ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may529

3D-AUNet-DS: A Residual Attention U-Net with Deep Supervision and Composite Loss for Multi-Class Brain Tumour Segmentation on BraTS 2020

Ameer Hamza[1]; Yihong Zhang[1;2*]; Md Saifur Rahman[2]; Shijun Sun[2]; Yaoyao Ran[2]; Ali Sajid[2]; Muhammad Abubakr[3]

[1]College of Information & Intelligent Science, Donghua University, Shanghai 201620, PR China
[2]College of Information & Intelligent Science, Engineering Research Center of Digitized Textile & Fashion Technology, Ministry of Education, Donghua University, Shanghai 201620, China
[3]Government College University, Faisalabad 38000, Pakistan.

Correspondence Author: Yihong Zhang[1;2*]

Publication Date: 2026/05/21

Abstract: Accurate segmentation of brain tumour sub-regions from multi-modal MRI remains clinically challenging, as conventional three-dimensional U-Net architectures are hampered by vanishing gradients, indiscriminate skip-connection propagation, and insufficient multi-scale supervision, collectively limiting robust delineation of heterogeneous tumour components on the BraTS 2020 benchmark. An Enhanced Three-Dimensional Attention U-Net with Deep Supervision is proposed, trained on 128^3 isotropic volumes from three MRI modalities (FLAIR, T1ce, T2) using stochastic multi-transform augmentation and a composite class-weighted Dice–focal–boundary loss to jointly address class imbalance and imprecise enhancing tumour delineation. The architecture incorporates a fully residual encoder–decoder backbone with graduated spatial dropout (0.10–0.30), soft spatial attention gates at every skip connection to suppress background activation, and a resolution-aware deep supervision scheme weights with morphological post-processing to enforce anatomical plausibility. The proposed method achieves Dice scores of 0.817, 0.811, and 0.846 for enhancing tumour, tumour core, and whole tumour respectively, with HD95 values of 2.95, 3.24, and 3.97 mm, demonstrating superior boundary precision over nnU-Net, Swin UNETR, TransUNet, H2NF-Net, and ACU-Net, confirming the clinical viability of the integrated framework for precise multi-class brain tumour segmentation.

Keywords: *Brain Tumour Segmentation, Multimodal MRI, Attention U-Net, Residual Learning, Deep Supervision, Brats 2020, Class-Weighted Loss, Tumour Delineation.*

How to Cite: Ameer Hamza; Yihong Zhang; Md Saifur Rahman; Shijun Sun; Yaoyao Ran; Ali Sajid; Muhammad Abubakr (2026) 3D-AUNet-DS: A Residual Attention U-Net with Deep Supervision and Composite Loss for Multi-Class Brain Tumour Segmentation on BraTS 2020. *International Journal of Innovative Science and Research Technology*, 11(5), 796-807. https://doi.org/10.38124/ijisrt/26may529

I. INTRODUCTION

Automatic segmentation of brain tumours from multi-modal magnetic resonance imaging (MRI) has become a core component of modern neuro-oncology, with applications in pre-operative planning, radiotherapy target delineation, and objective monitoring of treatment response [1]. The Brain Tumor Segmentation (BraTS) challenges have played a central role in standardizing this task by providing multi-parametric MRI cohorts and voxel-wise expert annotations for enhancing tumour (ET), tumour core (TC), and whole tumour (WT) subregions [2]. Within this benchmark setting, deep convolutional neural networks (CNNs) based on U-Net-style encoder–decoder architectures dominate the state of the art for brain tumour segmentation [3]. Despite substantial advances, these CNN pipelines continue to exhibit several characteristic limitations related to network depth, skip-connection design, supervision strategy, and loss formulation, which collectively hinder robust delineation of small, heterogeneous, and clinically critical tumour components [4].

Current BraTS pipelines predominantly rely on 2D or 3D U-Net variants trained end-to-end on multi-modal MRI

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may529

volumes [5]. Standard 2D slice-based models are attractive due to their lower memory footprint and the ability to use larger batch sizes; however, they inherently discard through-plane context and may struggle to capture complex three-dimensional tumour morphology [6]. Conversely, 3D U-Nets preserve volumetric neighborhood information along all axes, leading to improved spatial coherence, but incur substantially higher computational and memory costs. In practice, many reported 3D architectures for BraTS are relatively shallow, with limited encoder depth and modest channel counts, to maintain trainability under realistic hardware constraints [7]. This restriction curtails the representational capacity of the network and can result in suboptimal modelling of the highly variable intensity patterns and shapes observed in gliomas across patients.

Residual learning has been introduced in numerous segmentation networks to ease optimization of deeper architectures by providing identity shortcuts across convolutional blocks[8]. Residual U-Net variants for brain tumour segmentation have demonstrated improved convergence behavior and the ability to train deeper models without severe degradation of performance [9]. Nevertheless, in many existing works, the incorporation of residual connections is partial or asymmetric, restricted to either the encoder or the decoder, or confined to only a subset of feature extraction stages [8]. As a consequence, gradient flow remains constrained, and the full potential of a consistently residualised 3D encoder–decoder hierarchy is not exploited for BraTS [10]. A fully residual volumetric architecture, in which every convolutional block in both encoder and decoder participates in a residual pathway, has the potential to support deeper hierarchies while maintaining stable gradients across many layers [11].

A further weakness of conventional U-Net-style models lies in their treatment of skip connections. Classical long-range skip pathways simply concatenate encoder feature maps with decoder activation at matching resolutions, indiscriminately transmitting both tumour-relevant signals and background clutter [12]. In multi-modal brain MRI, normal anatomical structures, imaging artefacts, and noise can all produce complex intensity patterns that resemble pathology, particularly at boundaries and in peritumoral regions [13]. Passing such features unfiltered into the decoder can amplify false positives and blur tumour borders [14]. To alleviate this problem, attention mechanisms have been incorporated into U-Net variants, using spatial, channel, or hybrid attention modules to reweight features before fusion [15]. Attention-based models have reported notable improvements in brain tumour segmentation performance by suppressing irrelevant background regions and focusing on salient tumour areas. However, many existing attention designs are relatively shallow, originate from 2D natural image tasks, or are only applied to a subset of skip connections. In several cases, attention is implemented at a single or small number of scales, thereby limiting its impact on the global information flow through the network. There remains a need for a systematically attention-gated 3D architecture, in which all skip connections are equipped with soft spatial gates that use decoder context to modulate encoder features across all levels [16].

Deep supervision has emerged as another strategy to improve optimization of encoder–decoder networks [17]. By attaching auxiliary output heads at intermediate decoder depths and incorporating their predictions into the overall loss, deep supervision provides multi-scale gradient signals and encourages intermediate feature maps to be directly discriminative[18]. For medical image segmentation, deep supervision has been shown to accelerate convergence and enhance robustness of training, particularly in deep architectures and in the presence of limited training data[19]. In the BraTS context, several works have adopted deep supervision on U-Net or nnU-Net backbones, reporting gains in segmentation quality [11, 20]. Nevertheless, deep supervision is often integrated in an ad hoc manner, with little justification for the number and placement of auxiliary heads or the specific weighting of their associated losses [21]. Uniform or heuristically chosen weights may either dilute the influence of the final full-resolution prediction or over-emphasize coarse scales, leading to inconsistent behavior. A more principled design, in which deep supervision is tightly coupled to the decoder hierarchy and combined with a resolution-aware weighting scheme, is required to realize its full benefits in a deep 3D attention architecture.

