

Clinical Reliability and Failure-Risk Analysis of Deep Learning– Based Postoperative Glioblastoma Segmentation for Radiotherapy: A Multi-Metric Evaluation Using nnU-Net

Amirmohammad Soltaninejad

Isfahan University of Medical Sciences

Daryoush Shahbazi-Gahruei

`shahbazi@med.mui.ac.ir`

Isfahan University of Medical Sciences

Mahdie Soltaninejad

McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, 3801 University Street, Montreal, H3A 2B4, Quebec, Canada

Amir Khorasani

Isfahan University of Medical Sciences

Simin Hemati

Isfahan University of Medical Sciences

Article

Keywords: Glioblastoma, Radiotherapy Planning, Deep Learning, nnU-Net, Postoperative MRI, Automated Segmentation, Clinical Reliability, Failure Analysis

Posted Date: July 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10245900/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

Deep learning–based automated segmentation has the potential to improve radiotherapy planning workflows for glioblastoma; however, safe clinical integration requires not only high accuracy but also reliable performance across heterogeneous postoperative tumor regions. This study presents a clinically oriented reliability and failure-risk evaluation of nnU-Net v2 for automated multi-class segmentation of postoperative glioblastoma using multi-modal MRI. A total of 240 cases from the BraTS 2024 postoperative dataset were analyzed using four MRI modalities (T1, T1-Gd, T2, and FLAIR). Segmentation performance was assessed using a comprehensive multi-metric framework including Dice similarity coefficient, 95th percentile Hausdorff distance (HD95), mean surface distance (MSD), precision, and sensitivity. In addition to global accuracy, a clinically motivated failure-threshold analysis was performed to identify subregion-specific reliability patterns relevant to radiotherapy planning. The model demonstrated high and consistent performance for non-enhancing tumor regions, achieving Dice scores of 0.91 ± 0.05 for the non-enhancing tumor core and 0.90 ± 0.06 for surrounding FLAIR hyperintensity. In contrast, enhancing tumor and resection cavity segmentation showed reduced accuracy and greater variability, with Dice scores of 0.66 ± 0.12 and 0.57 ± 0.15 , respectively. Failure-threshold analysis revealed substantial reliability limitations in enhancing tumor detection and cavity delineation, highlighting radiotherapy-relevant segmentation risks. These findings indicate that automated segmentation with nnU-Net can reliably delineate non-enhancing postoperative tumor regions, whereas delineating enhancing tumor and resection cavity contours requires careful expert verification. This study emphasizes the importance of failure-aware, multi-metric evaluation frameworks for the safe clinical integration of deep learning–based segmentation tools into postoperative glioblastoma radiotherapy workflows.

Introduction

Accurate delineation of postoperative glioblastoma (GBM) subregions is a fundamental prerequisite for effective radiotherapy planning. (1). In contemporary neuro-oncological practice, radiotherapy target definition relies heavily on precise identification of residual tumor components, treatment-related changes, and postoperative anatomical structures (1). Inaccurate segmentation may directly influence target volume definition, dose distribution, and sparing of surrounding healthy brain tissue, thereby affecting both tumor control and treatment-related toxicity (2).

Magnetic resonance imaging (MRI) remains the primary imaging modality for postoperative assessment and radiotherapy planning in GBM (3). Clinically relevant postoperative subregions include enhancing tissue (ET), representing areas of blood–brain barrier disruption; non-enhancing tumor core (NETC), corresponding to necrotic or residual tumor components; surrounding non-enhancing FLAIR hyperintensity (SNFH), reflecting edema and infiltrative tumor; and postoperative resection cavities (RCs) (4). Accurate delineation of these subregions is essential for defining gross tumor volume (GTV), clinical target volume (CTV), and boost regions in modern radiotherapy protocols (1).

Despite its clinical importance, manual segmentation of postoperative GBM remains labor-intensive, time-consuming, and subject to substantial inter- and intra-observer variability (5). Surgically altered anatomy, postoperative hemorrhage, air pockets, and evolving cavity morphology introduce significant uncertainty into image interpretation. These factors complicate differentiation between residual tumor, postoperative enhancement, and treatment-related changes, thereby limiting the reproducibility of radiotherapy target delineation across institutions. Recent studies have demonstrated that postoperative glioblastoma segmentation remains substantially more challenging than preoperative tumor segmentation due to surgically altered anatomy and heterogeneous residual disease patterns (2).

Recent advances in deep learning have significantly improved automated brain tumor segmentation(6). In particular, the nnU-Net framework has demonstrated robust and adaptable performance across a wide range of biomedical image segmentation tasks by automatically configuring preprocessing, architecture, and training parameters to dataset characteristics (5). Numerous studies have reported high segmentation accuracy using nnU-Net and related convolutional neural network architectures in glioma imaging (5).

However, most existing investigations have focused predominantly on preoperative imaging, where anatomical structures remain relatively intact, and segmentation tasks are less complex. Postoperative glioblastoma segmentation presents substantially greater challenges due to surgically induced cavities, heterogeneous enhancement patterns, susceptibility artifacts, and dynamic postoperative tissue changes(4). These factors reduce segmentation reliability even among expert clinicians and are insufficiently addressed in current evaluation frameworks.

Moreover, the majority of prior studies rely primarily on overlap-based metrics such as the Dice similarity coefficient to assess segmentation performance. Although useful, such metrics may obscure clinically significant boundary deviations, detection failures, or subregion-specific limitations that directly affect radiotherapy planning. For safe clinical integration of automated segmentation tools, evaluation must extend beyond global accuracy metrics and incorporate boundary reliability, detection sensitivity, and failure-risk analysis of radiotherapy-relevant structures.

To address these gaps, the present study investigates the clinical reliability and failure-risk profile of deep learning–based automated segmentation of postoperative GBM subregions using nnU-Net v2. Using the BraTS 2024 postoperative dataset, we perform a comprehensive multi-metric and subregion-specific evaluation of segmentation performance across enhancing tissue, non-enhancing tumor core, FLAIR abnormalities, and resection cavities. Particular emphasis is placed on identifying clinically significant failure modes that may influence radiotherapy target delineation.

The main contributions of this study include:

- Failure-risk–oriented evaluation of deep learning segmentation in postoperative glioblastoma
- Multi-metric reliability assessment beyond Dice score

- Dedicated analysis of resection cavity segmentation performance
- Clinically interpretable evaluation from a radiotherapy planning perspective

Accurate delineation of postoperative glioblastoma remains a critical challenge in radiotherapy planning. While deep learning methods have shown promising segmentation performance, their clinical adoption requires not only accuracy but also reliability and failure-awareness. By systematically analyzing segmentation performance across multiple tumor subregions and identifying potential failure patterns, this study provides clinically interpretable evidence for the safe integration of automated segmentation into radiotherapy workflows. To our knowledge, this is one of the first studies to provide a dedicated, failure–risk–oriented evaluation of deep learning–based postoperative glioblastoma segmentation from a radiotherapy planning perspective.

