.; Cheng, T.; Liang, J. SwinBTS: A method for 3D multimodal brain tumor segmentation using Swin Transformer. *Brain Sci.* **2022**, *12*, 797. [CrossRef]
9. Isensee, F.; Jaeger, P.F.; Kohl, S.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef]
10. Chen, J.; Mei, J.; Li, X.; Lu, Y.; Yu, Q.; Wei, Q.; Luo, X.; Xie, Y.; Adeli, E.; Wang, Y.; et al. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Med. Image Anal.* **2024**, *97*, 103280. [CrossRef]
11. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.
12. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-UNet: U-Net-like pure transformer for medical image segmentation. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2022; pp. 205–218.
13. Hatamizadeh, A.; Tang, Y.; Nath, V.; Yang, D.; Myronenko, A.; Landman, B.; Roth, H.R.; Xu, D. UNetR: Transformers for 3d medical image segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 574–584.
14. Wang, W.; Chen, C.; Ding, M.; Yu, H.; Zha, S.; Li, J. TransBTS: Multimodal brain tumor segmentation using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer International Publishing: Cham, Switzerland, 2021; pp. 109–119.
15. Zhou, H.Y.; Guo, J.; Zhang, Y.; Yu, L.; Wang, L.; Yu, Y. nnFormer: Interleaved transformer for volumetric segmentation. *arXiv* **2021**, arXiv:2109.03201.
16. Zhang, Y.; Liu, H.; Hu, Q. Transfuse: Fusing transformers and CNNs for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer International Publishing: Cham, Switzerland, 2021; pp. 14–24.
17. Wei, C.; Ren, S.; Guo, K.; Hu, H.; Liang, J. High-resolution Swin transformer for automatic medical image segmentation. *Sensors* **2023**, *23*, 3420. [CrossRef]
18. Perera, S.R. Democratizing Artificial Intelligence Based Healthcare via Lightweight, Efficient and High-Performance Neural Networks. Doctoral Dissertation, The Ohio State University, Columbus, OH, USA, 2024.
19. Khan, S.; Mazhar, T.; Naz, N.S.; Ahmed, F.; Shahzad, T.; Ali, A.; Khan, M.A.; Hamam, H. Advanced feature selection techniques in medical imaging—A systematic literature review. *Comput. Mater. Contin.* **2025**, *85*, 2347–2401. [CrossRef]
20. 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.