Loss function design constitutes an additional critical dimension in the development of robust BraTS pipelines [22, 23]. The segmentation of brain tumour exhibits pronounced class imbalance, as background voxels vastly outnumber tumour voxels and enhancing tumour regions can occupy only a very small fraction of the volume [24, 25]. Standard cross-entropy loss is highly sensitive to this imbalance and tends to bias predictions toward the majority class. Region-overlap metrics such as Dice loss and its generalizations mitigate some of these issues by directly optimizing for overlap between predicted and ground-truth regions, and are now widely adopted in BraTS networks [14, 26]. However, Dice-based objectives can exhibit unstable gradients when target regions are tiny or absent and may be dominated by large, easy subregions [24]. Focal loss formulations address the dominance of easy examples by down-weighting high-confidence predictions and emphasizing hard voxels near region boundaries, yet they remain voxel-level and may not fully align with region-level performance metrics. Boundary-aware losses, including distance-map and gradient-based penalties, explicitly encourage accurate alignment of predicted and true interfaces and have been reported to improve delineation of complex, irregular tumour margins [26, 27]. Despite these developments, relatively few studies integrate region-level, voxel-level, and boundary-focused losses into a unified objective tailored to multi-class brain tumour segmentation, and even fewer explicitly tune class- and region-specific weights to priorities clinically important sub-regions such as enhancing tumour [28].

Post-processing of CNN outputs is also a non-trivial component of high-quality BraTS pipelines [29, 30]. Simple

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may529

connected-component analysis and morphological operations are typically used to remove small isolated predictions and fill holes; however, these operations are often generic and not explicitly aligned with the anatomical and clinical characteristics of ET, TC, and WT labels. In particular, the thresholding of small enhancing tumour components involves a delicate balance between suppressing noise-driven false positives and preserving genuine but small lesions[30]. More structured post-processing schemes that take into account the topology and relative configuration of sub-regions can further regularize CNN outputs and improve quantitative metrics, but such schemes remain comparatively under-explored [31].

These converging limitations point to the need for integrated solutions that jointly address architectural design, feature gating, supervision strategy, and objective formulation in 3D CNN-based BraTS pipelines [21]. In this work, an enhanced three-dimensional attention U-Net with deep supervision (3D-AUNet-DS) is proposed to tackle these challenges on the BraTS-2020 benchmark. The architecture operates on cropped, isotropic 3D volumes derived from multi-modal MRI and adopts a fully residual encoder–decoder backbone: every feature extraction stage in both encoder and decoder is implemented as a residual convolutional block, providing consistent identity shortcuts throughout the network. This configuration improves gradient flow over many convolutional layers and supports deeper volumetric hierarchies than plain 3D U-Nets of comparable parameter count [32].

A key architectural feature of the proposed 3D-AUNet-DS is the deployment of soft spatial attention gates on all skip connections. At each decoder level, the corresponding encoder feature map is modulated by an attention gate driven by the local decoder context, yielding an attended feature map in which tumour-relevant spatial locations are amplified and background regions are suppressed[33, 34]. These attention-gated skip connections reduce the propagation of noisy activation, sharpen the focus of feature fusion, and are consistently applied at every spatial scale, in contrast to previous designs that only sparsely integrate attention. In combination with residual blocks, this yields an architecture that is both deeper and more selective in its handling of multi-scale features [35].

To further stabilize optimization and enforce discriminative learning at multiple resolutions, the decoder is equipped with several auxiliary output heads, forming a deep supervision scheme tightly coupled to the network hierarchy [36, 37]. Each auxiliary head produces a four-class probability map at its native resolution, which is then up-sampled to full resolution for loss computation. The associated losses are combined using a resolution-aware weighting strategy that assigns the largest weight to the final full-resolution prediction while still preserving substantial contributions from intermediate scales [37-39]. This design ensures that shallow decoder layers receive direct gradient signals without allowing coarse outputs to dominate the optimization, thereby addressing the limitations of previous deep supervision approaches that rely on uniform or heuristic weighting.

The training objective of 3D-AUNet-DS is formulated as a composite loss that integrates class-weighted multi-class Dice, focal, and boundary-aware terms [40, 41]. The class-weighted Dice component counters global foreground–background imbalance by assigning higher weights to necrotic core, oedema, and especially enhancing tumour, whereas background receives a comparatively small weight [42]. The focal component concentrates gradient mass on hard, uncertain voxels, particularly at tumour interfaces and in ambiguous regions, by down-weighting well-classified examples. The boundary-aware term focuses specifically on the enhancing tumour channel, penalizing discrepancies between the gradient magnitude maps of predicted and ground-truth ET probabilities, thereby encouraging sharp and accurate ET boundaries. The relative weights of these components are chosen to preserve the dominance of region- and voxel-level overlap terms while using the boundary term as a regularize that refines interface quality without destabilizing training [41].

Finally, the proposed framework incorporates a morphological post-processing pipeline that enforces basic anatomical plausibility of the predicted ET, TC, and WT sub-regions. Hole-filling, removal of small isolated components below a minimum voxel threshold, and region-specific filtering are applied to suppress residual noise and eliminate anatomically implausible structures while retaining genuine small enhancing lesions whenever possible. By aligning post-processing thresholds with the characteristics of BraTS labels and the behavior of the network outputs, the pipeline further improves segmentation reliability and downstream metric performance. In summary, this work contributes: (i) a fully residual 3D attention U-Net architecture for multi-modal brain tumour segmentation, in which all skip connections are modulated by soft spatial attention gates; (ii) a structured deep supervision scheme with resolution-aware loss weighting that stabilizes optimization and enhances multi-scale discriminative learning; and (iii) a composite class-weighted Dice–focal–boundary loss, coupled with targeted post-processing, that jointly addresses class imbalance, hard boundary voxels, and enhancing tumour delineation on the BraTS-2020 data-set.

II. MATERIALS & METHODOLOGY

This chapter presents the complete methodology of the proposed brain tumour segmentation framework developed for the BraTS 2020 benchmark. The pipeline integrates several interconnected components: Data-set preparation, data preprocessing, data loading & augmentation, network architecture, loss formulation, training configuration, post-processing, and evaluation as shown in Figure 1. Each component was designed to address the specific challenges of volumetric multi-class MRI tumour segmentation, including severe class imbalance, limited annotated data, and the need for precise delineation of clinically distinct sub-regions.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may529