Despite rapid progress in automated glioma segmentation, most previous studies have focused primarily on global accuracy metrics and preoperative imaging scenarios. In the context of postoperative radiotherapy planning, segmentation reliability and failure-risk awareness are equally critical, as boundary inaccuracies or missed enhancing regions may directly affect target delineation and treatment margins.

Therefore, the primary objective of this study is not to propose a new segmentation architecture but to establish a clinically driven, failure-aware evaluation framework for automated postoperative glioblastoma segmentation. By integrating multi-metric analysis, subregion-specific evaluation, and clinically interpretable failure thresholds, this work provides a radiotherapy-oriented reliability assessment of deep learning segmentation performance in postoperative settings.

To our knowledge, few studies have provided a dedicated failure-risk–oriented evaluation of automated postoperative glioblastoma segmentation from a radiotherapy planning perspective.

Materials and Methods

Study Design

This retrospective methodological study was conducted to develop and evaluate a deep learning–based framework for automated multi-class segmentation of postoperative glioblastoma (GBM) subregions using multi-modal magnetic resonance imaging (MRI). The primary objective was to assess segmentation accuracy, boundary reliability, and subregion-specific performance of a state-of-the-art deep learning model in the postoperative setting, with particular relevance to radiotherapy planning. This study does not aim to propose a new segmentation architecture but rather to establish a clinically oriented failure-aware evaluation framework for postoperative radiotherapy segmentation.

All analyses were performed using publicly available anonymized imaging data. No patient re-identification was possible, and no additional data collection was conducted.

Dataset

Data Source

Imaging data were obtained from the Brain Tumor Segmentation (BraTS) 2024 postoperative glioblastoma dataset (4). This dataset is a curated, multi-institutional collection designed for postoperative brain tumor segmentation tasks and includes expert-annotated postoperative tumor subregions.

A total of 240 multi-modal postoperative MRI cases were included in this study. All data were fully anonymized and released for research purposes by the BraTS consortium.

MRI Modalities

Each case consisted of four co-registered MRI sequences routinely used in neuro-oncology and radiotherapy planning:

- T1-weighted (T1)
- Contrast-enhanced T1-weighted (T1-Gd)
- T2-weighted (T2)
- Fluid-attenuated inversion recovery (FLAIR)

These sequences provide complementary anatomical and pathological information necessary for accurate delineation of tumor subregions and postoperative cavities.

Ground-Truth Annotations

Ground-truth annotations were obtained directly from the publicly available BraTS 2024 postoperative dataset. The annotations were created according to the standardized BraTS postoperative labeling protocol by the BraTS consortium and its expert neuroradiologist annotators, as described in the original BraTS 2024 publication. No additional manual annotations were generated by the authors of the present study.

Four clinically relevant postoperative subregions were defined:

- **Enhancing tissue (ET)**: contrast-enhancing tumor components on post-contrast T1-weighted images
- **Non-enhancing tumor core (NETC)**: necrotic or residual non-enhancing tumor components
- **Surrounding non-enhancing FLAIR hyperintensity (SNFH)**: edema, infiltrative tumor, and treatment-related changes visible on FLAIR/T2
- **Resection cavity (RC)**: postoperative surgical cavities with variable signal characteristics

Data Partitioning and Validation Strategy

To ensure unbiased evaluation and methodological rigor, the dataset was divided into development and independent test cohorts:

- Development set: 200 cases
- Independent test set: 40 cases

The independent test cohort was completely excluded from all training and model selection procedures.

Within the development cohort, five-fold cross-validation was applied. In each fold:

- 160 cases were used for training
- 40 cases were used for validation

Each case served once as validation data across folds. Model selection was based on average validation performance. The final model was evaluated on the 40 cases of the independent test set.

Image Preprocessing

All preprocessing steps were performed using the standardized nnU-Net v2 automated pipeline to ensure reproducibility and eliminate operator-dependent variability (7).

Preprocessing included:

- Automated co-registration of MRI modalities
- Resampling to target voxel spacing determined from dataset statistics
- Z-score intensity normalization applied independently to each modality
- Automated foreground cropping to remove non-informative background regions

No additional manual preprocessing, skull stripping, or bias field correction beyond the default nnU-Net v2 pipeline was applied.

Deep Learning Framework

Network Architecture

Automated segmentation was performed using nnU-Net version 2 with the three-dimensional full-resolution (3D fullres) configuration. nnU-Net is a self-configuring deep learning framework that automatically adapts preprocessing, network topology, and training parameters to dataset characteristics(8, 9).

The architecture is based on a 3D U-Net encoder–decoder structure with skip connections that preserve spatial information across resolution levels. Multi-class segmentation of ET, NETC, SNFH, and RC was performed simultaneously(8, 9).

Training Configuration

Model training followed the default nnU-Net v2 configuration:

- Loss function: combined Dice loss and cross-entropy loss
- Data augmentation: random rotations, scaling, mirroring, Gaussian noise, and intensity perturbations
- Optimization: Adam
- Early stopping based on validation performance
- Fixed random seeds where applicable to improve reproducibility

Training and inference were performed on a workstation equipped with an NVIDIA Tesla V100 GPU (32 GB memory). Implementation was conducted using Python, PyTorch, and nnU-Net v2.

Table 1
Training configuration of nnU-Net v2 3D full-resolution model

Parameter	Setting
Framework	nnU-Net v2
Architecture	3D full-resolution U-Net
Input modalities	4 MRI sequences
Patch size	128×128×128
Batch size	2
Optimizer	Adam
Initial learning rate	0.001
Loss function	Dice + Cross-entropy
Data augmentation	Rotation, scaling, gamma, mirroring
Training folds	5-fold cross-validation
Hardware	NVIDIA Tesla V100 GPU
Inference mode	Full 3D inference

All experiments were conducted under identical training conditions with fixed random seeds, where applicable to enhance reproducibility.

Evaluation Metrics

Segmentation performance was evaluated separately for each subregion using complementary quantitative metrics:

Primary metrics

- Dice Similarity Coefficient (DSC)
- 95th percentile Hausdorff Distance (HD95)
- Mean Surface Distance (MSD)

Secondary metrics

- Precision
- Sensitivity (recall)
- Volumetric difference

All metrics were computed per case and summarized as the mean $\pm$ standard deviation across the independent test cohort (10).

Subregion-Specific Analysis

Performance was analyzed separately for ET, NETC, SNFH, and RC to evaluate subregion-dependent variability. Boundary accuracy and sensitivity were examined to identify systematic segmentation limitations, particularly in enhancing tissue and resection cavity regions.

A subgroup comparison between recent and chronic resection cavities was also performed to assess the influence of postoperative cavity characteristics on segmentation performance.

Statistical Analysis

Statistical analyses were conducted using Python scientific libraries. Normality of metric distributions was assessed using the Shapiro–Wilk test. Depending on distribution, either independent-samples t-tests or Mann–Whitney U tests were used for subgroup comparisons.

The Morphological irregularity index was calculated as the ratio between the lesion surface area and the equivalent spherical surface area derived from lesion volume.

A two-sided p-value < 0.05 was considered statistically significant. Effect sizes were calculated where appropriate to quantify the magnitude of observed differences.

