Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Deep Neural Network Architectures for Brain Metastasis Segmentation: A Comparison of 3D U-Net and 3D nnU-Net on FLAIR-Weighted MR Imaging

Mohamad Ali Kavehpour
Isfahan University of Medical Sciences
Daryoush Shahbazi-Gahrouei
shahbazi@med.mui.ac.ir

Isfahan University of Medical Sciences
Mahnaz Etehad Tavakol
Isfahan University of Medical Sciences
Simin Hemati
Isfahan Universi
Pooya Akbari
Isfahan Universi

Research Article

Keywords: Brain cancers, Deep learning, Magnetic resonance imaging, 3D nnU-Net, 3D U-Net

Posted Date: April 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8754688/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Brain metastasis is one of the most common intracranial tumors, and accurate diagnosis is crucial for effective treatment planning and overall patient survival. Manual segmentation of tumors and their sub-regions is often time-consuming and costly. In recent years, artificial intelligence has made remarkable advances in medical imaging. The study aims to compare two U-Net-based neural network architectures (3D U-Net and 3D nnU-Net) and to evaluate their performance in clinical settings using magnetic resonance imaging (MRI).

Methods

This study is a retrospective case-cohort. A dataset of 200 case images from the Brain Tumor Segmentation (BraTS) and 15 cases from multiple imaging centers was used for training and testing, respectively. Two neural network models (3D U-Net and 3D nnU-Net) were utilized and compared. Each case included FLAIR weighted MR imaging in Multiple center cases. Various metrics are Dice similarity coefficient (DSC), Hausdorff Distance 95 (HD95), precision, Jaccard index, and recall for tumor segmentation. Statistical analysis included descriptive statistics, Wilcoxon signed-rank testing, and agreement analysis ($P < 0.05$).

Results

The 3D U-Net showed a mean tumor DSC of 0.791 ± 0.01, while the 3D nnU-Net indicated a DSC of 0.764 ± 0.25. For 3D U-Net, HD95 was found to be 15.11 ± 1.33, while 13.74 ± 26.89 for 3D nnU-Net. The amount of 0.669 ± 0.001 and 0.607 ± 0.21 achieved a Jaccard index and 0.834 ± 0.01 and 0.7 ± 0.22 precision for 3D U-Net and 3D nnU-Net, respectively.

Conclusions

A visual evaluation of the segmentation outputs clearly demonstrated that 3D U-Net provided more consistent delineation of necrotic regions, whereas 3D nnU-Net showed higher robustness for segmenting edema and enhancing tumor regions. Overall, 3D nnU-Net demonstrated improved performance compared to the 3D U-Net model. However, these differences were not statistically significant.

1. Introduction

Brain metastases (BMs) constitute the most common type of intracranial (IC) malignancies and, despite substantial advancements in therapeutic and supportive care strategies, are still associated with

markedly poor overall survival (OS) (1). They generally arise from lung, breast tumors, and melanoma, which constitute the most frequent primary sources (2). Neurological manifestations in patients with BMs primarily arise from extensive peritumoral vasogenic edema, which can exceed the size of the lesion and disrupt neural function. Common symptoms include headache, visual disturbances, focal deficits, increased intracranial pressure, and epileptic seizures (3). Early diagnosis of BMs is clinically critical, as it enables prompt intervention and can lead to clinically meaningful improvements in patients' quality of life (4). Radiation therapy, surgery, chemotherapy, targeted agents, immunotherapy, or a combination of these constitute the principal therapeutic modalities for BMs, with treatment selection primarily determined by the number, size, and anatomical location of the lesions (5, 6).

Tumor segmentation is a foundational step in medical image analysis, aimed at the automated or semi-automated delineation of neoplastic regions from surrounding healthy brain tissue to achieve precise quantification of tumor location, volumetric extent, morphological characteristics, and spatially explicit boundaries (7). Accurate segmentation also helps prevent unnecessary and costly interventions. Therefore, developing non-invasive methods capable of automating tumor segmentation would provide substantial clinical value for diagnostic and adaptive treatment planning. Computed tomography (CT) is often used as the initial imaging modality, particularly in emergency settings where rapid assessment of neurological symptoms is required; however, its sensitivity is limited compared with Magnetic Resonance Imaging (MRI). High-resolution contrast-enhanced MRI has been well established as the gold standard for detecting BMs (8). Nevertheless, manual segmentation remains a labor intensive, time consuming, and operator dependent process, the inherent subjectivity of which can introduce substantial errors that adversely affect treatment planning.

Thresholding and k-means clustering are among the traditional automated approaches used for segmentation. Multilevel thresholding techniques have been proposed for brain metastasis segmentation, demonstrating superior performance metrics, including higher Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), compared with conventional baseline methods. However, these approaches exhibit suboptimal performance when applied to heterogeneous tumors or lesions of small volume (9). Basin-based methods delineate boundaries by exploiting gradient information. However, they are inherently sensitive to noise and often require substantial preprocessing to ensure stable performance (10).

Artificial intelligence (AI) has witnessed rapid growth within healthcare systems and is increasingly regarded as one of the most powerful tools for disease diagnosis and treatment (11). Deep learning, a prominent branch of artificial intelligence, has the capacity to learn task relevant representations directly from raw data, thereby substantially reducing the need for manual feature engineering. Its capabilities have been successfully demonstrated across a wide range of domains, including image recognition, speech processing, and natural language understanding (12). Building on these capabilities, recent studies have demonstrated that advanced deep learning models, such as hybrid U-Net architectures and CNN–LSTM frameworks, can further enhance medical image segmentation performance. By incorporating transfer learning, feature engineering, and temporal information, these approaches achieve

improved DSCs even when training data are limited (13). In addition, Naser et al. (14) investigated a hybrid architecture that combines VGG16 and U-Net for glioma detection, grading, and tumor segmentation. Regarding the application of deep learning in segmentation. Their model achieved a mean DSC of 0.84 for tumor segmentation. However, an independent and multicenter dataset was not used as the final evaluation in this study. Moreover, in this study, the DSC was calculated only between a single neural network model and the manually masked. Another study, Rasool et al. (15), used the Brain Tumor Segmentation (BraTS) 2018 dataset for the segmentation of high-grade glioma tumors and the architecture of their network is based on 3D U-Net model. The DSCs for the whole tumor, tumor core, and enhancing tumor were 0.89, 0.95, and 0.90, respectively. However, no post-processing techniques were applied, and the employed 3D U-Net architecture was entirely baseline, without any architectural modification. After that, Ahsan et al. (16) conducted a comparative analysis of several neural network architectures for pixel-wise tumor segmentation. All models were trained on two publicly available benchmark datasets, namely Brain Tumor Figshare (BTF) and BraTS 2018. They concluded that the combination of YOLOv5 with a 2D U-Net achieved the highest tumor-segmentation performance on the BTF dataset and outperformed both the standard 2D U-Net and Mask R-CNN architectures. The analysis was conducted on a limited number of tumor classes, which may affect the generalizability of the results, despite the inclusion of the BraTS 2018 dataset for glioma grading. In addition, the proposed framework relies on two separately trained models for detection and segmentation, increasing computational complexity. Consequently, Tabassum et al. (17) utilized a hybrid model combining 3D nnU-Net and transfer learning for the segmentation of various brain tumors, including meningiomas, brain metastases, and gliomas. The results of this study achieved DSC of 0.86 and 0.81 for meningioma and metastasis, respectively. The study lacks validation in clinical settings, which may affect the assessment of its practical applicability. Additionally, data limitations and heterogeneity are among the concerns in this study.

With the rapid advancement of medical technologies and their expanding role in clinical practice, early detection and timely management of cancer have become essential priorities. As mentioned above, MRI, a safe imaging modality, plays a valuable role in detecting and distinguishing soft tissues, masses, blood vessels, and edema. Its integration with neural network architectures has contributed to significant medical advances. This work aims to assess the performance of 3D U-Net and 3D nnU-Net architectures by combining MRI sequences with comprehensive clinical data for testing and to compare their effectiveness in the automated segmentation of BMs. Finally, by incorporating population-specific factors such as ethnic background, environmental exposures, and regional climatic conditions, this research aims to establish a locally adapted and robust predictive framework capable of improving clinical decision-making for patients with BMs.

2. Methods and Materials

2.1. Study Design and Ethical Approval

This retrospective study utilized patient data from multi-MR imaging centers (Ethical codes: IR.MUI.MED.REC.1403.499 and IR.ARI.MUI.REC.1404.037). All procedures were performed in accordance with the Declaration of Helsinki. The requirement for informed consent was waived due to the retrospective nature of the study. Data confidentiality was rigorously maintained.