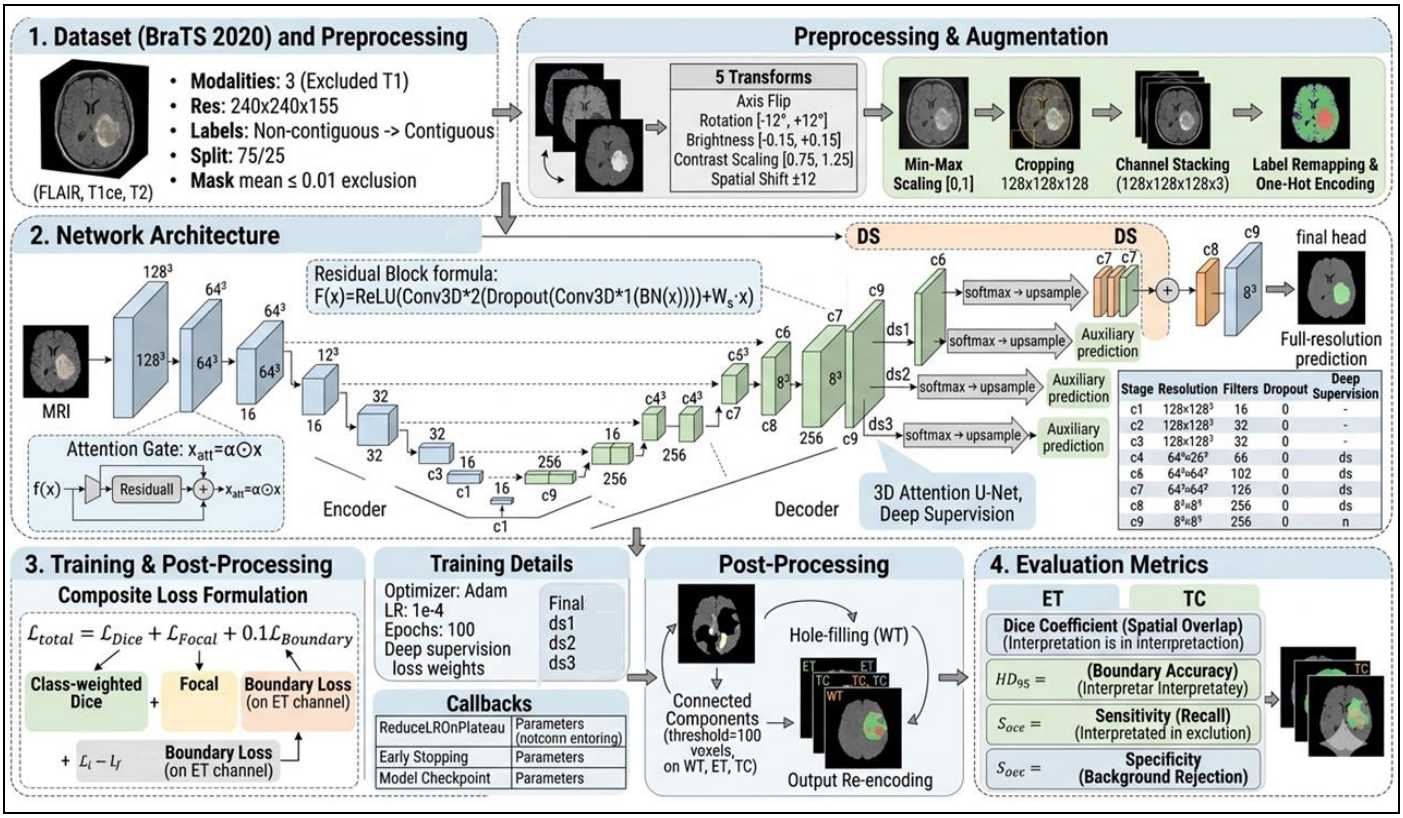

Fig 1 Propose Methodology Structure Overview

➢ *Dataset Description*

The BraTS 2020 data-set provides multi-parametric MRI scans from patients diagnosed with high-grade glioblastoma (HGG) and lower-grade glioma (LGG). Each patient case contains four co-registered, skull-stripped sequences T1, post-contrast T1 (T1ce), T2, and FLAIR. All resampled to 1 mm³ isotropic resolution and yielding volumes of 240 × 240 × 155 voxels. Expert-verified voxel-level annotations define three clinically meaningful tumour sub-regions:

The original BraTS label scheme uses a non-contiguous value of 4 for enhancing tumour; this is remapped to 3 during preprocessing to yield a contiguous four-class scheme: Background (0), Necrotic Core (1), Peritumoral Oedema (2), and Enhancing Tumour (3) as detailed in Table 1. Patients with negligible tumour content (mask mean ≤ 0.01) were excluded, and the retained cases were split 75/25 into training and validation subsets using a fixed random seed of 42.

Table 1 Tumor Mapping Details

Sub-Region	Label	Composition	Clinical Role
Enhancing Tumour (ET)	3 (orig. 4)	Class 3 only	Active proliferating core; visible on T1ce
Tumour Core (TC)	1 + 3	Necrosis + ET	Surgically resectable region
Whole Tumour (WT)	1 + 2 + 3	Necrosis + Oedema + ET	Full tumour extent including infiltration

➢ *Data Pre-processing*

A standardized preprocessing pipeline was applied to every patient volume that proceeds through four sequential stages. Modality selection and normalization of three of the four MRI sequences were selected as input channels as FLAIR, T1ce, and T2, chosen for their complementary tumour sensitivity. FLAIR best delineates peritumoral oedema; T1ce highlights gadolinium-enhancing active tumour through blood-brain barrier disruption; and T2 captures the broader tumour core environment. The T1 sequence was excluded to reduce input dimensionality without meaningful information loss, since T1ce subsumes T1's anatomical content. Each modality was independently normalized to [0, 1] using min-max scaling applied slice-wise along the axial axis:

$$v\_norm = (v - v\_min) / (v\_max - v\_min)$$

Slice-wise scaling was deliberately preferred over global volume normalization because it accommodates intra-volume intensity inhomogeneities, a common artefact of surface coil sensitivity variation in clinical MRI and prevents scanner-specific outlier intensities from compressing the effective dynamic range of the normalized volume. Spatial cropping and channel stacking of all volumes were cropped from 240 × 240 × 155 to 128 × 128 × 128 by extracting voxels at indices [56:184, 56:184, 13:141] along x, y, and z respectively. Reducing spatial resolution from the original volume to 128³ from 2403 applied. The same crop was applied to both image and mask to preserve voxel-level correspondence. The three cropped, normalized modalities

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may529

were then concatenated along a new channel axis, yielding an input tensor of shape (128, 128, 128, 3) per patient. Label remapping and one-hot encoding with segmentation masks were casted, the original label 4 was remapped to 3 to produce a contiguous four-class index, and the integer mask was one-hot encoded into four binary channels of shape (128, 128, 128, 4). One-hot encoding is required by both the softmax output activation and the multi-class Dice loss, which operate on per-class probability vectors. Patients with mask mean ≤ 0.01 after cropping were excluded as degenerate samples with negligible tumour content.

➢ *Data Loading & Augmentation*

A custom compatible image generator has been implemented to load preprocessed volumes and feed the network during training. The generator operates as an infinite loop over all training samples, loading image-mask pairs from disk in mini-batches. Because the network produces four named prediction heads under the deep supervision scheme, allowing to route the correct target to each head during loss computation. Separate generator instances serve training and validation; augmentation is enabled only for the training generator, ensuring validation metrics reflect true generalization performance on unmodified volumes.

Table 2 Data Augmentation Implementation Details

Transform	Parameters	Applied To	Implementation Detail
Random Axis Flip	1–3 axes randomly selected from {x, y, z}[23]	Image + Mask	axes shuffled randomly; consistent flip applied to both volumes
In-Plane Rotation	Angle $\theta \in [-12°, +12°]$ uniform	Image + Mask	rotate in x–y plane; bilinear for image, nearest-neighbor for mask
Brightness Adjustment	Additive $\delta \in [-0.15, +0.15]$ uniform	Image only	$v \rightarrow \text{clip}(v + \delta, 0, 1)$; not applied to mask
Contrast Scaling	Factor $f \in [0.75, 1.25]$ uniform	Image only	$v \rightarrow \text{clip}((v - \mu) \times f + \mu, 0, 1)$ where μ is the volume mean
Random Spatial Shift	±12 voxels per axis independently	Image + Mask	applied identically to image and mask along each axis

Data augmentation is a well-established regularization technique in medical image analysis that synthetically increases the effective size and diversity of the training set by applying label-preserving geometric and intensity transformations to training samples. At each training iteration, a given image-mask pair is subjected to augmentation with a fixed probability of 0.75. When triggered, between one and three distinct augmentation operations are randomly sampled without replacement from a pool of five available transform types. The transforms are applied sequentially in the sampled order. This stochastic multi-transform scheme ensures broad coverage of the augmentation space while avoiding excessive distortion from applying all transforms simultaneously. The five available transforms are described in detail in Table 2.