Definition of Segmentation Failure Threshold

In the present study, segmentation failure was defined as a Dice similarity coefficient (DSC) < 0.70 for a given subregion. This threshold was selected based on methodological, statistical, and clinically interpretable considerations supported by prior literature.

First, in medical image segmentation research, a DSC of 0.70 has frequently been considered the lower boundary of acceptable spatial agreement for clinically meaningful structures, particularly in radiotherapy planning contexts where contour accuracy directly affects target definition. In glioma segmentation benchmarks such as the Brain Tumor Segmentation Challenge, high-performing models typically achieve Dice scores above 0.80–0.90 for whole tumor regions, while values below 0.70 are generally associated with visibly incomplete or structurally inconsistent segmentations.

Second, from a geometric perspective, a DSC below 0.70 implies substantial spatial mismatch. Given the mathematical formulation of Dice:

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

Dice similarity coefficient remains one of the most widely used segmentation evaluation metrics, although metric selection should be interpreted in the context of target structure characteristics and clinical objectives (11). Recent recommendations for image-analysis validation emphasize that performance metrics should be selected according to the underlying clinical problem rather than used in isolation (12).

A value below 0.70 typically corresponds to either substantial boundary deviation or partial detection failure. Previous methodological studies have shown that overlap-based metrics such as the Dice similarity coefficient become increasingly sensitive to clinically meaningful contour discrepancies as overlap decreases, particularly in small or morphologically complex structures. Consequently, DSC values below 0.70 are often associated with observable under-segmentation or over-segmentation rather than minor contour variability (11, 12).

Third, in radiotherapy contouring studies, inter-observer variability for gross tumor volume (GTV) delineation in glioblastoma generally corresponds to Dice values ranging from approximately 0.75 to 0.85, depending on region complexity. Therefore, a DSC < 0.70 falls below typical human inter-observer agreement levels and can reasonably be interpreted as performance inferior to acceptable clinical variability.

Finally, the chosen threshold was supported empirically in the present dataset. Visual inspection of cases with DSC < 0.70 consistently demonstrated one or more of the following failure patterns:

- Missed small enhancing regions
- Incomplete resection cavity delineation
- Substantial boundary displacement (> 10 mm HD95)

- Large volumetric bias on Bland–Altman analysis

These qualitative observations confirmed that $DSC < 0.70$ corresponded to practically meaningful segmentation errors rather than minor contour fluctuations.

To ensure robustness, sensitivity analyses were performed using alternative thresholds ($DSC < 0.65$ and $DSC < 0.75$), which demonstrated similar predictor trends in logistic regression modeling, confirming the stability of failure-risk associations.

Accordingly, a DSC threshold of 0.70 was selected as a conservative and clinically meaningful cutoff for defining segmentation failure in postoperative glioblastoma radiotherapy planning.

Results

Overall Segmentation Performance

The nnU-Net v2 (3D full-resolution configuration) model was successfully trained and evaluated on the BraTS 2024 postoperative dataset. Quantitative evaluation was performed on the independent test cohort of 40 cases to ensure unbiased performance assessment. Segmentation accuracy was analyzed separately for enhancing tissue (ET), non-enhancing tumor core (NETC), surrounding non-enhancing FLAIR hyperintensity (SNFH), and resection cavity (RC).

Overall, the model demonstrated high and consistent performance for non-enhancing tumor regions, particularly NETC and SNFH. In contrast, segmentation of enhancing tissue and resection cavities showed lower accuracy and higher inter-case variability across evaluation metrics.

Table 2. Characteristics of the BraTS 2024 dataset used in this study.

Parameter	Value
Total number of subjects	240
Training cases	200
Validation cases	40
Validation Type	5-fold-cross-validation
MRI modalities	T1, T1-Gd, T2, FLAIR
Annotation type	Expert manual segmentation
Tumor subregions	NETC, SNFH, ET, RC
Image format	NIfTI
Voxel spacing	Heterogeneous (resampled by nnU-Net)
Preprocessing	Automatic (nnU-Net v2 pipeline)

Implementation Details

The 3D full-resolution nnU-Net v2 framework was trained on the preprocessed dataset. The nnU-Net pipeline automatically configured preprocessing steps, network architecture, and baseline training parameters based on dataset characteristics, ensuring a standardized and reproducible training setup. Additional training settings were defined to ensure controlled and comparable experimentation.

A five-fold cross-validation scheme was applied to the training cohort. For each fold, the model was trained for a maximum of 100 epochs using a composite loss function combining Dice loss and cross-entropy loss. Training was performed with a batch size of 2 and a patch size of $128 \times 128 \times 128$ voxels. Optimization was conducted using the Adam optimizer with an initial learning rate of 0.001 and a cosine learning rate decay schedule.

Data augmentation strategies included random rotations, scaling, elastic deformations, intensity shifts, and Gaussian noise to improve model robustness and generalization. Early stopping was applied if no improvement in validation performance was observed for 20 consecutive epochs, reducing the risk of overfitting.

Inference was performed using a sliding-window approach to generate full-volume predictions for each MRI case.

Hardware and Software Environment

Training and inference were conducted on a high-performance computing node equipped with an NVIDIA Tesla V100 GPU (32 GB VRAM), an AMD EPYC 7502 CPU, and 64 GB system RAM. The model was implemented in Python 3.10.8 using PyTorch 2.7.1 with CUDA 12.6 support. The average training time per epoch was approximately 60 seconds.

Quantitative Performance by Subregion

Non-Enhancing Tumor Core (NETC)

Segmentation of the non-enhancing tumor core achieved the highest accuracy among all evaluated subregions. The mean Dice similarity coefficient (DSC) reached 0.913 ± 0.054 , indicating strong spatial agreement with expert reference annotations. Boundary accuracy was also high, with a mean 95th percentile Hausdorff distance (HD95) of 3.85 ± 1.27 mm and a mean surface distance of 1.97 ± 0.45 mm.

Precision and sensitivity were well balanced (0.922 ± 0.076 and 0.909 ± 0.061 , respectively), demonstrating reliable detection and delineation of tumor core structures. Volumetric agreement between predicted and reference segmentations remained stable across cases, with low relative volume differences and limited variability.

Surrounding Non-Enhancing FLAIR Hyperintensity (SNFH)

Segmentation of SNFH also demonstrated strong performance. The mean Dice score was 0.899 ± 0.056 , reflecting high overlap with reference annotations. This region exhibited the best boundary accuracy among all subregions, with a mean HD95 of 2.47 ± 1.22 mm and a mean surface distance of 1.20 ± 0.56 mm.

Precision was high (0.940 ± 0.058), while sensitivity was slightly lower (0.866 ± 0.080), suggesting a tendency toward conservative delineation in regions with diffuse FLAIR hyperintensity. Despite the heterogeneous nature of SNFH, volumetric differences remained low and consistent across the independent test cohort.

Enhancing Tissue (ET)

Enhancing tissue segmentation demonstrated substantially lower accuracy and greater variability compared with non-enhancing regions. The mean Dice coefficient for ET was 0.655 ± 0.12 , with marked inter-subject dispersion. This variability reflected the heterogeneous and often fragmented appearance of postoperative contrast enhancement.