2.2. Datasets and Sample size

Two datasets were used to develop and evaluate the proposed artificial intelligence–based segmentation framework.

2.2.1. Public training dataset:

A total of 200 cases from the Brain Tumor Segmentation (BraTS) 2025 Challenge were used for model training. It comprises several disease-specific challenges, including adult glioma, intracranial meningioma, brain metastases, pediatric brain tumors, and a cross-entity generalizability task. In this study, we exclusively used data from Brain Metastases. Each case included FLAIR weighted. Four subregions are annotated for each case, encompassing edema, tumor core, necrosis, and background. These data were obtained from the Synapse platform, the official hosting environment for the BraTS Challenges (https://www.synapse.org).

2.2.2. Clinical evaluation dataset:

The clinical dataset comprised MRI scans from 15 patients diagnosed with BMs, collected from multiple imaging centers. Each case included FLAIR modality and corresponding expert annotations. A rigorous quality control (QC) process was conducted to ensure complete modality availability, sufficient image quality, and the absence of artifacts. All MRI volumes and segmentation masks were independently reviewed by a radiologist using ITK-SNAP to verify anatomical consistency and labeling accuracy (18).

2.3. Image Preprocessing

Initially, the curated BraTS dataset undergoes verification to confirm the presence of requisite MRI modality, including FLAIR weighted as well as the corresponding segmentation channels for the three tumor sub-regions, such as tumor core, necrosis, and edema. These procedures were implemented in Google Colab, where the SimpleITK, MONAI, and NiBabel libraries were executed to facilitate data processing and preprocessing workflows.

All images were reoriented to a common RAS anatomical convention and resampled to an isotropic voxel size of 1×1×1 mm^3. To eliminate non-informative background and reduce computational load, foreground cropping was performed using the FLAIR volume, followed by symmetric padding or cropping to standardize all modalities and segmentation masks to 240×240×155 voxels.

A subject-specific brain mask was automatically extracted from the FLAIR volume using Otsu thresholding and applied for modality wise normalization. In-mask voxels were z-score normalized, whereas out of mask voxels were set to zero. The resulting mask was saved as an additional NIfTI file

for each case. The segmentation label map was explicitly co-registered to the FLAIR affine to ensure strict geometric consistency across volumes (19).

A three-dimensional Gaussian smoothing kernel was applied to estimate a low-frequency approximation of the underlying bias field. Within the confines of the brain mask, voxel-wise intensity standard deviations were computed for both the native MRI volume and its smoothed bias approximation. The Global Non-Uniformity (GNU), smaller values reflect higher intensity homogeneity, whereas elevated values indicate pronounced bias-related variation. For any modality in which the GNU metric exceeded the predefined instability threshold (GNU ≥ 0.15), a more structurally robust and bias-resistant mask was reconstructed, replacing the previous mask that exhibited significant intensity inhomogeneity. Figure 1 illustrates the preprocessing of the raw data before training the neural network.

Figure 1

2.4. Train/test/validation

For 3D U-Net baseline and optimization, the dataset was split into training, testing, and validation sets with 80% allocated for training, 10% for validation, and 10% for testing. To prevent model overfitting, this form of data splitting is a common and effective practice in machine-learning studies, particularly in deep learning (20).

In addition, 3D nnU-Net neural network architecture employed a fully automated 5-fold cross-validation strategy, consistent with its self-configuring design, and eliminating the need for manual data partitioning.

2.5. Data augmentation

In this study, integration of on-the-fly GAN-based augmentation into the training pipeline of deep learning architectures by employing GliGAN to generate synthetic tumors during training was performed. This augmentation requires no additional storage and improves lesion diversity and sensitivity to small or multi-lesion cases, while validation data remain unchanged (21).

2.6. GPU specification

All model training experiments were conducted on a high-performance computing (HPC) system equipped with an NVIDIA Tesla V100-SXM2 GPU (32 GB memory) and 64 GB of system RAM. The CUDA and driver versions were 12.2 and 535.86.10, respectively. Model implementation was performed using Python 3.10.8, PyTorch 2.7.1, and MONAI.

Data preprocessing and clinical inference were additionally performed in Google Colab Premium using an NVIDIA A100-SXM4 GPU with 80 GB of memory (CUDA 12.4, driver version 550.54.15).

2.7. 3D U-Net architecture

The U-Net architecture is an encoder–decoder convolutional neural network specifically designed for medical image segmentation (22). In 3D U-Net used in this study, the input MRI volumes of size 240×240×160 are fed into the network, and with each encoder level, the spatial dimensions are reduced by a factor of two. Following each pair of convolutional layers, a ReLU activation is applied, succeeded by a 3D max pooling operation (2×2×2) with a stride of 2 in all dimensions. In the final layer, a 1×1×1 convolution is applied to reduce the number of output channels to the desired number of segmentation labels. After training, the model contained a total of 5,652,097 parameters, of which 5,649,153 were trainable, and 2,944 were non-trainable. Two separate training configurations were evaluated.

2.7.1. First training configuration:

Model was trained for 300 epochs using a batch size of 2 and a learning rate of 1×10^{-4}. A combined Dice and binary cross-entropy loss function was used for supervision network training. The Adam optimizer was employed for parameter optimization due to its high computational efficiency, stability, and reduced need for manual hyperparameter tuning (23).

2.7.2. Second training configuration (optimized):

For the second training, several hyperparameters were modified. The loss function was modified from a combination of DSC and cross-entropy to a composite DSC, Focal, and boundary loss. Dice loss function calculates the overlap between predicted and ground truth segmentation masks. To avoid division by zero, a smoothing term (smooth = 1.0) is applied. The focal loss formula is calculated by:

$$\text{Focal Loss} = -\alpha_t \cdot (1 - p_t)^{\gamma} \cdot \log(p_t)$$

which, gamma (γ) controls the focusing power (default value: 2.0), p_t is softmax probabilities, and alpha (α), which adjusts the weight for each class (default value). Binary loss is defined by:

$$\text{Boundary Loss} = \sum(\text{Probabilities} \times \text{Distance Map})$$

Which distance map measures the distance to the nearest object boundary, encouraging the model to focus on accurately predicting object boundaries.

To improve the efficiency of learning from small lesions, we adopted 3D patch-based training with subvolumes (128×128×128) instead of full 240×240×155 volumes. The network capacity was increased by doubling the base number of feature channels from 16 to 32. Furthermore, the optimizer was switched from Adam to AdamW (learning rate = 1×10^{-4}, weight decay = 1×10^{-4}), and a learning rate scheduler (cosine annealing or Reduce LR On Plateau) was introduced to stabilize convergence.

The optimized 3D U-Net architecture consists of four encoder-decoder blocks with feature map progression doubling at each layer, starting with 32 channels (BASE_CH = 32), followed by BatchNorm3d layers and ReLU activations. The final output layer uses a 1x1 convolution to produce the segmentation output for the four classes (NUM_CLASSES = 4).

2.7.3 3D nnU-Net

3D nnU-Net is a deep learning framework endowed with fully self-configuring capabilities. It automatically adapts its preprocessing, training, and post-processing pipelines to the characteristics of the input data (24). A total of 200 label maps and 800 MRI volumes (comprising the four standard modalities) are provided to the model. One of the key advantages of 3D nnU-Net is its fully automated preprocessing pipeline. The framework autonomously configures critical hyperparameters such as batch size (batch = 2), patch size ([128, 160, 112]), median image size in voxels ([134.0, 172.0, 134.0]), and Z-score intensity normalization. The architecture employs 3×3×3 convolutional kernels with feature map dimensions of (32, 64, 128, 256, 320, 320) across successive layers. Training was conducted for 720 epochs, with the learning rate decreasing from 0.0099 to 0.00318. Total training time was approximately 12 hours and 2 minutes.

2.8. Manual segmentation

Clinical data is segmented into four regions, including background, necrosis, edema, and tumor, by a radiologist. A separate mask is generated for each class. All these masks are generated in a binary format (0 and 1). In the next step, the mask for each class, created by the radiologist, is compared with the corresponding class mask generated by two trained neural network architectures. To achieve better alignment of the masks, they must be compared in terms of intensity, orientation, and dimensions, then adjusted to account for any differences.