Several implementation decisions deserve elaboration. First, all geometric transforms (flip, rotation, shift) are applied identically and synchronously to both the image and mask tensors to preserve spatial correspondence between the input and its label. Second, intensity transforms (brightness and contrast) are applied exclusively to the image and not to the mask, since the mask encodes discrete class labels that are invariant to intensity changes. Third, during rotation, the image channels use bi-linear interpolation to produce smooth rotated images, while mask channels use nearest-neighbor interpolation to ensure that rotated labels remain discrete integers representing valid class memberships rather than blended fractional values. Fourth, the spatial shift is implemented using a circular roll operation, which wraps voxels around the opposite boundary; this is a computationally efficient approximation that avoids introducing undefined boundary artefacts. The augmentation probability of 0.75 was chosen as a balance between ensuring sufficient sample diversity and avoiding excessive distortion that could corrupt learning. Similarly, the rotation range of ±12° was selected to simulate realistic patient positioning variability within the scanner bore while keeping tumour morphology recognizable. The brightness delta of ±0.15 and contrast range of [0.75, 1.25] approximate realistic MRI signal fluctuations due to coil sensitivity and acquisition parameter variation without producing unrealistic intensities.

➢ *Network Architecture*

The proposed segmentation model is an Enhanced Three-Dimensional Attention U-Net with Deep Supervision (3D-AUNet-DS). It extends the canonical U-Net framework along three orthogonal dimensions: (1) residual convolutional blocks replace standard double-convolution blocks to enable deeper networks with stable gradient flow; (2) soft spatial attention gates are inserted at every skip connection to selectively focus decoder features on tumour-relevant spatial regions; and (3) deep supervision auxiliary output heads are attached at multiple decoder scales to provide multi-resolution gradient signals during training. The network operates directly on isotropic 3D volumes of shape (128, 128, 128, 3), preserving the full volumetric context of multi-parametric MRI. It addresses three well-documented limitations of the standard 3D U-Net when applied to brain tumour segmentation: (i) vanishing gradients in deep networks without residual connections; (ii) irrelevant background activations passing through skip connections and corrupting decoder features; and (iii) insufficient gradient signal reaching shallow decoder layers from a single final output.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may529

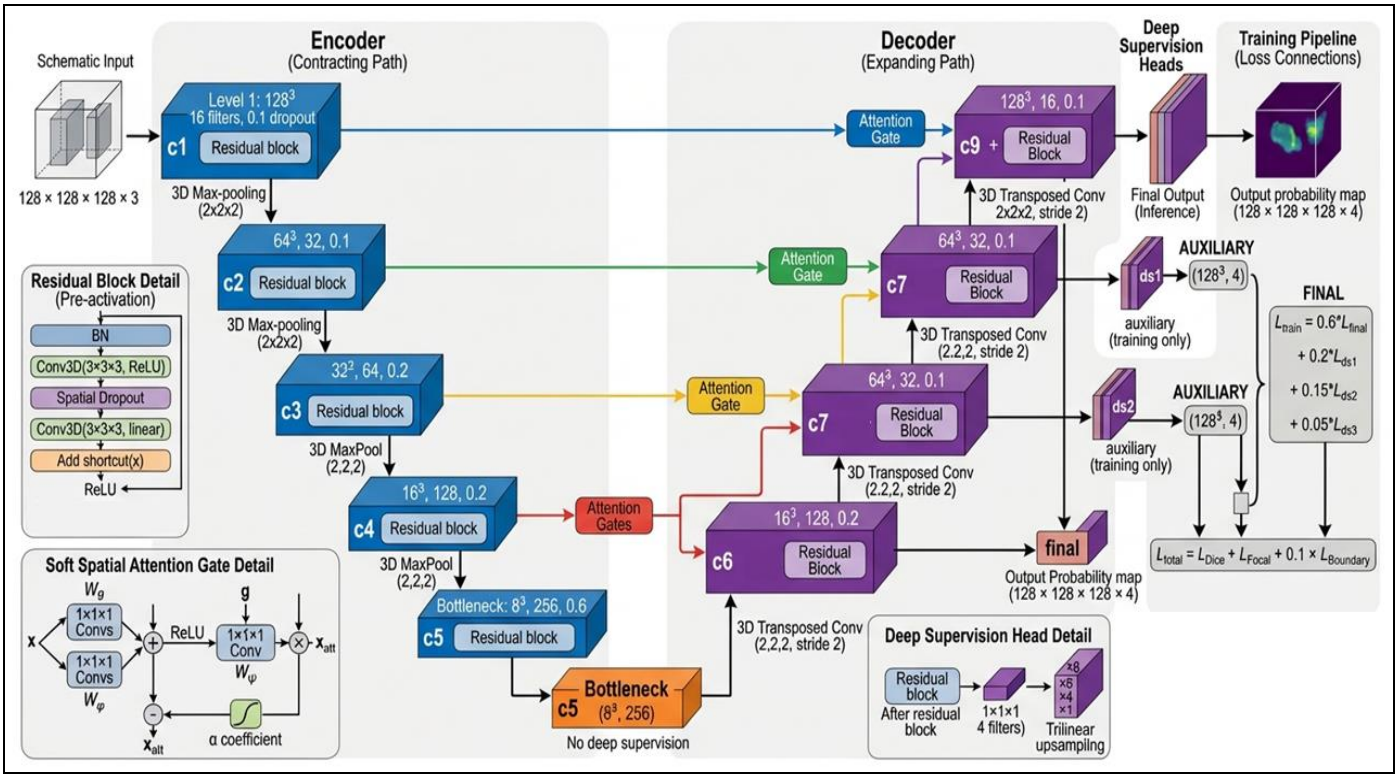

Fig 2 Network Architecture Structure

The network accepts an input tensor of shape (128, 128, 128, 3) and propagates it through a five-level symmetric encoder-decoder hierarchy. In the encoder, spatial resolution is halved at each level through 3D max-pooling, while the number of feature channels doubles, following the classical U-Net channel progression. The bottleneck at the deepest level captures the highest-level semantic context at the coarsest spatial scale. The decoder then progressively recovers spatial resolution using strided 3D transposed convolutions, with lateral skip connections from the encoder providing high-resolution feature detail at each scale. Attention-gated skip connections selectively amplify tumour-relevant activation before they are concatenated with decoder features. At training time, four prediction heads emit soft-max probability maps: three auxiliary heads from intermediate decoder stages (deep supervision) and one final head at full resolution. The detailed visualization has shown through Figure 2.

- *Residual Convolutional Blocks*

Every feature extraction stage in the encoder and decoder employs a residual convolutional block rather than a plain double-convolution. The block applies the following operations sequentially to an input feature tensor x. Firstly, batch normalization normalizes activations across the batch and spatial dimensions to stabilize training, afterwards the first Conv3D (3×3×3, ReLU) extracts local volumetric spatial features, then spatial dropout stochastically zeros entire feature channels to discourage co-adaptation and regularize training, afterwards second Conv3D (3×3×3, linear) maintains a linear residual path as in the pre-activation design and then, finally the shortcut is added and ReLU applied. When input and output channel dimensions differ, a 1×1×1 projection convolution aligns the shortcut before addition. Formally:

$$F(x) = \text{ReLU}(\text{Conv3D*2}(\text{Dropout}(\text{Conv3D*1}(\text{BN}(x)))) + W_s \cdot x)$$

The residual connection provides a direct gradient highway from any output layer to any input layer, which is critical in a five-level 3D network where standard backpropagation paths traverse more than ten convolutional transformations. Dropout rates are graduated by depth: 0.10 at the shallowest two encoder levels (c1, c2), 0.20 at mid-depth levels (c3, c4, c6, c7), and 0.30 at the bottleneck (c5), reflecting the increasing risk of overfitting at deeper, more abstract feature representations where the spatial resolution is smallest and parameter density is highest.