Boundary-based evaluation revealed higher errors relative to other subregions, with a mean HD95 of 5.91 ± 3.86 mm and a mean surface distance of 4.54 ± 6.69 mm. Precision (0.72 ± 0.382) and sensitivity (0.626 ± 0.297) were both reduced compared with NETC and SNFH, indicating challenges in detecting small or irregular enhancing regions. Volumetric differences were also more variable, particularly in cases with minimal residual enhancement.

Resection Cavity (RC)

Resection cavity segmentation presented the greatest challenge for the automated framework. The mean Dice score was 0.570 ± 0.15 , accompanied by the highest boundary variability among all subregions. The mean HD95 reached 13.44 ± 14.73 mm, and the mean surface distance was 5.49 ± 7.63 mm.

While precision remained relatively high (0.70 ± 0.320), sensitivity was considerably lower (0.548 ± 0.348), indicating under-segmentation in a subset of cases. Large volumetric discrepancies were observed in cases with irregular or heterogeneous postoperative cavities. Performance variability reflected substantial differences in cavity morphology and signal characteristics across patients.

Table 3. Quantitative segmentation performance across tumor subregions.

Subregion	Dice (mean $\pm$ SD)	HD95 (mm)	MSD (mm)	Precision	Sensitivity
NETC	0.91 $\pm$ 0.04	3.85 $\pm$ 1.27	1.97 $\pm$ 0.45	0.92 $\pm$ 0.076	0.90 $\pm$ 0.061
SNFH	0.89 $\pm$ 0.05	2.47 $\pm$ 1.22	1.20 $\pm$ 0.56	0.91 $\pm$ 0.058	0.87 $\pm$ 0.080
ET	0.65 $\pm$ 0.12	5.91 $\pm$ 3.86	4.54 $\pm$ 6.69	0.72 $\pm$ 0.382	0.62 $\pm$ 0.297
RC	0.57 $\pm$ 0.15	13.44 $\pm$ 14.73	5.49 $\pm$ 7.63	0.70 $\pm$ 0.320	0.52 $\pm$ 0.348

Abbreviations: NETC = Non-Enhancing Tumor Core; SNFH = Surrounding Non-Enhancing FLAIR Hyperintensity; ET = Enhancing Tumor; RC = Resection Cavity; HD95 = 95th percentile Hausdorff Distance; MSD = Mean Surface Distance.

Boundary Accuracy Analysis

Boundary-based metrics demonstrated consistently low boundary errors for NETC and SNFH, indicating stable delineation of larger non-enhancing regions. In contrast, ET and RC exhibited significantly higher boundary deviations and greater variability. This effect was particularly pronounced in cases with small enhancing lesions or complex resection cavity geometries, where minor spatial deviations resulted in larger relative boundary errors.

Effect Size

Effect size analysis revealed substantial performance disparities among anatomical subregions. Compared to NETC segmentation, RC segmentation showed a very large negative effect size, indicating substantially inferior overlap accuracy. Similarly, ET segmentation exhibited a large effect size relative to both NETC and SNFH regions.

These findings confirm that performance differences are not only statistically significant but also clinically meaningful in magnitude.

Bland–Altman

Bland–Altman analysis revealed substantial volumetric disagreement in ET and RC regions. For enhancing tumor (ET), the mean volumetric bias was -758 mm^3 , with wide 95% limits of agreement (-9710 to $+8193 \text{ mm}^3$), indicating considerable inter-case variability.

Resection cavity (RC) segmentation demonstrated a larger systematic underestimation (mean bias -6954 mm^3) and markedly broader limits of agreement (-42155 to $+28248 \text{ mm}^3$), reflecting substantial volumetric instability.

In contrast, NETC and SNFH regions exhibited narrower limits of agreement and minimal systematic bias, consistent with their higher Dice stability.

Volumetric Agreement

Volumetric analysis revealed close agreement between predicted and reference volumes for NETC and SNFH, as reflected by stable Dice scores and low volumetric differences. Conversely, ET and RC segmentation showed greater volumetric variability, driven primarily by small ground-truth volumes and sensitivity to boundary deviations.

Subgroup Analysis of Resection Cavities

A subgroup analysis comparing recent and chronic resection cavities demonstrated significantly lower segmentation performance for recent postoperative cavities across both overlap- and boundary-based metrics ($p < 0.05$). Recent cavities exhibited heterogeneous signal characteristics and irregular morphology, contributing to increased segmentation uncertainty. Chronic cavities with cerebrospinal fluid-like signal characteristics were segmented more consistently.

Failure Prediction Modeling

To identify independent predictors of segmentation failure, multivariable logistic regression analysis was performed separately for enhancing tissue (ET) and resection cavity (RC) segmentation. Segmentation failure was defined as a Dice similarity coefficient (DSC) < 0.70 .

Enhancing Tissue (ET)

Univariable analysis demonstrated that a smaller ground-truth volume was significantly associated with segmentation failure ($p < 0.001$). Variables with $p < 0.10$ in univariable analysis were included in the multivariable model.

In the multivariable logistic regression model, ground-truth ET volume remained an independent predictor of failure (Odds Ratio [OR] = 0.82 per 1000 mm³ increase; 95% CI: 0.74–0.90; $p < 0.001$). This indicates that for each 1000 mm³ increase in enhancing tumor volume, the odds of segmentation failure decreased by approximately 18%.

The model demonstrated good discriminative ability with an area under the receiver operating characteristic curve (AUC) of 0.86 (95% CI: 0.75–0.96). The Hosmer–Lemeshow test showed no evidence of poor model fit ($p = 0.47$).

Resection Cavity (RC)

For resection cavity segmentation, both cavity volume and morphological irregularity index were associated with failure in univariable analysis ($p < 0.01$).

In multivariable analysis:

- Smaller cavity volume independently predicted failure (OR = 0.89 per 1000 mm³ increase; 95% CI: 0.81–0.97; p = 0.008).
- Higher morphological irregularity index significantly increased failure risk (OR = 1.42; 95% CI: 1.12–1.80; p = 0.003).

The final model achieved an AUC of 0.83 (95% CI: 0.71–0.94), indicating good predictive performance.

Variance inflation factor (VIF) analysis confirmed the absence of multicollinearity (all VIF < 2.0).

Summary of Findings

In summary, the nnU-Net v2 framework achieved high and reproducible segmentation performance for non-enhancing postoperative tumor regions, particularly NETC and SNFH. Enhancing tissue and resection cavity segmentation remained more challenging, with increased variability and reduced sensitivity. These findings highlight subregion-specific performance differences and emphasize the importance of multi-metric evaluation when assessing automated postoperative glioblastoma segmentation.

Table 4. Observed segmentation failure patterns.

Subregion	Common error type	Clinical relevance	Failure rate (%)
NETC	Minor boundary deviation	Low	0%
SNFH	Peripheral under-segmentation	Low	0%
ET	Missed small enhancing regions	High	73%
RC	Irregular cavity boundaries	High	47%

Abbreviations: NETC = Non-Enhancing Tumor Core; SNFH = Surrounding Non-Enhancing FLAIR Hyperintensity; ET = Enhancing Tumor; RC = Resection Cavity.

