BraTS 2024 Post-Treatment MRI Segmentation: Comprehensive Project Report

1. Summary & Objectives

1.1. Problem Description

Automated Segmentation of Post-Treatment Glioma in Multi-Modal MRI Using Deep Learning

This study addresses the significant challenge of automating the segmentation of post-treatment glioma sub-regions. In contrast to pre-treatment scenarios, the post-treatment MRI landscape is characterized by high heterogeneity, including surgically induced cavities, resection margins, scar tissue, and radiation-induced leukoencephalopathy, all of which confound standard segmentation algorithms.

1.2. Research Objectives

The primary objectives of this research are to:

1. **Develop an automated segmentation framework** utilizing the nnU-Net architecture to accurately delineate multiple tumor sub-regions in post-treatment glioma patients using multi-modal MRI scans.
2. **Address the "Resection Cavity" (RC) segmentation challenge** introduced in the BraTS 2024 dataset, distinguishing surgical voids from residual non-enhancing tumor core (NETC) and necrosis.
3. **Evaluate the efficacy of the 3D full-resolution U-Net** configuration in handling severe class imbalance and anatomical variability inherent in post-operative neuroimaging.
4. **Conduct a comprehensive performance assessment** employing voxel-wise metrics (Dice Similarity Coefficient, IoU, Precision, Recall) to ensure clinical relevance and reliability.

2. Medical Background & Annotation Protocol

A fundamental understanding of MRI physics and the specific annotation protocols utilized in BraTS 2024 is prerequisite for interpreting the model's performance characteristics.

2.1. MRI Modalities and Principles

- **T1-weighted (T1n):** The "Anatomical Map."
  - *Principle:* Sensitive to longitudinal relaxation times (T1). Adipose tissue appears hyperintense, while water and Cerebrospinal Fluid (CSF) appear hypointense.
  - *Clinical Role:* Serves as the baseline for subtraction with T1-Gd to isolate true contrast enhancement. Crucial for identifying "intrinsic T1 hyperintensity" (e.g., methemoglobin from post-operative hemorrhage), which designates a region as Non-enhancing Tumor Core (NETC) rather than active Enhancing Tissue.
- **T2-weighted (T2w):** The "Pathology Map."
  - *Principle:* Sensitive to transverse relaxation times (T2). Pathological tissues with high water content (tumor, edema) and CSF appear hyperintense.
  - *Clinical Role:* The primary modality for delineating the **Resection Cavity (RC)**, which typically presents as a fluid-filled, T2-hyperintense surgical void.
- **FLAIR (T2-FLAIR):** The "Edema Map."
  - *Principle:* T2-weighted sequence with CSF signal suppression (Inversion Recovery).
  - *Clinical Role:* The gold standard for defining **Surrounding Non-enhancing FLAIR Hyperintensity (SNFH)**. By suppressing the CSF signal, FLAIR ensures that peritumoral edema and infiltrative tumor are distinguishable from ventricles and sulci.

- **T1-Gd (T1-weighted with Gadolinium):** The "Active Tumor Map."
  - *Principle:* Paramagnetic contrast agents (Gadolinium) accumulate in regions with Blood-Brain Barrier (BBB) disruption.
  - *Clinical Role:* The definitive sequence for identifying **Enhancing Tissue (ET)**, representing active tumor proliferation.

2.2. Label Definitions (BraTS 2024)

The segmentation targets are formally defined as follows:

Label	Class Name	Color	Definition	Key MRI Features
1	**Enhancing Tissue (ET)**	Blue	Active, proliferating tumor tissue.	Hyperintense on T1-Gd relative to T1n. Strictly excludes intravascular contrast and intrinsic T1 hyperintensity.
2	**Non-enhancing Tumor Core (NETC)**	Red	Necrotic core, cysts, or non-enhancing tumor parenchyma.	Hypointense on T1-Gd; may encompass regions of intrinsic T1 hyperintensity (e.g., blood products).
3	**Surrounding FLAIR Hyperintensity (SNFH)**	Green	Edema, infiltration, gliosis, and radiation effects.	Hyperintense on FLAIR. Represents the "whole tumor" extent typically including vasogenic edema.
4	**Resection Cavity (RC)**	Yellow	Surgical cavity following tumor resection.	**Novel Class.** Fluid-filled cavity; hyperintense on T2, hypointense on T1/FLAIR.