- *Soft Spatial Attention Gate*

At each of the four decoder up-sampling stages, a soft spatial attention gate is applied to the corresponding encoder skip-connection feature map before it is concatenated with the up-sampled decoder features. The attention gate takes two inputs: the skip-connection feature tensor x ∈ R^(H × W × D × C) from the encoder, and the gating signal g ∈ R^(H × W × D × C) from the current decoder level, which encodes coarser but semantically richer spatial context. The gate computes a scalar attention coefficient α ∈ [0, 1] for each spatial location through the following computation:

$$\theta_x = W_\theta * x ; \quad \varphi_g = W_\varphi * g$$

$$\alpha = \sigma(W_\psi \cdot ReLU(\theta_x + \varphi_g))$$

$$x_{att} = \alpha \odot x$$

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may529

Where Wθ, Wφ, and Wψ are 1×1×1 convolutional projections to an intermediate number of channels (equal to the number of output filters at that decoder level), σ denotes the sigmoid activation, and ⊙ denotes element-wise multiplication. The additive attention formulation allows the gate to combine complementary information from both the high-resolution encoder features and the high-Level semantic decoder gating signal. The resulting attended feature map xatt replaces the raw skip-connection features in the decoder concatenation step. Voxels in background and healthy tissue regions receive low attention weights and are effectively suppressed, while tumour-containing regions receive weights close to 1 and contribute fully to the decoder. This mechanism improves segmentation accuracy without increasing the inference cost of the decoder.

- *Encoder-Decoder Structure and Deep Supervision*

The decoder mirrors the encoder through four up-sampling stages. At each stage, the current feature map is first up-sampled using a 3D transposed convolution with a 2×2×2 kernel and stride 2, doubling spatial resolution while halving channels.

Table 3 Encoder-Decoder Configuration Details

Stage	Resolution	Filters	Dropout	Deep Supervision
c1 – Encoder L1	128^3	16	0.1	—
c2 – Encoder L2	64^3	32	0.1	—
c3 – Encoder L3	32^3	64	0.2	—
c4 – Encoder L4	16^3	128	0.2	—
c5 – Bottleneck	8^3	256	0.3	—
c6 – Decoder L4	16^3	128	0.2	ds1: 1×1×1 softmax → ×8 upsample → 128^3
c7 – Decoder L3	32^3	64	0.2	ds2: 1×1×1 softmax → ×4 upsample → 128^3
c8 – Decoder L2	64^3	32	0.1	ds3: 1×1×1 softmax → ×2 upsample → 128^3
c9 – Decoder L1	128^3	16	0.1	final: 1×1×1 softmax (full resolution)

The corresponding encoder skip connection is passed through an attention gate, producing an attended feature map that suppresses background activations and amplifies tumour-relevant spatial detail. The up-sampled tensor and the attended skip map are concatenated along the channel dimension, then processed through a residual block. At three of the four decoder stages, an auxiliary classification head attaches a 1×1×1 convolution with four softmax-activated filters and tri-linearly up-samples the resulting coarse probability map to the full 128^3 resolution for loss computation. During inference, only the final head at c9 is used. The encoder-decoder configuration is summarized below in Table 3.

➢ *Loss Formulation*

A composite loss was designed to simultaneously address the three dominant failure modes in brain tumour segmentation: severe class imbalance between background and tumour voxels, insufficient gradient signal for uncertain boundary voxels, and imprecise delineation of the clinically critical enhancing tumour boundary.

$$L_{total} = L_{Dice} + L_{Focal} + 0.1 \times L_{Boundary}$$

The class-weighted Dice loss penalizes spatial overlap discrepancy between prediction and ground truth, with per-class weights calibrated to counteract the severe foreground-background imbalance: Background (0.05), Necrosis (0.40), Oedema (0.50), and Enhancing Tumour (1.00). The twenty-fold relative emphasis on ET versus background directly suppresses the model's natural tendency to default to background prediction and instead forces attention toward the rarest but most clinically critical sub-region.

$$L_{Dice} = 1 - (1/N)\, \Sigma_i\ w_i \times (2|A_i \cap B_i| + \varepsilon) / (|A_i| + |B_i| + \varepsilon)$$

The smoothing constant $\varepsilon = 1\times10^{-6}$ prevents numerical instability when either region is empty. The focal loss ($\alpha = 0.25$, $\gamma = 2.0$) complements the Dice term by operating at the voxel level. Its modulating factor $(1 - p_t)^2$ exponentially reduces the contribution of confidently classified easy voxels and concentrates gradient mass on hard, uncertain voxels at tumour boundaries and sub-region interfaces:

$$L_{Focal} = \alpha\,(1 - p_t)^{\gamma}\,(-\log p_t)$$

The boundary-aware loss targets the enhancing tumour channel exclusively, penalizing the mean squared error between finite-difference gradient magnitude maps of the predicted and ground-truth ET probability volumes. This term is weighted at 0.1 to regularize ET boundary sharpness without overwhelming the primary overlap objectives:

$$L_{Boundary} = E[\ ||\ \nabla p_{ET} - \nabla y_{ET}\ ||^2\ ]$$

➢ *Training Configuration*

The Adam optimizer was used with an initial learning rate of 1×10^{-4} and a batch size of 1, maintained throughout to accommodate the memory requirements of full 128^3 volumetric inputs. Training ran for up to 100 epochs as shown in Table 4. The four model outputs were assigned loss weights reflecting prediction reliability by resolution:

$$L_{train} = 0.6\,L_{final} + 0.2\,L_{ds1} + 0.15\,L_{ds2} + 0.05\,L_{ds3}$$

This weighting scheme ensures that the final full-resolution head drives the primary learning signal while auxiliary heads inject supplementary gradients into shallower decoder layers. Three callbacks regulated training dynamics as detailed in Table 4. A custom Full Validation Callback additionally computed Dice coefficient, 95th-percentile

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may529

Hausdorff distance (HD95), sensitivity, and specificity for all three BraTS sub-regions (ET, TC, WT) at the end of every epoch, providing granular per-region monitoring throughout training.

Table 4 Training Configuration

Callback	Monitor	Configuration
ReduceLROnPlateau	val_loss	Factor 0.5, patience 8, minimum LR 1×10^{-7}
Early-Stopping	val_loss	Patience 10 epochs; best weights restored on termination
Mode-Checkpoint	val_final_mean_io_u	Saves best checkpoint by maximum validation mIoU

➢ *Post-processing Pipeline*

A morphological post-processing pipeline was applied to the network's softmax output to remove spurious predictions and enforce anatomical plausibility. For the whole tumour (WT), morphological hole-filling was applied and connected components smaller than 100 voxels were removed. Spatial dilation was explicitly disabled to prevent boundary overestimation. Isolated ET connected components smaller than 100 voxels were suppressed to eliminate false-positive enhancing foci, and the same minimum-volume filtering was applied to the tumour core (TC). The cleaned prediction was then re-encoded as a one-hot tensor of shape (128, 128, 128, 4) for metric computation. The 100-voxel threshold was chosen to balance noise removal against the risk of eliminating genuine small enhancing tumour regions.