Discussion

The present study investigated the clinical reliability and failure-risk profile of deep learning–based automated segmentation of postoperative glioblastoma subregions using the nnU-Net v2 framework. By focusing specifically on postoperative anatomy and radiotherapy-relevant structures, this work provides a clinically contextualized assessment of segmentation performance beyond conventional accuracy reporting.

From a radiotherapy planning perspective, segmentation reliability is not solely determined by global overlap metrics. Even small boundary deviations in enhancing tumor or resection cavity regions can influence the definition and planning of clinical target volumes. Therefore, failure-risk awareness is essential before integrating automated segmentation into clinical workflows. The present findings suggest that deep learning–based segmentation can serve as a decision-support tool, but expert supervision remains necessary, particularly in anatomically complex postoperative regions.

Interpretation of Overall Performance

The findings demonstrate that automated segmentation using nnU-Net v2 can achieve high and consistent accuracy for non-enhancing tumor regions, particularly the non-enhancing tumor core (NETC) and surrounding non-enhancing FLAIR hyperintensity (SNFH). These regions exhibited strong spatial overlap, stable boundary accuracy, and consistent volumetric agreement across the independent test cohort. The relative homogeneity and larger spatial extent of non-enhancing regions likely facilitate feature learning and contribute to robust model performance.

From a radiotherapy planning perspective, reliable delineation of non-enhancing tumor components is of particular importance in the postoperative setting, where residual disease may not demonstrate prominent contrast enhancement. Accurate segmentation of these regions supports more consistent target volume definition and may help reduce inter-observer variability in contouring clinical target volumes. The balanced precision and sensitivity observed for NETC and SNFH suggest that automated segmentation could be integrated as a supportive tool for these structures with appropriate clinical verification.

Challenges in Enhancing Tissue Segmentation

In contrast, segmentation of enhancing tissue demonstrated substantially lower accuracy and greater variability. Fragmented postoperative enhancement, small residual tumor volumes, and confounding treatment-related changes contribute to increased segmentation difficulty. The observed sensitivity deficits for enhancing tissue are particularly relevant for radiotherapy planning, as under-segmentation of viable tumor regions may result in underestimation of boost volumes or biologically active disease.

These findings highlight that enhancing tissue segmentation represents the most clinically critical failure mode identified in this study. While moderate over-segmentation may be acceptable from a dosimetric perspective, missed enhancing regions could have direct implications for local control. Consequently, automated segmentation outputs for enhancing tissue should be interpreted cautiously and require expert review before clinical use.

Resection Cavity Delineation and Boundary Uncertainty

Resection cavity segmentation exhibited the greatest variability across all evaluated metrics. Postoperative cavities present substantial heterogeneity in size, morphology, and signal characteristics depending on surgical technique, presence of blood products, and time since surgery. This variability complicates feature extraction and leads to increased boundary uncertainty, particularly in recently formed cavities.

The low Dice performance in RC segmentation should be interpreted in the context of highly irregular postoperative geometries and the known sensitivity of overlap-based metrics to boundary perturbations in cavity-like structures.

Accurate delineation of resection cavities is clinically relevant for radiotherapy planning, as cavity margins often influence boost volume definition and dose gradients near critical brain structures. Boundary deviations in cavity segmentation may therefore affect local dose distribution, especially in highly conformal treatment techniques such as VMAT or stereotactic radiotherapy. The observed performance variability underscores the need for manual verification of automated cavity contours before clinical implementation.

Importance of Failure-Aware Multi-Metric Evaluation

A central contribution of this study is the implementation of a failure-aware and multi-metric evaluation framework tailored to radiotherapy-relevant segmentation reliability. Many previous investigations have relied primarily on overlap-based metrics such as the Dice similarity coefficient, which may mask clinically significant boundary deviations or detection failures. By incorporating boundary-based metrics, sensitivity analysis, and subregion-specific evaluation, the present study provides a more comprehensive understanding of segmentation reliability.

This approach enables identification of clinically meaningful failure modes that are not apparent from aggregate accuracy metrics alone. In particular, sensitivity deficits in enhancing tissue detection and boundary uncertainty in resection cavity delineation emerged as the most relevant limitations for radiotherapy applications. Such insights are essential for safe clinical translation of automated segmentation tools, where understanding model limitations is as important as achieving high average accuracy.

Comparison with Existing Literature

Previous deep learning–based glioblastoma segmentation studies have predominantly focused on preoperative imaging, where tumor morphology remains relatively intact and segmentation tasks are less complex. Under these conditions, high Dice similarity coefficients have been consistently reported for whole tumor and edema regions. However, postoperative segmentation presents substantially greater challenges due to surgically altered anatomy and heterogeneous imaging characteristics. Similar reductions in segmentation performance for postoperative enhancing tumor regions have been reported in recent postoperative glioblastoma segmentation studies(13).

The present findings align with emerging evidence that segmentation performance decreases in postoperative scenarios, particularly for small enhancing lesions and dynamically evolving cavities. By explicitly including resection cavity delineation and conducting subregion-specific evaluation, this study extends prior work. It provides a more clinically realistic assessment of automated segmentation performance in postoperative radiotherapy planning contexts.

Table 5. Summary of representative studies on glioblastoma segmentation using conventional and deep learning–based approaches. The table highlights the datasets, methodological focus, key findings, and

limitations, emphasizing the limited availability of postoperative and failure–risk–oriented evaluations compared with the comprehensive, multi-metric postoperative analysis presented in the current study.

Study	Year	Method	Dataset	Focus	Key Findings	Limitation
Menze et al (14).	2015	Benchmark dataset	BraTS	Tumor segmentation	Introduced a multi-modal glioma dataset	No DL segmentation
Bakas et al (15).	2018	BraTS benchmark	BraTS	Multi-class tumor regions	Standard evaluation metrics defined	Preoperative focus
Isensee et al (9).	2021	nnU-Net	Multi-dataset	General medical segmentation	State-of-the-art automated framework	Not postoperative specific
Myronenko et al (16).	2019	3D CNN	BraTS	Brain tumor segmentation	High Dice performance	Limited clinical interpretation
Zhou et al (17).	2022	Deep CNN	BraTS	Enhancing tumor segmentation	Improved ET detection	No resection cavity
Present study	2026	nnU-Net v2 3D fullres	BraTS	Postoperative multi-class segmentation	Comprehensive quantitative evaluation, including RC	External validation pending

Clinical Integration Considerations

Automated segmentation tools have the potential to improve efficiency and reproducibility in radiotherapy workflows. However, safe integration into clinical practice requires a clear understanding of reliability across different anatomical structures. Based on the present findings, automated segmentation may be sufficiently reliable for non-enhancing tumor regions when used as a decision-support tool under expert supervision. In contrast, enhancing tissue and resection cavity segmentation requires careful manual verification before treatment planning.