2.9. Evaluation metrics

To evaluate the performance of neural network architectures and compare them, various metrics such as DSC, HD95, precision, Jaccard index, and recall are used. The DSC for the whole tumor is calculated by computing the DSC for the three regions encompassing edema, necrosis, and enhancing tumor, while the DSC for the central tumor is calculated by computing the DSC for necrosis and enhancing tumor for each case. HD95 is a metric used to measure the largest discrepancy between two sets of points, typically represented by pixel or voxel sets in images. Precision indicates the proportion of instances that are truly positive among all instances predicted as positive by the model. Also, the Jaccard index is a metric used to measure the similarity between two sets. Additionally, Recall is a segmentation evaluation parameter that indicates the model's ability to identify all positive instances (25).

Statistical analysis included descriptive statistics, Wilcoxon signed-rank testing, and agreement analysis ($P < 0.05$).

3. Results

3.1. Quantitative Segmentation Performance

In this study, the quantitative performance of two different U-Net models for tumor segmentation is evaluated using the DSC, HD95, precision, Jaccard index, and recall. The results obtained from tumor

segmentation are demonstrated in Table 1.

Table 1
Validation Dice results.

Model	DSC WT	DSC TC	Mean tumor dice	HD95	Precision	Recall	Jaccard index	Epoch
3D U-Net **(baseline)**	0.703 ± 0.01	0.773 ± 0.01	0.791 ± 0.01	Class 1: 17.950 ± 2.84 Class 2: 15.110 ± 1.33 Class 3: 18.280 ± 1.25	Class 1: 0.723 ± 0.03 Class 2: 0.834 ± 0.01 Class 3: 0.824 ± 0.02	Class 1: 0.776 ± 0.01 Class 2: 0.770 ± 0.01 Class 3: 0.761 ± 0.01	Class 1: 0.624 ± 0.02 Class 2: 0.669 ± 0.001 Class 3: 0.680 ± 0.01	300
3D U-Net **(optimize configuration)**	0.738 ± 0.07	0.717 ± 0.03	0.762 ± 0.03	Class 1: 20.870 ± 2.70 Class 2: 31.550 ± 5.67 Class 3: 35.260 ± 4.93	Class 1: 0.690 ± 0.19 Class 2: 0.840 ± 0.05 Class 3: 0.813 ± 0.05	Class 1: 0.616 ± 0.01 Class 2: 0.755 ± 0.01 Class 3: 0.645 ± 0.08	Class 1: 0.478 ± 0.10 Class 2: 0.654 ± 0.05 Class 3: 0.557 ± 0.04	200
3D 3D nn U-Net	0.720 ± 0.02	0.770 ± 0.02	0.764 ± 0.25	Class 1: 19.510 ± 39.00 Class 2: 13.740 ± 26.89 Class 3: 20.310 ± 31.31	Class 1: 0.610 ± 0.32 Class 2: 0.664 ± 0.26 Class 3: 0.700 ± 0.22	Class 1: 0.637 ± 0.34 Class 2: 0.676 ± 0.27 Class 3:	Class 1: 0.404 ± 0.28 Class 2: 0.532 ± 0.26 Class 3: 0.6067 ± 0.21	720

*: WT: whole tumor and TC: tumor core

Model	DSC WT	DSC TC	Mean tumor dice	HD95	Precision	Recall	Jaccard index	Epoch
						0.700 ± 0.25		
*: WT: whole tumor and TC: tumor core								

3.1.1. 3D U-Net

Baseline configuration:

3D U-Net model baseline achieved a DSC of 0.703 ± 0.01 for Whole Tumor (WT) segmentation, 0.773 ± 0.01 for Tumor Core (TC), and a mean tumor DSC of 0.791 ± 0.01 (Fig. 2A, B). The best values for Dice per class (classes 1, 2, and 3) achieved 0.7682 at epoch 141, 0.8018 at epoch 179, and 0.8089 at epoch 154, respectively (Fig. 2C). In addition, the HD95, precision, Jaccard index, and recall values for all three classes are reported in Table 1. Figure 2D presents box plots of the HD95 values for all three classes. The best values for the Jaccard index per class 1, 2, and 3 saved 0.6237 (at epoch 141), 0.6691 (at epoch 179), and 0.6803 (at epoch 154), respectively (Fig. 5A).

Optimized configuration:

The optimized 3D U-Net trained for 200 epoch, performed with a DSC of 0.738 ± 0.07 for WT, 0.717 ± 0.03 for TC, and a mean tumor DSC of 0.762 ± 0.03 (Fig. 3A, 3B). The best values for DSC per class (classes 1, 2, and 3) achieved 0.7220 at epoch 179, 0.8184 at epoch 165, and 0.7505 at epoch 188, respectively (Fig. 3C). Indeed, the HD95 values, precision, and Recall scores per class are calculated (Table 1). In addition, the best Jaccard index for classes 1, 2, and 3 was recorded at epoch 170, epoch 165, and epoch 200, respectively (Table 1 and Fig. 5B).

Table 1

Figure 2

Figure 3

3.1.2. 3D nnU-Net

3D nnU-Net model was trained for 720 epochs. Early stopping was performed to prevent overfitting. The segmentation performance of this architecture refers to WT and TC DSC, 0.720 ± 0.02 and 0.770 ± 0.02, respectively (Fig. 4A). The best mean tumor DSC was derived as 0.764 ± 0.25. Consequently, the findings obtained from DSC, HD95, precision, and recall for each class are presented in Table 1. DSC for classes 1, 2, and 3 were obtained 0.8481 (by epoch 105), 0.8720 (by epoch 672), and 0.8095 (by epoch 705), respectively (Fig. 4C). As shown in Fig. 5, Jaccard index Box plots were generated for 40 validation cases per class.

Figure 4

Figure 5

3.2. Training and Validation Loss results

3.2.1. 3D U-Net (baseline)

After training the 3D U-Net, the training and validation losses started at 1.2461 and 1.1688, respectively (Fig. 6A). At the beginning of the stages, a sharp decrease indicates that the models improve quickly. At the end of training, the training and validation losses were 0.1116 and 0.1763, respectively. The lowest loss occurred at epochs 0.1082 and 0.1755, and at epoch 291, while signs of overfitting emerged after epoch 140.

3.2.2. 3D U-Net (with configuration)

The training and validation losses are evaluated as 0.7622 and 0.6195, respectively. At the end of the training model, the training and validation losses were 0.1682 and 0.1788. The lowest train and validation were 0.1682 at epoch 200 and 0.1738 at epoch 185, with the overfitting beginning around epoch 115 (Fig. 6B).

3.2.3. 3D nnU-Net

The training and validation losses start at -0.0381 and – 0.2186, respectively. As indicated in Fig. 6C, the minimum training loss occurs at epoch 676 with a value of -0.7904, and the minimum validation loss occurs at epoch 668 with a value of -0.6614. As for overfitting, it appears to start around epoch 668.

Figure 6

3.3. Clinical Test Cohort Characteristics

After training 3D U-Net and 3D nnU-Net models, the data from 15 clinical cases, obtained from various imaging centers, were used as test data for the models .74 metastatic tumors were identified across the evaluated clinical cases. The mean age of the patients is 47.93. Ten male and five female patients were included as test subjects in the study. The input data for the study mainly consists of tumors originating from the lungs and prostate.

3.4. MRI Scanners Specification

Table 2 summarizes the specifications of the MRI scanners used in this study. The table presents the Field of View (FOV) in millimeters and the corresponding MRI machine brand for each device. The FOV values are provided for different MRI scanners with varying configurations.

Table 2
MRI Scanners Characteristic.

MRI Scanner Brand	FOV (in millimiters)	Slice Thickness (in millimiters)
Philips	250 × 250	1
Symphony	512 × 448	5.5
Avanto	230 × 230	5
Interna	220 × 220	5

Table 2

3.5. Visual Assessment

As the radiologist's mask is drawn based on the FLAIR image, 3D U-Net and 3D nnU-Net models were trained using this modality, which served as the reference modality. It should be noted that, for this purpose, a 3D U-Net baseline, which demonstrated relatively better performance than the modified model, is compared with 3D nnU-Net, whose hyperparameters are automatically optimized at each epoch. In contrast, 3D nnU-Net model exhibited high sensitivity to domain shift.

The masks manually delineated by a radiologist for each class, alongside those generated by the neural network architectures, are presented. Also, Table 4 shows masks for the same cases listed in Table 3. The DSC for each class, as well as for 3D U-Net and 3D nnU-Net neural networks, is presented separately in Tables 5 and 6, respectively.

Table 5
Results of DSC for each case (3D U-Net).