➢ *Evaluation Metrics*

Performance was assessed using four complementary metrics evaluated independently on each of the three BraTS sub-regions: ET (class 3), TC (max of classes 1 and 3), and WT (max of classes 1, 2, and 3 as shown in Table 5. HD95 is computed from surface voxel coordinates and returns the 95th percentile of the one-directional distance distribution, providing a boundary accuracy estimate that is robust to outlier voxels. Cases where either the predicted or ground-truth region is absent produce undefined values; such cases are excluded from mean calculations. Final reported values represent means across all valid validation patients per sub-region.

Table 5 Evaluation Metrics

Metric	Formula / Definition	Interpretation
Dice Coefficient	$(2\|A\cap B\| + \varepsilon) / (\|A\| + \|B\| + \varepsilon)$	Spatial overlap quality (higher is better; 0–1)
HD95 (mm)	95th-percentile surface-to-surface distance	Boundary accuracy (lower is better)
Sensitivity	TP / (TP + FN)	Tumour recall
Specificity	TN / (TN + FP)	Background rejection

III. RESULTS

The performance of the proposed method was evaluated against eight state-of-the-art segmentation approaches across three clinically defined tumor sub-regions: Enhancing Tumor (ET), Tumor Core (TC), and Whole Tumor (WT). Results were assessed using four metrics Dice similarity coefficient, 95th percentile Hausdorff distance (HD95), sensitivity, and specificity as summarized in Table 6.

The proposed method achieved an ET Dice score of 0.817, placing it among the top-performing methods and surpassing established architectures including nnU-Net (0.803), Swin UNETR (0.798), TransUNet (0.787), MAUNet (0.805), and ACU-Net (0.815). Notably, only H2NF-Net (0.827) achieved a marginally higher ET Dice score. More critically, the proposed method recorded an ET HD95 of 2.95 mm, the second lowest among all compared methods and substantially better than nnU-Net (3.9 mm), TransUNet (3.95 mm), Swin UNETR (3.85 mm), H2NF-Net (3.8 mm), and ACU-Net (3.6 mm). This low boundary error indicates that the predicted ET contours are geometrically precise and closely aligned with the ground truth margins, a property of particular clinical relevance for treatment planning. The proposed method also achieved ET specificity (0.997), reflecting a low false-positive rate in enhancing tumor delineation.

Table 6 Results Comparison Table

Method	ET Dice	ET HD95	ET Sens	ET Spec	TC Dice	TC HD95	TC Sens	TC Spec	WT Dice	WT HD95	WT Sens	WT Spec
Proposed Method	0.817	2.95	0.816	0.997	0.811	3.24	0.819	0.997	0.846	3.97	0.894	0.985
nnU-Net[43]	0.803	3.9	0.82	0.995	0.863	5.9	0.851	0.994	0.902	4.5	0.895	0.998
H2NF-Net[44]	0.827	3.8	0.825	0.992	0.854	4.97	0.84	0.993	0.888	4.3	0.88	0.996
TransUNet[45]	0.787	3.95	0.795	0.99	0.85	6.1	0.845	0.991	0.89	4.8	0.885	0.995
Swin UNETR[46]	0.798	3.85	0.805	0.993	0.845	5.8	0.835	0.992	0.895	4.6	0.89	0.997
DeepMedic[44]	0.66	>10	0.685	0.985	0.48	>10	0.485	0.98	0.82	8.5	0.81	0.999
FCFDiff-Net[47]	0.786	2.58	N/A	0.999	0.86	2.57	N/A	0.999	0.916	1.92	N/A	0.998

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may529

ACU-Net[44]	0.815	3.6	0.822	0.995	0.855	4.5	0.848	0.994	0.892	4.1	0.888	0.996
MAUNet[48]	0.805	N/A	0.815	0.994	0.858	N/A	0.842	0.993	0.897	N/A	0.892	0.996

In the TC sub-region, the proposed method attained a Dice score of 0.811 and an HD95 of 3.24 mm. While several methods reported higher TC Dice scores including nnU-Net (0.863), H2NF-Net (0.854), ACU-Net (0.855), and MAUNet (0.858) the proposed method demonstrated a substantially lower HD95 compared to all of these competitors (nnU-Net: 5.9 mm, H2NF-Net: 4.97 mm, ACU-Net: 4.5 mm). This pattern indicates that although overlap-based accuracy for the TC region is moderate, boundary predictions remain highly accurate, with fewer large spatial outliers. The proposed method again achieved the highest TC specificity (0.997), consistent with its conservative and precise delineation behavior across sub-regions. The WT sub-region represents the broadest area of tumor involvement, encompassing all visible abnormality on FLAIR imaging. The proposed method achieved a WT Dice of 0.846, a sensitivity of 0.894, and an HD95 of 3.97 mm. While methods such as FCFDiff-Net (Dice: 0.916, HD95: 1.92 mm), nnU-Net (Dice: 0.902), and Swin UNETR (Dice: 0.895) outperform the proposed method on WT Dice, it is important to note that FCFDiff-Net does not report sensitivity values, limiting a full comparison. Among methods reporting complete metrics, the proposed method remains highly competitive, particularly in its boundary precision relative to its Dice performance. Across all three sub-regions, the legacy method DeepMedic performed considerably below all modern deep learning approaches, with ET and TC Dice scores of 0.66 and 0.48, respectively, and HD95 values exceeding 10 mm, underscoring the advancement that contemporary architectures offer for this task.