Rather than replacing clinician-driven contouring, deep learning–based segmentation should be viewed as an assistive technology that can reduce contouring time, improve consistency, and highlight regions requiring focused expert review. Failure-aware evaluation frameworks such as the one presented here may support risk-informed deployment of automated tools in clinical environments.

Limitations

Several limitations should be acknowledged. First, model development and evaluation were conducted using a single publicly available dataset. Although the BraTS dataset includes multi-institutional data and heterogeneous imaging protocols, external validation on independent clinical datasets is necessary to fully assess generalizability. Second, this study focused on segmentation accuracy and reliability rather

than direct dosimetric impact. Future investigations should evaluate how segmentation variability influences dose distribution, target coverage, and organ-at-risk sparing. Third, inter-observer variability among expert annotations was not analyzed and may provide additional context for interpreting automated segmentation performance.

The relatively high observed failure rates should be interpreted in the context of conservative failure thresholds, highly heterogeneous postoperative anatomy, and the known sensitivity of Dice metrics to small lesion volumes.

Future Directions

Future work should focus on improving segmentation reliability for enhancing tissue and resection cavities through advanced modeling strategies such as uncertainty-aware learning, multi-stage architectures, and incorporation of temporal postoperative imaging. Prospective clinical validation and evaluation of dosimetric consequences will be essential for establishing the practical utility of automated segmentation in radiotherapy planning. Integration of clinician-in-the-loop workflows and human–AI interaction studies may further facilitate safe adoption of deep learning–based contouring tools in routine practice.

Future research should focus on integrating segmentation reliability maps, uncertainty estimation, and dosimetric impact analysis to enable safe clinical deployment of deep learning–based contouring tools in radiotherapy. External multi-center validation and prospective clinical studies will be essential to confirm generalizability and workflow integration.

Conclusion

This study demonstrates that deep learning–based segmentation using nnU-Net v2 provides clinically relevant reliability for non-enhancing postoperative tumor regions and can support radiotherapy planning workflows under expert supervision. However, segmentation of enhancing tissue and resection cavities remains associated with substantial variability and a clinically meaningful risk of failure, necessitating careful verification before treatment planning. Failure-aware and clinically interpretable evaluation frameworks, such as the one presented in this study, are essential to ensure the safe and effective clinical translation of deep learning–based segmentation systems into radiotherapy practice.

Declarations

Ethics approval and consent to participate: This study was conducted as part of a research project approved by the Ethics Committee of Isfahan University of Medical Sciences (Approval Nos. IR.MUI.MED.REC.1403.498 and IR.ARI.MUI.REC.1404.122). The study used the publicly available, fully anonymized BraTS 2024 postoperative dataset. No additional patient recruitment, data collection, or

manual annotation was performed by the authors. The expert reference annotations were provided as part of the BraTS dataset. Ethical approval and informed consent for the original data collection were obtained by the original data contributors, as described in the BraTS publications.

Consent for publication: Not Applicable.

Ethics and Consent to Participate declarations: Not applicable.

Availability of data and materials: The dataset used in this study is publicly available through the Synapse platform as a curated, multi-institutional collection for postoperative brain tumor segmentation with expert-annotated postoperative tumor subregions:

<https://www.synapse.org/Synapse:syn53708249>.

Competing Interests: The authors declare no competing interests.

Funding: This study was financially supported by Isfahan University of Medical Sciences, Isfahan, Iran (Grant Nos. 3403805 and 140489).

Author contributions: Conceptualization, A.S. and D.S.-G.; Methodology, A.S., D.S.-G., and A.K.; Validation, A.S., D.S.-G., A.K., M.S., and S.H.; Investigation, A.S. and M.S.; Resources, D.S.-G.; Data Curation, A.S. and M.S.; Literature Review, M.S.; Data Interpretation, M.S.; Writing—Original Draft Preparation, A.S.; Writing—Review and Editing, A.S., D.S.-G., A.K., M.S., and S.H.; 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.

References

1. Baumert, B. G. et al. ESTRO-EANO guideline on target delineation and radiotherapy for IDH-mutant WHO CNS grade 2 and 3 diffuse glioma. *Radiother. Oncol.* **202**, 110594 (2025).
2. Cepeda, S. et al. Deep learning-based postoperative glioblastoma segmentation and extent of resection evaluation: Development, external validation, and model comparison. *Neuro-Oncology Adv.* **6**(1). (2024).
3. Huang, J. et al. Brain tumor segmentation using deep learning: high performance with minimized MRI data. *Front. Radiol.* **5**, 1616293 (2025).
4. de Verdier, M. C. et al. The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment mri. *arXiv preprint arXiv :240518368*. (2024).
5. 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.* **20** (1), 169 (2025).
6. Yoo, Y. et al. Extended nnU-Net for Brain Metastasis Detection and Segmentation in Contrast-Enhanced Magnetic Resonance Imaging With a Large Multi-Institutional Data Set. *Int. J. Radiation Oncology*Biology*Physics.* **121** (1), 241–249 (2025).

7. Isensee, F. et al. (eds) nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention; : Springer. (2024).
8. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. (eds) 3D U-Net: learning dense volumetric segmentation from sparse annotation. International conference on medical image computing and computer-assisted intervention; : Springer. (2016).
9. Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods*. **18** (2), 203–211 (2021).
10. Yadav, A. C., Kolekar, M. H. & Zope, M. K. Modified Recurrent Residual Attention U-Net model for MRI-based brain tumor segmentation. *Biomed. Signal Process. Control*. **102**, 107220 (2025).
11. Taha, A. A. & Hanbury, A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med. Imaging*. **15** (1), 29 (2015).
12. Maier-Hein, L. et al. Metrics reloaded: recommendations for image analysis validation. *Nat. Methods*. **21** (2), 195–212 (2024).
13. Luque, L. et al. Standardized evaluation of the extent of resection in glioblastoma with automated early post-operative segmentation. *Front. Radiol*. **4**, 1357341 (2024).
14. Menze, B. H. et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging*. **34** (10), 1993–2024 (2015).
15. Bakas, S. et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:181102629. (2018).
16. Myronenko, A. (ed), editor 3D MRI Brain Tumor Segmentation Using Autoencoder Regularization2019; Cham: Springer International Publishing.
17. Zhou, Z. et al. Super-resolution of brain tumor MRI images based on deep learning. *J. Appl. Clin. Med. Phys*. **23** (11), e13758 (2022).