Case	DSC class 1	DSC class 2	DSC class 3
1	0.8	0.6	0.75
2	0.84	0.476	0.6315
3	-	0.19	0.6
4	0.85	0.35	0.6

Table 6
Results of DSC for each case (3D nnU-Net).

case	DSC class 1	DSC class 2	DSC class 3
1	0.65	0.6	0.7
2	0.6842	0.7458	0.842
3	-	0.22	0.55
4	0.85	0.55	0.6

3.6. Statistical Analysis

Table 7 indicates the statistical comparison of DSC for each class and mean tumor between 3D U-Net and 3D nnU-Net models. As shown in Table 6, the 3D U-Net achieved slightly higher mean DSC for Class 1, whereas 3D nnU-Net demonstrated higher DSC values for Classes 2 and 3. However, the Wilcoxon signed-rank test revealed that none of these differences were statistically significant ($p > 0.05$).

Table 7
Statistical comparison of Dice similarity coefficients per class using the Wilcoxon signed-rank test.

Dice Similarity Coefficients	3D U-Net	3D nnU-Net	p-value ($p < 0.05$)
Class 1	0.771 ± 0.15	0.71 ± 0.15	0.063
Class 2	0.571 ± 0.25	0.65 ± 0.23	0.167
Class 3	0.62 ± 0.2	0.676 ± 0.17	0.443
Mean tumor	0.652 ± 0.19	0.672 ± 0.18	0.776
Values are reported as mean ± standard deviation, and the Wilcoxon signed-rank test was used to assess statistical significance.			

Table 3

Table 4

Table 5

Table 6

Table 7

4. Discussion

In this study, 3D U-Net and 3D nnU-Net neural network models were trained using the BraTS 2025 dataset to evaluate and compare these models for the segmentation of different regions of brain metastasis. After training, the models were assessed in terms of performance using clinical MRI data.

According to Table 1, the highest average DSC for the tumor class was observed in the baseline 3D U-Net model around the midpoint of the training process. However, 3D U-Net optimized model, after several adjustments to its code, showed a gradual and consistent increase in the DSC (Fig. 3A) compared to the baseline 3D U-Net (Fig. 2A). The highest DSC for this model was recorded at 0.760 during the final epoch of training. Notably, the overall DSC for the tumor class was considerably higher than the baseline model, although the DSC for the central tumor was still lower. These results indicate that the baseline 3D U-Net model performed better in capturing the masses compared to the adjusted version. According to related works (15, 26), it has been shown that the 3D U-Net model demonstrates better performance in training and segmenting glioma tumors and their various regions compared to the present study. In general, brain metastasis tumors are often smaller and more dispersed than glioma tumors, making their detection significantly more challenging. Additionally, studies conducted for tumor segmentation have frequently employed hybrid deep learning models, transfer learning, or hybrid versions, which have yielded higher results. However, the calculation of HD95 showed that the modifications made to 3D U-Net model (Table 1, Fig. 2D and 3D) not only reduced the value but actually increased it. However, by evaluating the precision metric, 3D U-Net model with modifications demonstrated better performance in detecting positive instances of edema compared to the baseline model, with the maximum precision observed particularly towards the end of the training. This indicates that the modified model has the potential to achieve significantly better results compared to the baseline model in longer training sessions.

The baseline 3D U-Net model demonstrates higher precision for Classes 1 and 3, while the modified model shows superior precision for Class 2. Evaluating recall and the results, the classic 3D U-Net model outperforms the modified architecture, especially because all three classes show a more significant decrease. This indicates that the baseline model exhibits higher sensitivity across all three classes, with the highest sensitivity observed for Class 2. Reviewing the study by Beers et al. (27) a 3D U-Net was used for brain tumor segmentation with input data from the BraTS 2018 dataset, achieving DSCs of 0.732 and 0.73 for the enhancing tumor and tumor core, respectively. The model performed better on the tumor core, with a DSC of 0.773 ± 0.01, while the DSC for the whole tumor decreased slightly. Reviewing the relevant literature to this study (17), which achieved higher DSC in brain tumor segmentation, the use of a hybrid version consisting of 3D nnU-Net architecture and transfer learning has enabled better performance in class detection. However, this model had lower performance in Class 1. This is again because the cases with necrosis were fewer in number. According to the analysis conducted on the validation data, the number of lesions corresponding to Classes 1, 2, and 3 was 100, 216, and 273, respectively. Of these, the model successfully identified 75, 149, and 160 lesions for each class, respectively. The DSC calculated for all three classes using the 3D nnU-Net architecture was higher than that of both the baseline 3D U-Net model and the modified version, indicating the strong performance of this model in detecting all three classes. The DSC for both 3D U-Net architectures (Fig. 2C and 3C) is significantly higher for Class 2 compared to Classes 1 and 3. This is due to its larger extent and easier segmentation for the trained model. Another reason is that some brain metastasis cases did not have necrosis or had such small areas of it that detection by the model was difficult. Class 3, which lies between Class 1 and Class 2, has a smaller size compared to Class 2. Although it was expected that 3D

nnU-Net model, with its self-adjusting architecture in both preprocessing and model training, would perform better, the results presented in Table 1 indicated a lower DSC for the whole tumor compared to the modified 3D U-Net model, although it outperforms the baseline 3D U-Net model. However, the Dice for the tumor core shows comparable results to the baseline 3D U-Net. Despite this, the average DSC of the tumor is lower than the baseline model. The HD95 values are better than the optimized configuration of 3D U-Net for all classes, but Classes 1 and 3 still show higher results than Class 2. These lower results are because the model used is based on a purely foundational architecture without any additional adjustments or integration with other architectures.

As shown in Table 1, the Jaccard index is reported separately for all three classes and for each neural network architecture. The Jaccard index for the baseline 3D U-Net architecture is significantly higher than that of the optimized 3D U-Net and 3D nnU-Net. In addition, for each architecture, the Jaccard index of Class 2 is higher than that of the other two classes. This metric was reported to be comparable to that of the baseline 3D U-Net architecture and the study related to the present work that employed the DeepLab V3 network by Grøvik et al. (28). However, its value for 3D nnU-Net is somewhat lower.

According to the results obtained from Fig. 6A and 6B, the loss function baseline model showed higher performance during training, whereas the modified model showed improved performance during validation. This suggests that, with continued training, the second model has the potential to yield more favorable results. The baseline 3D U-Net enters the learning process with relatively high loss values (1.24 for training and 1.17 for validation). However, the sharp decrease in these early losses shows that the network quickly grasps initial features. By the end of training, losses converge approximately 0.1116 (training) and 0.1763 (validation), indicating a reasonable balance between learning and Generalizability. After about epoch 140, the gap between the training and validation curves widens, suggesting the model is becoming overly tuned to the training data. In other words, this architecture performs steadily and predictably, reaching a sensible compromise between accuracy and robustness despite its slower loss reduction.

By computing the DSC between the radiologist-generated masks and the masks produced by the neural network models, we obtained the results presented in Tables 5 and 6 for a subset of the evaluated cases. Case 1, which contains two large tumors, for which the 3D U-Net model demonstrated superior performance across all three classes compared to the 3D nnU-Net. Similarly, in Case 2, the 3D U-Net again outperformed 3D nnU-Net; however, in this instance, the model showed particularly improved detection of the tumor region (Class 3). Case 3 contains a very large number of small tumors, for which both models demonstrated low accuracy in segmenting Classes 2 and 3. However, 3D nnU-Net achieved slightly better performance in segmenting edema, whereas the 3D U-Net showed higher accuracy in detecting the tumor region. In Case 4, the 3D U-Net demonstrated superior segmentation performance for Classes 1 and 3, whereas the 3D nnU-Net achieved higher accuracy for Class 2.

These findings further support the observation that 3D nnU-Net consistently performs better in identifying peritumoral edema and enhancing tumor, while the 3D U-Net provides more accurate

delineation of necrosis regions. This distinction is likely attributable to the architectural modifications applied to the 3D U-Net, which enhanced its sensitivity to necrotic components. It is important to note that, in this study, 3D nnU-Net architecture was employed in its standard self-configuring form, without any supplementary architectural refinements or enhanced preprocessing strategies. Therefore, it is reasonable to expect that applying similar modifications, such as optimized preprocessing pipelines, architectural refinements, hybrid model integration, or the use of transfer learning, could further improve the performance of both models. In particular, these strategies have the potential to substantially strengthen 3D nnU-Net's capabilities, enhancing not only its training efficiency but also its robustness and accuracy when evaluated on heterogeneous clinical datasets.