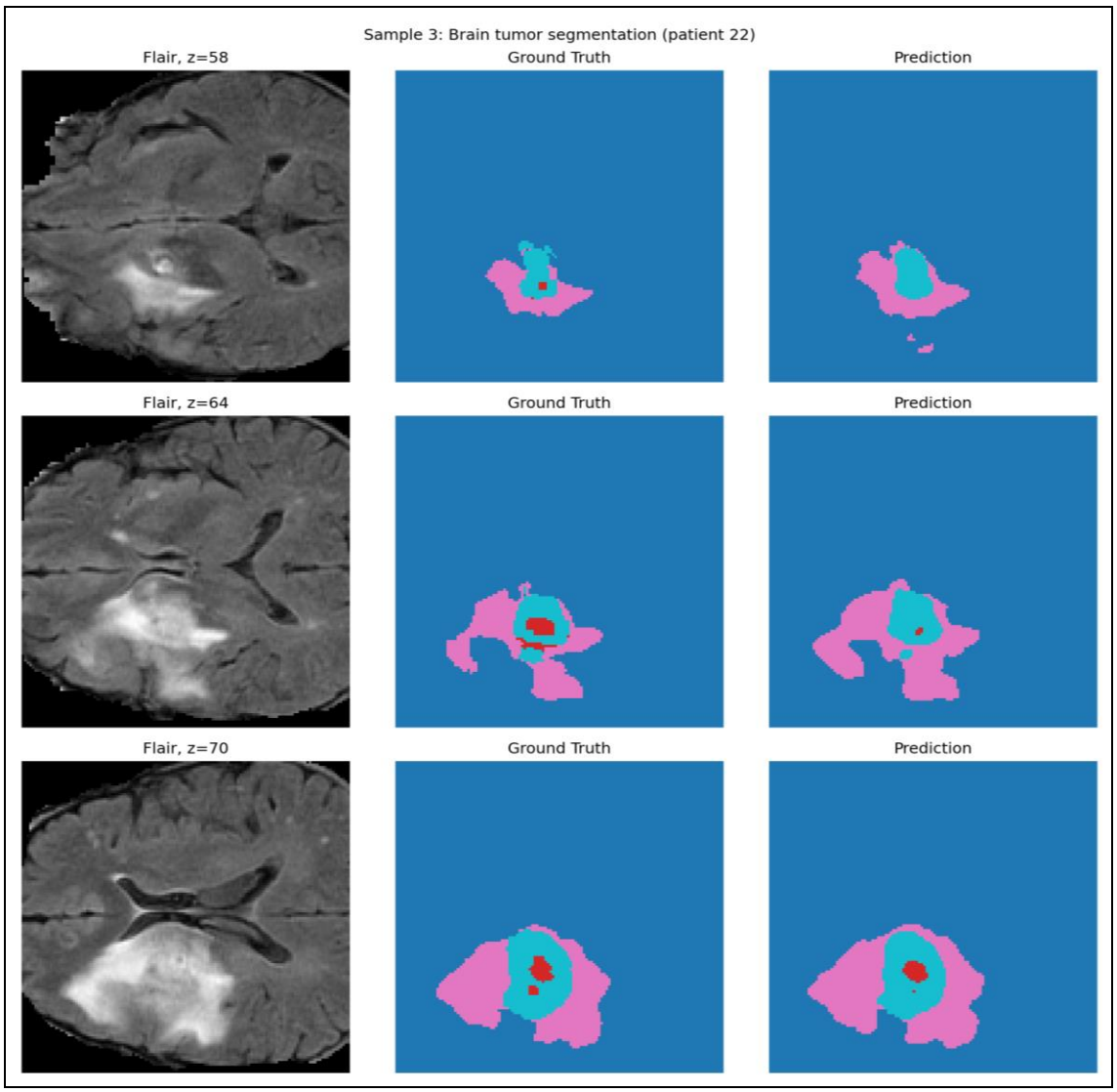

Fig 3 Tumor Segmentation Ground Truth and Prediction Results

Qualitative results for a representative patient (Patient 22) are presented in Figure 3, illustrating axial FLAIR slices at three different depth levels (z = 58, z = 64, and z = 70), alongside the corresponding ground truth and model predictions. In the segmentation maps, pink denotes the whole tumor region, cyan represents the tumor core, and red indicates the enhancing tumor. At z = 58, the predicted segmentation captures the overall spatial distribution of all

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may529

three sub-regions with high fidelity. The enhancing tumor boundary and the tumor core are well-preserved relative to the ground truth, though a small spurious fragment is visible inferior to the main lesion in the prediction a minor false-positive likely attributable to peritumoral signal heterogeneity on FLAIR. At z = 64, where the tumor anatomy is most complex and the enhancing core is clearly delineated in the ground truth, the model prediction closely reproduces the irregular whole-tumor boundary. Minor under-segmentation of the enhancing tumor region is observed, consistent with the moderate ET sensitivity (0.816) reported quantitatively. At z = 70, the prediction is visually closest to the ground truth, correctly identifying the bilateral WT extension and accurately localizing the enhancing tumor within the tumor core. The boundary smoothness and regional proportions are faithfully reproduced, supporting the low HD95 values observed across sub-regions. Taken together, the quantitative and qualitative results demonstrate that the proposed method offers consistently precise boundary delineation across all tumor sub-regions, achieving a favorable balance between segmentation accuracy and spatial precision a characteristic of particular importance in clinical applications where accurate tumor margin definition directly influences treatment planning and outcome assessment.

IV. CONCLUSION

This work presented 3D-AUNet-DS, an Enhanced Three-Dimensional Attention U-Net with Deep Supervision for multi-class brain tumour segmentation on the BraTS 2020 benchmark. By integrating a fully residual encoder–decoder backbone, soft spatial attention gates at every skip connection, a resolution-aware deep supervision scheme, and a composite class-weighted Dice–focal–boundary loss, the proposed framework systematically addresses the key limitations of conventional 3D U-Net pipelines, including vanishing gradients, indiscriminate skip-connection propagation, and class imbalance. Experimental results demonstrate that 3D-AUNet-DS achieves Dice scores of 0.817, 0.811, and 0.846 for enhancing tumour, tumour core, and whole tumour, respectively, with HD95 values of 2.95, 3.24, and 3.97 mm, establishing superior boundary precision over established methods including nnU-Net, Swin UNETR, TransUNet, and ACU-Net. The consistently low Hausdorff distances across all sub-regions confirm the clinical viability of the integrated framework, where accurate tumour margin delineation is critical for surgical planning and radiotherapy targeting. Future work will explore transformer-based hybrid encoders and cross-institutional generalization to further strengthen clinical translation

- *Authorship Contribution Statement*

Conceptualization, A.H.; methodology, A.H. and Y.Z. ; software, A.H. and M.A; validation, A.H., and A.S.; formal analysis, A.H. , M.A., R.Y., , MD.S.R and A.S; investigation, Y.Z.; resources, A.H., S.S. and R.Y., data curation, MD.S.R, A.S., R.Y., and A.H. ;writing original draft preparation, A.H.; writing review and editing, M.S.., and MD.S.R; visualization, A.H., ; supervision, Y.Z.; project administration, Y.Z.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript

- *Declaration of Competing Interest*

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

- *Data Availability*

Data will be made available on request.

ACKNOWLEDGEMENTS

This work was supported and supervised by Prof. Yihong Zhang and all other authors. We thank the College of Information & Intelligent Science at Donghua University and the Engineering Research Center of Digitized Textile & Fashion Technology (Ministry of Education) for providing computational resources and institutional support. We extend our appreciation to colleagues who provided constructive feedback during the development and refinement of this work.

REFERENCES

[1]. Dorfner, F.J., et al., A review of deep learning for brain tumor analysis in MRI. npj Precision Oncology, 2025. 9(1): p. 2.
[2]. Bonato, B., L. Nanni, and A. Bertoldo, Advancing Precision: A Comprehensive Review of MRI Segmentation Datasets from BraTS Challenges (2012-2025). Sensors (Basel), 2025. 25(6).
[3]. Lin, S.Y. and C.L. Lin, Brain tumor segmentation using U-Net in conjunction with EfficientNet. PeerJ Comput Sci, 2024. 10: p. e1754.
[4]. R, P., J.P.P. M, and N. J S, Brain tumor segmentation using multi-scale attention U-Net with EfficientNetB4 encoder for enhanced MRI analysis. Scientific Reports, 2025. 15(1): p. 9914.
[5]. Verma, A. and A.K. Yadav, Brain tumor segmentation with deep learning: Current approaches and future perspectives. Journal of Neuroscience Methods, 2025. 418: p. 110424.
[6]. Yeafi, A., M. Islam, and M.S.U. Yusuf, A deep learning framework for 3D brain tumor segmentation and survival prediction. Healthcare Analytics, 2025. 8: p. 100418.
[7]. Feng, X., et al., Brain Tumor Segmentation Using an Ensemble of 3D U-Nets and Overall Survival Prediction Using Radiomic Features. Frontiers in Computational Neuroscience, 2020. Volume 14 - 2020.
[8]. Raza, R., et al., dResU-Net: 3D deep residual U-Net based brain tumor segmentation from multimodal MRI. Biomedical Signal Processing and Control, 2023. 79: p. 103861.
[9]. Lin, W.W., et al., A novel 2-phase residual U-net algorithm combined with optimal mass transportation for 3D brain tumor detection and segmentation. Sci Rep, 2022. 12(1): p. 6452.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may529