Figures

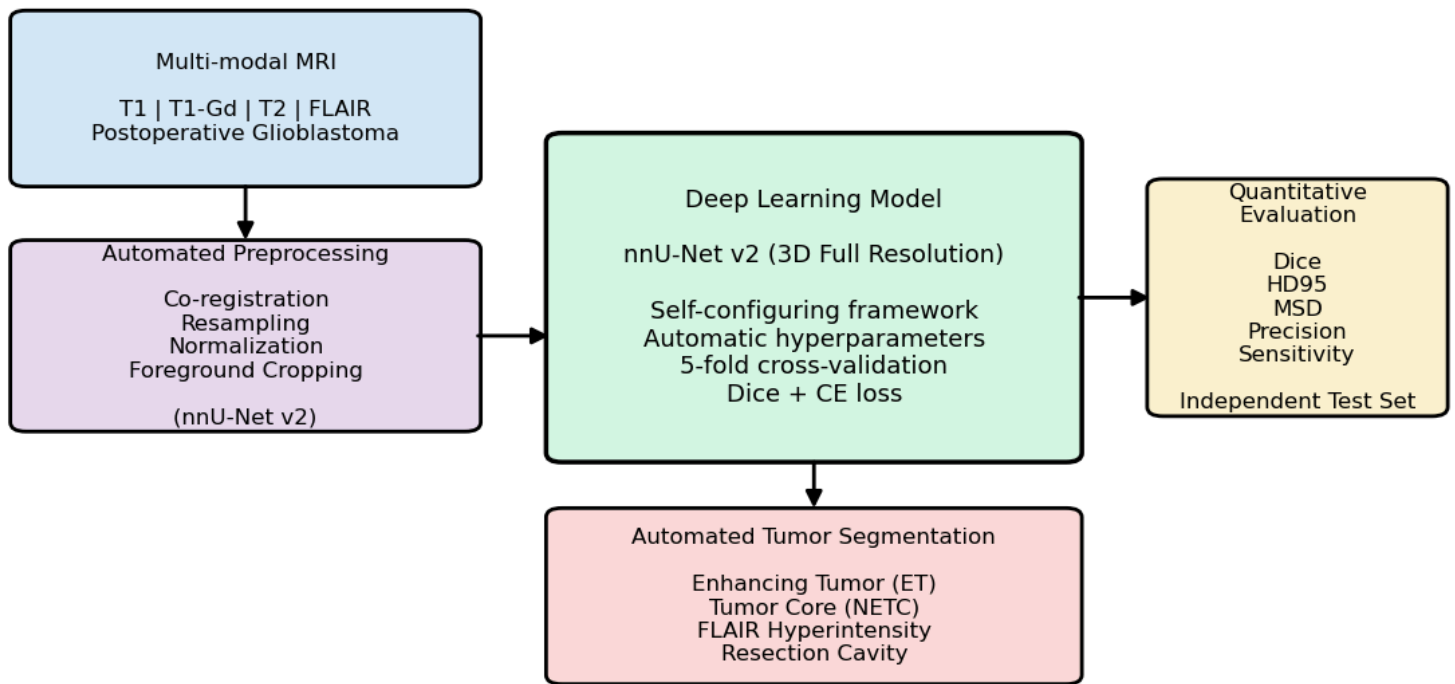


Figure 1

Overview of the proposed nnU-Net–based segmentation pipeline for automatic delineation of glioma subregions from multi-modal MRI. The workflow includes image acquisition, preprocessing, and normalization of four MRI modalities (T1, T1-Gd, T2, and FLAIR), network training using the nnU-Net framework, and automated prediction of tumor subregions, including NETC, SNFH, enhancing tumor (ET), and resection cavity (RC). Quantitative evaluation was performed using the Dice similarity coefficient, HD95, MSD, precision, and sensitivity metrics.

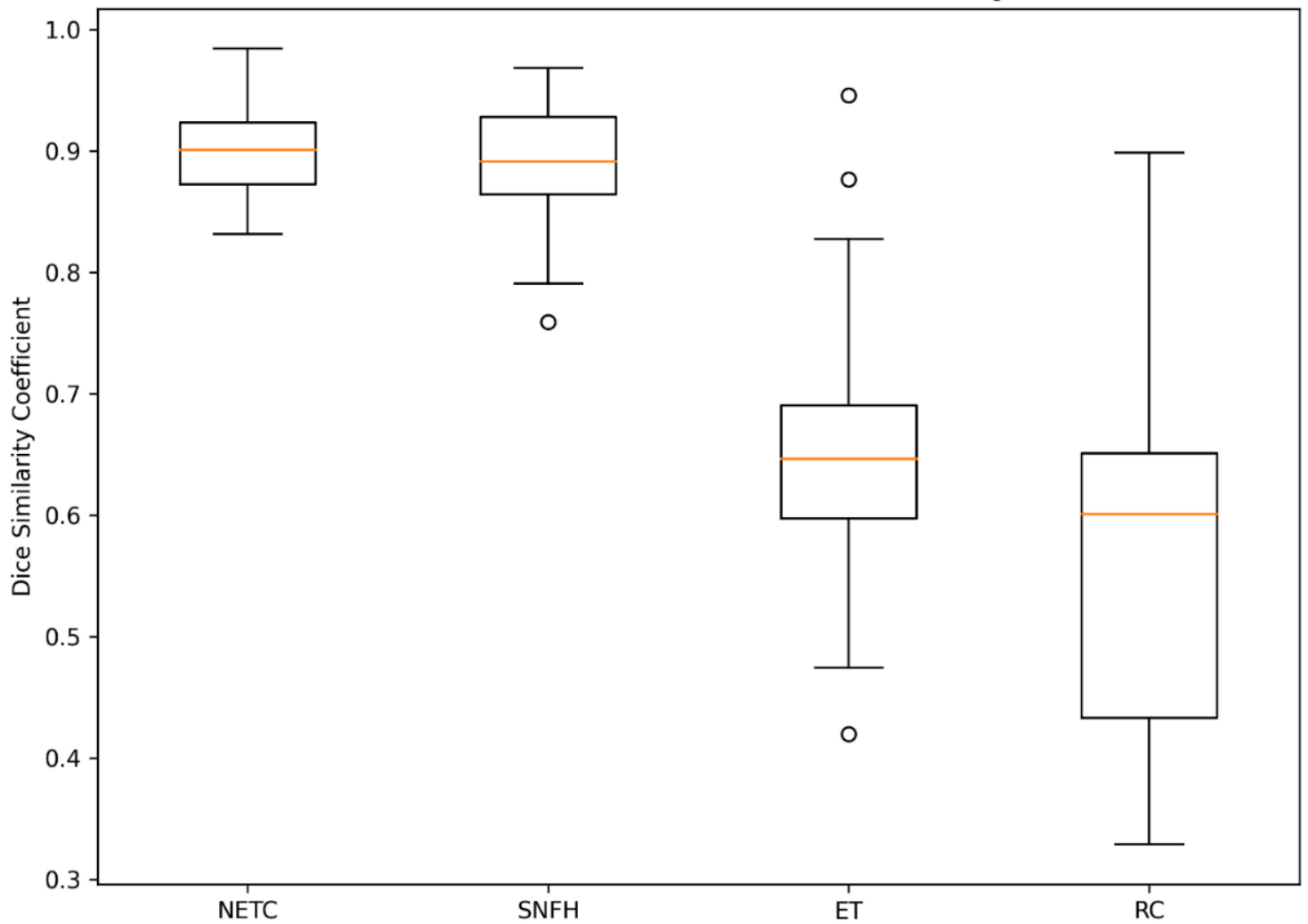


Figure 2

Boxplot distribution of Dice similarity coefficients across tumor subregions. The model achieved the highest and most consistent performance for the NETC and SNFH regions, while the enhancing tumor (ET) and resection cavity (RC) regions demonstrated greater variability and lower segmentation accuracy, highlighting ongoing challenges in complex anatomical regions.