Based on quantitative evaluation metrics, visual inspection, and statistical analysis of clinical test cases, 3D nnU-Net architecture exhibited more favorable performance compared to 3D U-Net. This observation may be attributed to several factors. One explanation is that hyperparameter tuning in 3D U-Net, including preprocessing and training settings, was performed manually. In contrast, 3D nnU-Net employs a fully automated configuration strategy, which resulted in improved performance. This automation enhances not only general applicability but also leads to better accuracy across different clinical scenarios. Furthermore, the 3D U-Net architecture supports multi-modality integration; nnU-Net demonstrated competitive segmentation performance when applied to a single-modality configuration.

5. Limitations and Future Directions

Despite recent progress in convolutional architectures, accurate segmentation of brain metastases remains challenging. Models such as 3D U-Net, 3D nnU-Net, and their derivatives have improved delineation of enhancing tumor, necrotic-core, and edematous regions, yet several methodological and practical limitations.

First, collecting large amounts of clinical data from imaging centers was one of the challenges. Many of the datasets were either unavailable or incomplete, leading to their exclusion from the study. Another limitation is the difference in the number of slices per modality. The clinical data had a slice range between 16 and 21, whereas the collected BraTS data contained about 150 slices. Even though resizing and patching were performed to match the number of slices to those on which the model was trained, this introduces some bias. Another issue was that the neural network architectures were not able to accurately segment tumors located at the brain's edges or those in the highest slices. Finally, domain shift posed a significant challenge. The multi-center data that were collected varied in resolution, noise, intensity, scanner type, and voxel size.

Additionally, the 3D nnU-Net settings caused the fusion of masks on the images to be inferior to that of the 3D U-Net. Following the training of the neural network models, clinical testing is conducted to assess their accuracy, reliability, and applicability in real-world clinical environments. The cases included in this study were obtained from multiple medical centers, each utilizing different MRI scanners from various manufacturers (e.g., Philips, Symphony, Siemens, and others), as well as exhibiting variations in slice

thickness, field of view (FOV), and image resolution. These heterogeneities ensure a more rigorous evaluation of the model's performance across diverse acquisition settings and clinical conditions. Additionally, both architectures demonstrated difficulty in accurately segmenting small lesions and tumors located at the brain periphery, highlighting an ongoing limitation of convolutional neural networks in handling extreme spatial variability.

Several approaches are considered for future work. First, integrating different neural network architectures to achieve improved segmentation performance is of interest. For example, employing a VGG16-based U-Net architecture could be explored. Alternatively, a hybrid approach that combines a 2D VGG16 model with a 3D network may be investigated to provide improved volumetric representation of tumors. Second, a potential approach involves extracting tumor-related and other class specific information using radiomics. Third, employing a substantially larger dataset is recommended to enable more rigorous validation of the trained model on clinical data. Finally, applying task-specific enhancements to nnU-Net, such as customized preprocessing or transfer learning, may further improve its performance in heterogeneous clinical environments.

6. Conclusions

In this study, 3D U-Net and 3D nnU-Net neural network architectures were evaluated for the segmentation of different classes of brain metastases. Both models were trained using the BraTS 2025 benchmark dataset and tested on clinical MRI data collected from multiple imaging centers, highlighting their applicability to heterogeneous real-world imaging conditions. The use of FLAIR modality plays a significant role in the detection and differentiation of these lesions, and its integration with deep learning substantially enhances segmentation performance.

Based on the quantitative results, 3D U-Net architecture achieved a higher DSC than nnU-Net (0.791 ± 0.01 vs. 0.764 ± 0.25). However, observations from clinical cases indicate that 3D nnU-Net is better suited to clinical environments. Further work on segmentation accuracy improvement by integrating different network architectures or by increasing the size and diversity of the training dataset is recommended. At the end, based on both quantitative outcomes and visual assessment of clinical cases, 3D nnU-Net demonstrated improved performance compared to the 3D U-Net model. However, these differences were not statistically significant.

Declarations

Author Contributions: Conceptualization, D.S.-G., M.K.; Methodology, D.S.-G., M.K., M.E., S.H., and P.A.; Validation, D.S.-G., M.K., M.E., and S.H.; Investigation, M.K.; Resources, D.S.-G.; Data Curation, D.S.-G., M.K., M.E., S.H., and P.A.; Writing—Original Draft Preparation, M.K.; Writing—Review and Editing, D.S.-G., M.E.; Supervision, D.S.-G.; Project Administration, D.S.-G.; Funding Acquisition, D.S.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded financially (Grant Numbers: 3403806 and 1403341) by Isfahan University of Medical Sciences, Isfahan, Iran.

Institutional Review Board Statement: This article contains no studies with human participants or animals performed by the authors.

Ethics approval and consent to participate: This article contains no studies with human participants or animals performed by the authors. Ethical codes: IR.MUI.MED.REC.1403.499 and IR.ARI.MUI.REC.1404.037 by Isfahan University of Medical Sciences, Isfahan, Iran.

Consent for publication: Not Applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

Ethical Considerations and AI Use Declaration: All procedures complied with the IRB and Declaration of Helsinki standards. This manuscript was reviewed using Grok 3 solely for grammar correction and language refinement, as the authors are non-native English speakers. The intellectual content, data analysis, and conclusions are the sole responsibility of the authors.

References

1. Zuccato JA, Mamatjan Y, Nassiri F, Ajisebutu A, Liu JC, Muazzam A, et al. Prediction of brain metastasis development with DNA methylation signatures. Nat Med. 2025;31(1):116–25.
2. Jiang K, Parker M, Materi J, Azad TD, Kamson DO, Kleinberg L, et al. Epidemiology and survival outcomes of synchronous and metachronous brain metastases: a retrospective population-based study. NeuroSurg Focus. 2023;55(2):E3.
3. Le Rhun E, Weller M, Anders C, Larkin J, Li J, Moss NS, et al. Symptomatic melanoma brain metastases: A call for clear definitions and adoption of standardized tools. Eur J Cancer. 2024;208:114202.
4. Mirmoeeni S, Azari Jafari A, Shah M, Salemi F, Hashemi SZ, Seifi A. The clinical, diagnostic, therapeutic, and prognostic characteristics of brain metastases in prostate cancer: a systematic review. Prostate Cancer. 2022;2022(1):5324600.
5. Nieblas-Bedolla E, Nayyar N, Singh M, Sullivan RJ, Brastianos PK. Emerging immunotherapies in the treatment of brain metastases. Oncologist. 2021;26(3):231–41.
6. Bodensohn R, Kaempfel A-L, Boulesteix A-L, Orzelek AM, Corradini S, Fleischmann DF, et al. Stereotactic radiosurgery versus whole-brain radiotherapy in patients with 4–10 brain metastases: A nonrandomized controlled trial. Radiother Oncol. 2023;186:109744.

7. Rajendran S, Rajagopal SK, Thanarajan T, Shankar K, Kumar S, Alsubaie NM, et al. Automated segmentation of brain tumor MRI images using deep learning. IEEE Access. 2023;11:64758–68.
8. Soltaninejad A, Shahbazi-Gahrouei D, Khorasani A, Hemati S. Evaluation of CNN-based deep learning models for auto-contouring in Glioblastoma radiotherapy: A review. Radiat Oncol. 2026;20:169.
9. Sharma SR, Alshathri S, Singh B, Kaur M, Mostafa RR, El-Shafai W. Hybrid multilevel thresholding image segmentation approach for brain MRI. Diagnostics. 2023;13(5):925.
10. Mohammed AS, Mihoub Z, editors. A Review of Image Segmentation Strategies from Classical Methods to Deep Learning. 2024 Conference of Young Researchers in Electrical and Electronic Engineering (ElCon); 2024: IEEE.
11. Rezaei M, Rahmani E, Khouzani SJ, Rahmannia M, Ghadirzadeh E, Bashghareh P, et al. Role of artificial intelligence in the diagnosis and treatment of diseases. Kindle. 2023;3(1):1–160.
12. Chen X, Hu X, Huang Y, Jiang H, Ji W, Jiang Y, et al. Deep learning-based software engineering: progress, challenges, and opportunities. Sci China Inform Sci. 2025;68(1):111102.
13. Etehadtavakol M, Etehadtavakol M, Ng E. Optimizing thyroid nodule segmentation in thermal imaging with temporal sequences and advanced deep Learning backbones. Expert Syst Appl. 2025;296:129105.
14. Naser MA, Deen MJ. Brain tumor segmentation and grading of lower-grade glioma using deep learning in MRI images. Comput Biol Med. 2020;121:103758.
15. Rasool N, Bhat J. Multimodal Brain Tumor Segmentation using 3D-U-Net. 2023.
16. Ahsan R, Shahzadi I, Najeeb F, Omer H. Brain tumor detection and segmentation using deep learning. Magn Reson Mater Phys Biol Med. 2025;38(1):13–22.
17. Tabassum M, Di Ieva A, Liu S. Meta transfer learning for brain tumor segmentation using nnUNet in meningioma and metastasis cases. Sci Rep. 2025;15.
18. Yushkevich PA, Yang G, Gerig G. ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. Annu Int Conf IEEE Eng Med Biol Soc. 2016;2016:3342–5.
19. Huang J, Yagmurlu B, Molleti P, Lee R, VanderPloeg A, Noor H, et al. Brain tumor segmentation using deep learning: high performance with minimized MRI data. Front Radiol. 2025;5:1616293.
20. Sivakumar M, Parthasarathy S, Padmapriya T. Trade-off between training and testing ratio in machine learning for medical image processing. PeerJ Comput Sci. 2024;10:e2245.
21. Jain I, Willems S, Latre S, De Schepper T. On-the-Fly Data Augmentation for Brain Tumor Segmentation. arXiv preprint arXiv:250924973. 2025.
22. Agrawal P, Katal N, Hooda N. Segmentation and classification of brain tumor using 3D-UNet deep neural networks. Int J Cogn Comput Eng. 2022;3:199–210.
23. Zhang Z, editor. Improved adam optimizer for deep neural networks. 2018 IEEE/ACM 26th international symposium on quality of service (IWQoS); 2018: Ieee.
24. Luu HM, Park S-H, editors. Extending nn-UNet for brain tumor segmentation. International MICCAI brainlesion workshop. Springer; 2021.

25. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. Sci Rep. 2024;14(1):6086.
26. Jiang H, Imran M, Muralidharan P, Patel A, Pensa J, Liang M, et al. MicroSegNet: A deep learning approach for prostate segmentation on micro-ultrasound images. Comput Med Imaging Graph. 2024;112:102326.
27. Beers A, Chang K, Brown J, Sartor E, Mammen C, Gerstner E et al. Sequential 3d u-nets for biologically-informed brain tumor segmentation. arXiv preprint arXiv:170902967. 2017.
28. Grøvik E, Yi D, Iv M, Tong E, Nilsen LB, Latysheva A, et al. Handling missing MRI sequences in deep learning segmentation of brain metastases: a multicenter study. NPJ Digit Med. 2021;4(1):33.

Table 3 and 4

Table 3 and 4 are available in the Supplementary Files section.

Figures

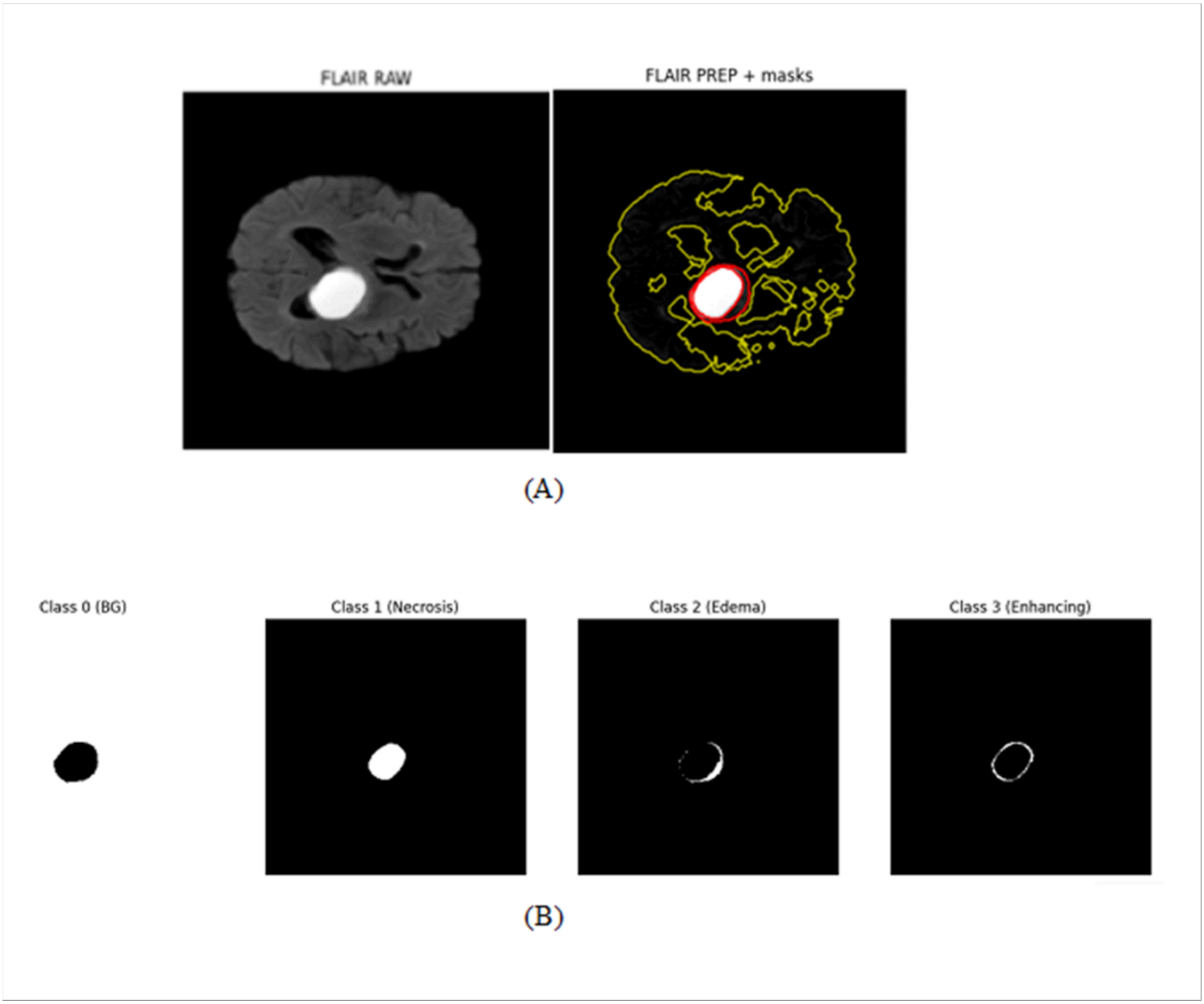


Figure 1

(A) Raw MRI data before preprocessing and Preprocessed MRI data (B) Distinct segmentation classes for a representative case (from left to right: background, necrosis, edema, and enhancing lesion)