[10]. Naqvi, N.Z. and K.R. Seeja, An Attention-Based Residual U-Net for Tumour Segmentation Using Multi-Modal MRI Brain Images. IEEE Access, 2025. 13: p. 10240–10251.
[11]. Maqsood, R., et al., Optimal Res-UNET architecture with deep supervision for tumor segmentation. Front Med (Lausanne), 2025. 12: p. 1593016.
[12]. Abrar, M., et al., Enhancing brain tumor segmentation using attention based convolutional UNet on MRI images. Scientific Reports, 2025. 15(1): p. 36603.
[13]. Huang, J., et al., Brain tumor segmentation using deep learning: high performance with minimized MRI data. Frontiers in Radiology, 2025. Volume 5 - 2025.
[14]. Aumente-Maestro, C., et al., BTS U-Net: A data-driven approach to brain tumor segmentation through deep learning. Biomedical Signal Processing and Control, 2025. 104: p. 107490.
[15]. Chen, W., et al., MAUNet: a mixed attention U-net with spatial multi-dimensional convolution and contextual feature calibration for 3D brain tumor segmentation in multimodal MRI. Frontiers in Neuroscience, 2025. Volume 19 - 2025.
[16]. Talukder, M.A., M. Tabassum, and M. Khalid, Self-attention U-Net (SAU-Net): An attention-driven U-Net framework for precise brain tumor segmentation using multimodal magnetic resonance imaging. Digit Health, 2026. 12: p. 20552076261426312.
[17]. Mishra, S., et al., Data-Driven Deep Supervision for Medical Image Segmentation. IEEE Trans Med Imaging, 2022. 41(6): p. 1560–1574.
[18]. Fu, Z., et al., Deep supervision feature refinement attention network for medical image segmentation. Engineering Applications of Artificial Intelligence, 2023. 125: p. 106666.
[19]. Turečková, A., et al., Improving CT Image Tumor Segmentation Through Deep Supervision and Attentional Gates. Frontiers in Robotics and AI, 2020. Volume 7 - 2020.
[20]. Maqsood, R., et al., Optimal Res-UNET architecture with deep supervision for tumor segmentation. Frontiers in Medicine, 2025. Volume 12 - 2025.
[21]. Gao, Y., et al., Medical Image Segmentation: A Comprehensive Review of Deep Learning-Based Methods. Tomography, 2025. 11(5).
[22]. Ma, J., et al., Loss odyssey in medical image segmentation. Medical Image Analysis, 2021. 71: p. 102035.
[23]. Ma, J., et al., Loss odyssey in medical image segmentation. Med Image Anal, 2021. 71: p. 102035.
[24]. Li, B., et al., Region-related focal loss for 3D brain tumor MRI segmentation. Med Phys, 2023. 50(7): p. 4325–4339.
[25]. Yeung, M., et al., Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Comput Med Imaging Graph, 2022. 95: p. 102026.
[26]. Visalakshi, G. and L. Mohan, A novel sub-differentiable hausdorff loss combined with BCE for MRI brain tumor segmentation using UNet variants. Scientific Reports, 2025. 15(1): p. 45136.
[27]. Zhou, T., S. Ruan, and B. Lei, BUFNet: Boundary-aware and uncertainty-driven multi-modal fusion network for MR brain tumor segmentation. Medical Image Analysis, 2026. 107: p. 103855.
[28]. Yang, L., et al., MUNet: a novel framework for accurate brain tumor segmentation combining UNet and mamba networks. Frontiers in Computational Neuroscience, 2025. Volume 19 - 2025.
[29]. Kofler, F., et al., BraTS Toolkit: Translating BraTS Brain Tumor Segmentation Algorithms Into Clinical and Scientific Practice. Frontiers in Neuroscience, 2020. Volume 14 - 2020.
[30]. Jiang, Z., et al., ENHANCING GENERALIZABILITY IN BRAIN TUMOR SEGMENTATION: MODEL ENSEMBLE WITH ADAPTIVE POST-PROCESSING. Proc IEEE Int Symp Biomed Imaging, 2024. 2024.
[31]. Banerjee, S. and S. Mitra, Novel Volumetric Sub-region Segmentation in Brain Tumors. Frontiers in Computational Neuroscience, 2020. Volume 14 - 2020.
[32]. Hardani, D.N.K., H.A. Nugroho, and I. Ardiyanto. An Automatic Brain Tumor Segmentation Using 3D Residual U-Net. in 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE). 2022.
[33]. Schlemper, J., et al., Attention gated networks: Learning to leverage salient regions in medical images. Medical Image Analysis, 2019. 53: p. 197–207.
[34]. Alshomrani, S., M. Arif, and M.A. Al Ghamdi, SAA-UNet: Spatial Attention and Attention Gate UNet for COVID-19 Pneumonia Segmentation from Computed Tomography. Diagnostics (Basel), 2023. 13(9).
[35]. He, X., et al., Deep Convolutional Neural Network With a Multi-Scale Attention Feature Fusion Module for Segmentation of Multimodal Brain Tumor. Frontiers in Neuroscience, 2021. Volume 15 - 2021.
[36]. Rehman, A., et al., Selective Deeply Supervised Multi-Scale Attention Network for Brain Tumor Segmentation. Sensors (Basel), 2023. 23(4).
[37]. Yang, J., et al., MSDS-UNet: A multi-scale deeply supervised 3D U-Net for automatic segmentation of lung tumor in CT. Computerized Medical Imaging and Graphics, 2021. 92: p. 101957.
[38]. Park, M., et al., ES-UNet: efficient 3D medical image segmentation with enhanced skip connections in 3D UNet. BMC Med Imaging, 2025. 25(1): p. 327.
[39]. Yuan, D., et al., μ-Net: Medical image segmentation using efficient and effective deep supervision. Computers in Biology and Medicine, 2023. 160: p. 106963.
[40]. Huang, D., et al., Learning rich features with hybrid loss for brain tumor segmentation. BMC Med Inform Decis Mak, 2021. 21(Suppl 2): p. 63.
[41]. Montazerolghaem, M., et al., U-Net Architecture for Prostate Segmentation: The Impact of Loss Function

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26may529

on System Performance. Bioengineering (Basel), 2023. 10(4).
[42]. Liu, Z., et al., Innovative multi-class segmentation for brain tumor MRI using noise diffusion probability models and enhancing tumor boundary recognition. Scientific Reports, 2024. 14(1): p. 29576.
[43]. Isensee, F., et al., Automated design of deep learning methods for biomedical image segmentation. arXiv preprint arXiv:1904.08128, 2019.
[44]. Abrar, M., et al., Enhancing brain tumor segmentation using attention based convolutional UNet on MRI images. Sci Rep, 2025. 15(1): p. 36603.
[45]. Chen, J., et al., TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Medical Image Analysis, 2024. 97: p. 103280.
[46]. Hatamizadeh, A., et al. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. in International MICCAI brainlesion workshop. 2021. Springer.
[47]. Wu, X., et al., FCFDiff-Net: full-conditional feature diffusion embedded network for 3D brain tumor segmentation. Quant Imaging Med Surg, 2025. 15(5): p. 4217–4234.
[48]. Chen, W., et al., MAUNet: a mixed attention U-net with spatial multi-dimensional convolution and contextual feature calibration for 3D brain tumor segmentation in multimodal MRI. Frontiers in Neuroscience, 2025. Volume 19 - 2025.