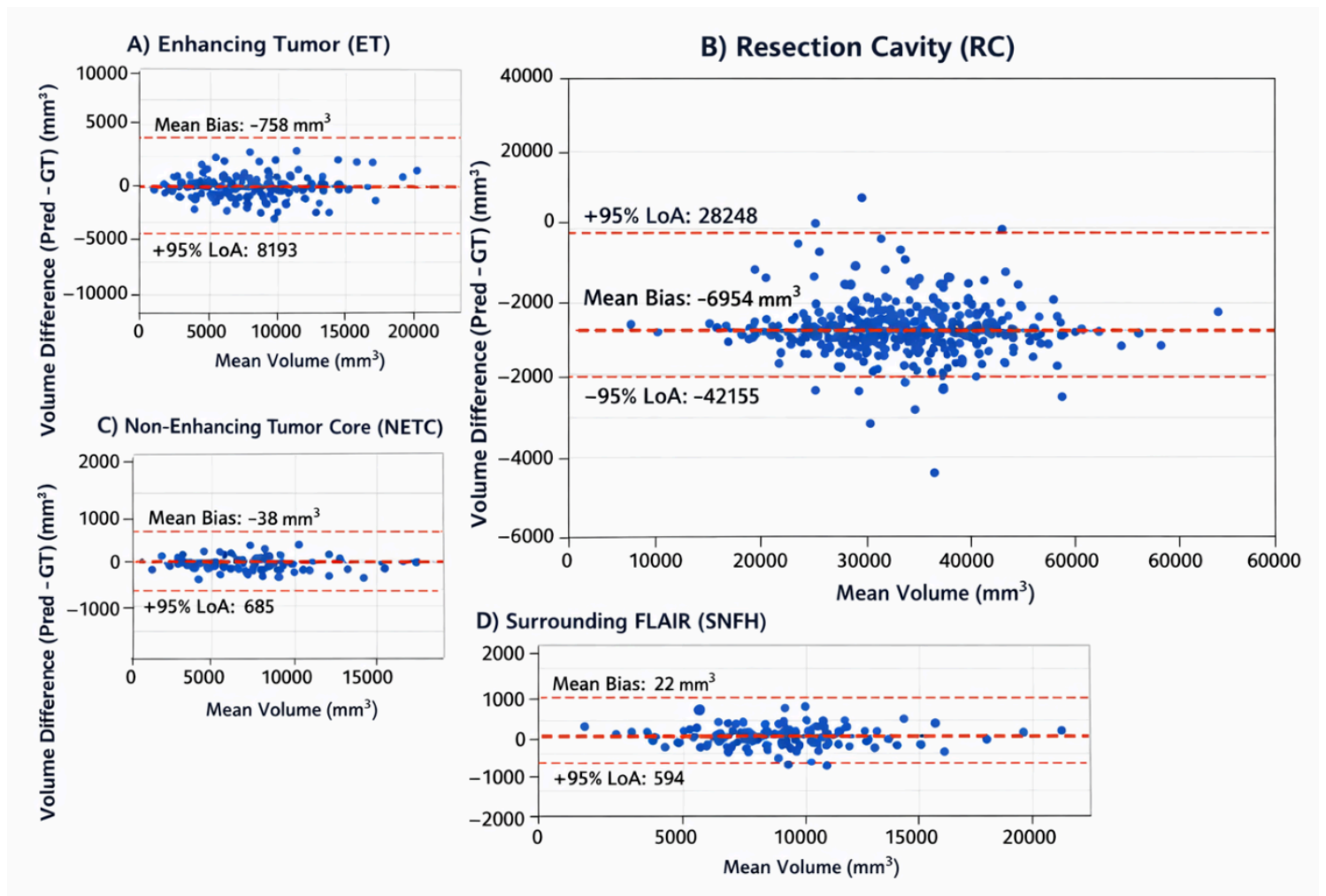


Figure 3

Bland–Altman agreement analysis across postoperative tumor subregions. Enhancing tumor (ET) and resection cavity (RC) demonstrate wide limits of agreement and systematic underestimation, whereas NETC and SNFH exhibit stable volumetric agreement.

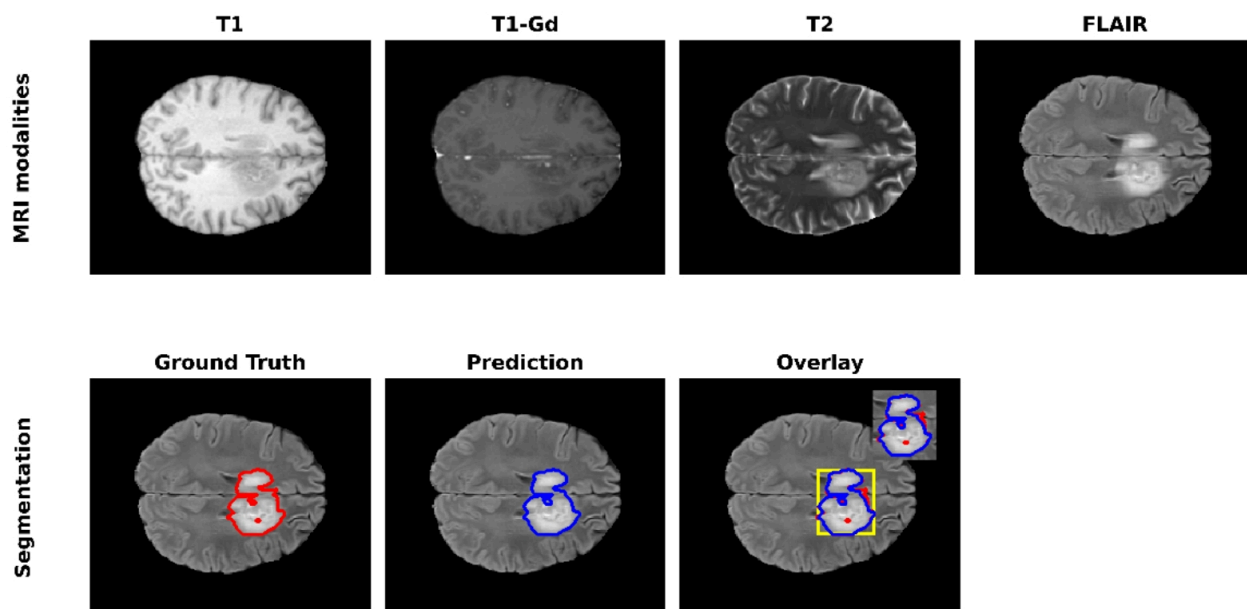


Figure 4

Representative example of automatic glioma subregion segmentation using the nnU-Net model on multi-modal MRI. Four MRI modalities (T1, T1-Gd, T2, and FLAIR) are shown alongside the corresponding expert ground-truth annotations (GT), model predictions, and overlay visualizations. The overlay demonstrates spatial agreement between manual and automated delineations across tumor subregions, with higher concordance observed in NETC and SNFH regions and greater discrepancies in enhancing tumor (ET) and resection cavity (RC).