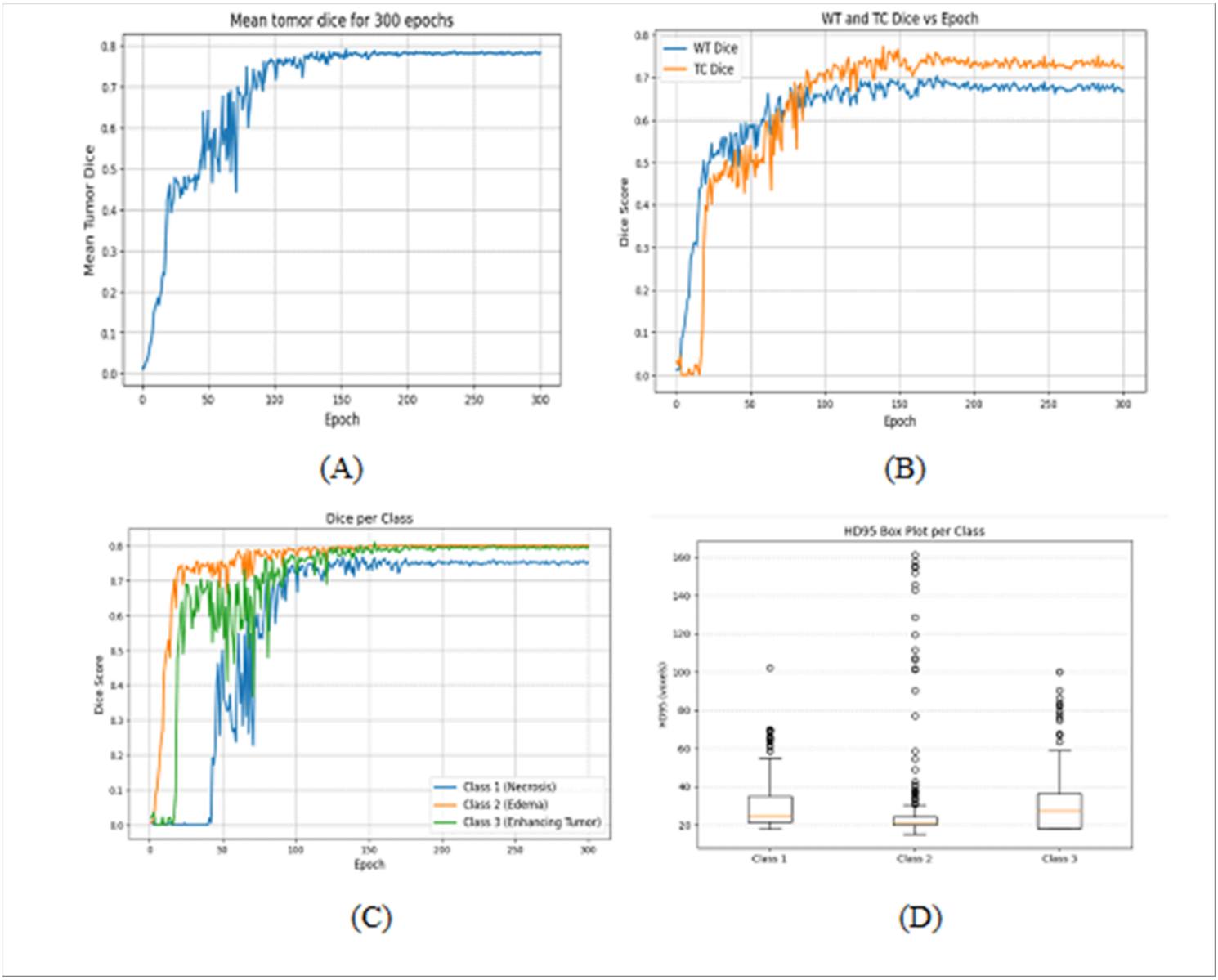


Figure 2

(A) Mean tumor dice for 300 epochs after training 3D U-Net classic (B) Whole tumor and tumor core dice for 300 epochs after training 3D U-Net classic (C) Dice per class for training of 3D U-Net baseline (D) HD95 per class after training 300 epochs of classical 3D U-Net.

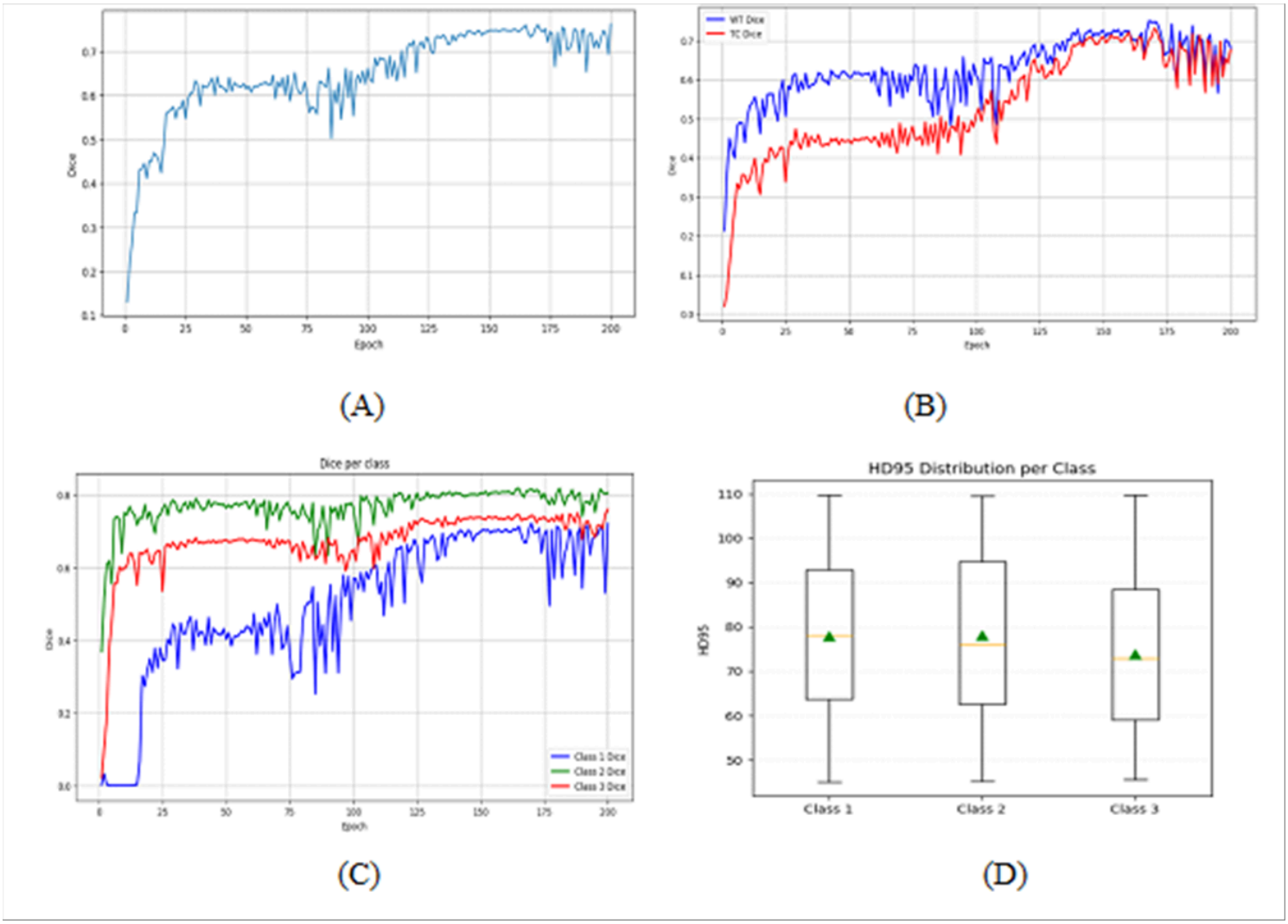


Figure 3

(A) mean tumor dice (training on 200 epoch) (B) WT and TC dice for second configuration 3D U-Net (training on 200 (C) Dice per class trained on 200 epoch (second configuration 3D U-Net), (D) HD95 per class (trained on second configuration 3D U-Net).

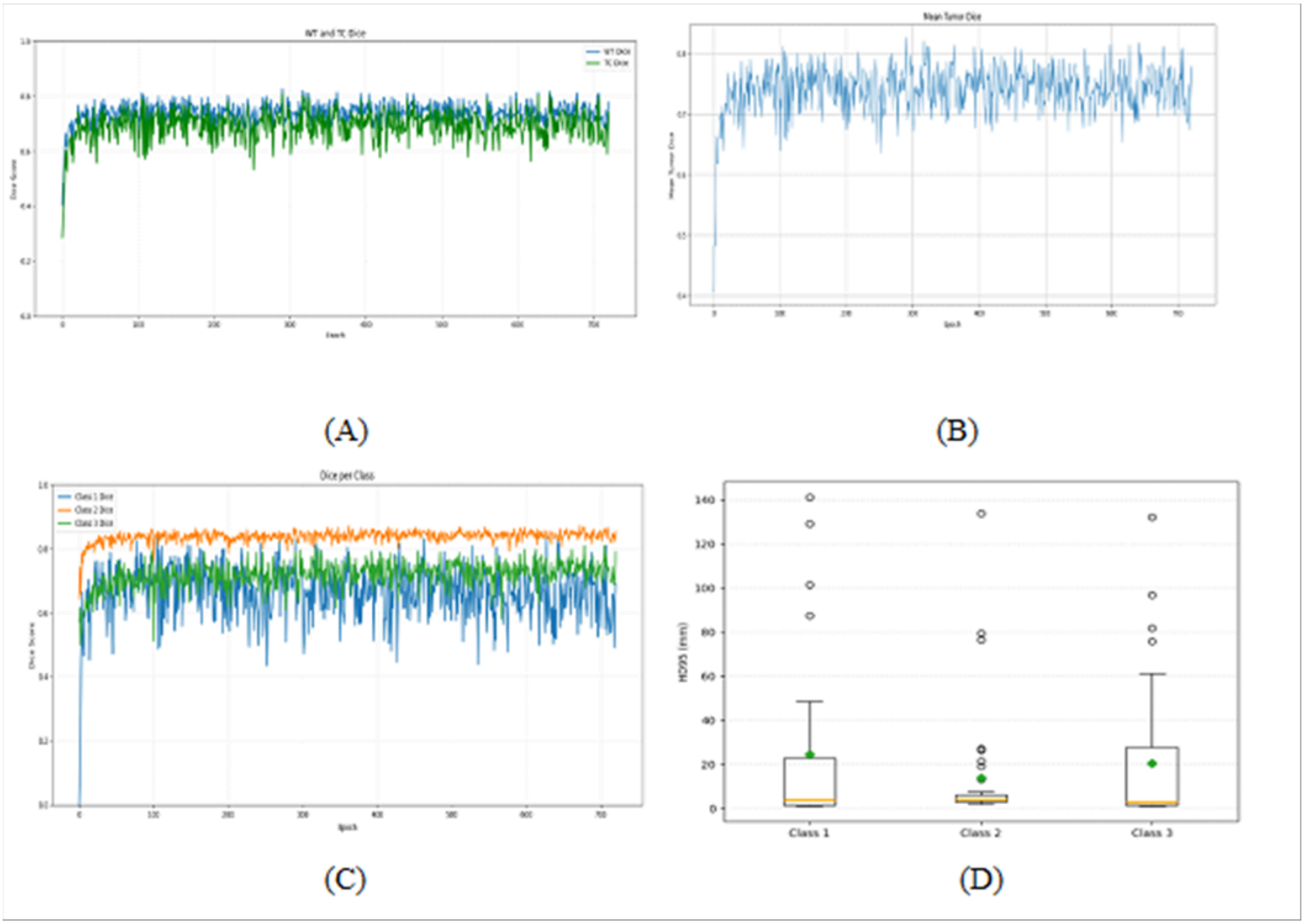


Figure 4

(A) WT and TC dice for 3D nnU-Net (B) Mean tumor dice training on 3D nnU-Net (C) DSC per class (D) HD95 per class for 40 validation cases.

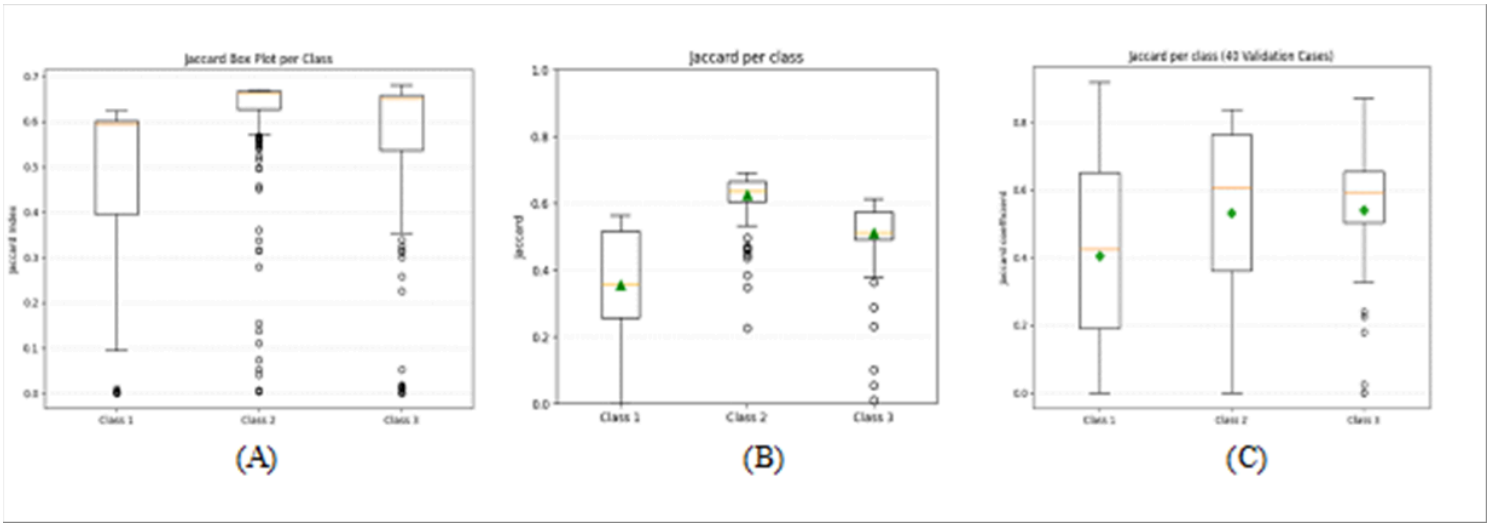


Figure 5

Jaccard index box plot per class. (A), (B), and (C) correspond to the baseline 3D U-Net, the optimized 3D U-Net, and the 3D nnU-Net, respectively.

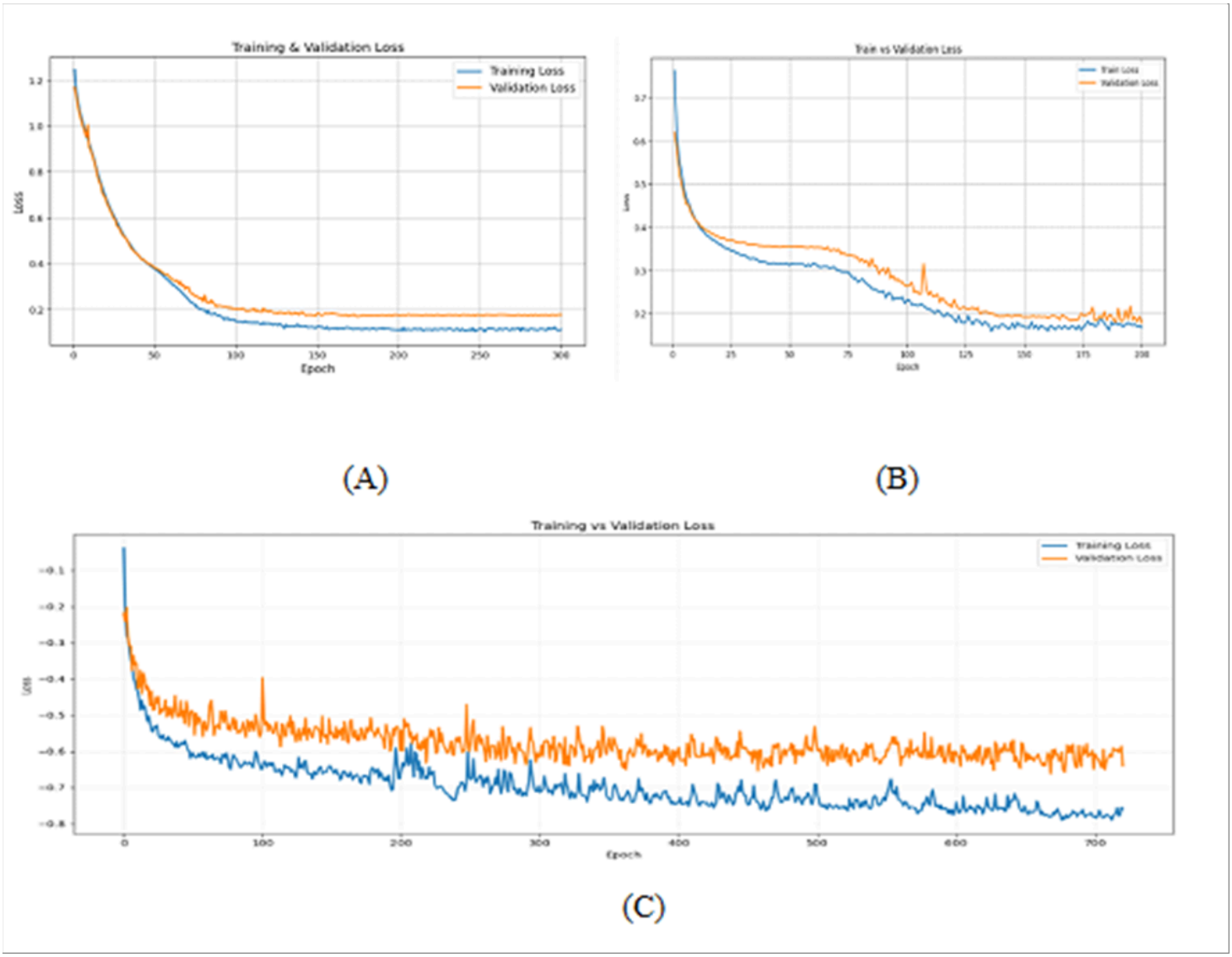


Figure 6

training and validation loss curve (A) 3D U-Net (300 epoch), (B) 3D U-Net (200 epoch), and (C) 3D nnU-Net.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table34.docx