2.3. Common Segmentation Pitfalls

Expert annotation review highlights recurring challenges for automated systems:

1. **Vascular Mimicry:** Intracranial vessels exhibit contrast enhancement (T1-Gd hyperintensity) similar to ET. *Distinction:* Vessels display linear/tubular morphology, whereas tumor tissue is typically nodular.
2. **Choroid Plexus Enhancement:** The choroid plexus lacks a blood-brain barrier and naturally enhances. *Distinction:* Anatomical localization within the ventricles allows for differentiation.
3. **Intrinsic T1 Hyperintensity:** Post-operative blood products appear bright on both pre- and post-contrast T1 images. Models relying solely on T1-Gd intensity may misclassify these as ET. *Distinction:* Comparison with T1n confirms the signal is intrinsic, classifying it as NETC or RC.
4. **Microvascular Ischemic Changes:** Elderly patients often present with white matter lesions (leukoaraiosis) that are hyperintense on FLAIR. *Distinction:* These lesions are typically symmetric, patchy, and spatially distinct from the glioma mass.

3. Dataset Characterization

3.1. Dataset Snapshot

This study utilizes the official BraTS 2024 Post-Treatment Glioma dataset, standardized to the NIfTI format.

Cohort	Cases	Modalities	Resolution	Orientation
Training	1,350	T1, T1-Gd, T2, FLAIR	1.0 mm^3 Isotropic	LAS (Left-Anterior-Superior)
Additional	271	T1, T1-Gd, T2, FLAIR	1.0 mm^3 Isotropic	LAS
Total	**1,621**	-	-	-

3.2. Intensity and Geometric Standardization

- **Geometry:** All volumes are rigid-registered to a common atlas space with dimensions of 182 × 218 × 182 voxels. This spatial normalization facilitates the learning of anatomical priors by the 3D CNN.
- **Intensity:** Substantial inter-scanner variability is observed (T1c 99th percentile values range from approximately 92 to 12,000).
    - *Implication:* Robust preprocessing, specifically percentile-based clipping (1st–99th percentile) and Z-score normalization, is mandatory to harmonize the input distribution.

3.3. Class Imbalance Analysis

The dataset demonstrates severe class imbalance, a critical factor influencing model convergence and performance:

- **Background:** >98% of total voxels.
- **NETC (Label 2):** ~67% of the tumor mass (Dominant foreground class).
- **RC (Label 4):** ~19% of the tumor mass.
- **SNFH (Label 3):** ~11% of the tumor mass.
- **ET (Label 1): ~2.3% of the tumor mass.** (Extremely rare, posing the greatest segmentation challenge).

4. Methodology

4.1. Preprocessing Pipeline

To ensure data consistency and optimal network performance, the following preprocessing steps were applied:

1. **Reorientation:** All volumes were realigned to the standard RAS (Right-Anterior-Superior) coordinate system.
2. **Resampling:** A strictly isotropic voxel spacing of 1.0 × 1.0 × 1.0 mm^3 was enforced to maintain volumetric consistency.
3. **Cropping:** Foreground cropping was performed to eliminate non-informative background regions, optimizing computational efficiency.
4. **Normalization:** Channel-wise Z-score normalization was applied to non-zero voxels to standardize intensity distributions across modalities.

4.2. Model Architecture: nnU-Net 3D Full-Resolution

This study employs the **nnU-Net framework** configured for **3D Full-Resolution** segmentation. The architecture is a dynamically configured PlainConvUNet derived from the dataset fingerprint (Dataset501_BraTSPostTx).

- **Architecture Class:** dynamic_network_architectures.architectures.unet.PlainConvUNet
- **Topological Structure:**
    - **Depth:** The network consists of **6 stages** in both the encoder (contracting path) and decoder (expanding path).

    - **Feature Maps:** The number of features per stage is hierarchically organized as [32, 64, 128, 256, 320, 320].
    - **Convolutional Blocks:** Each stage contains 2 convolutional blocks (n_conv_per_stage: [2, 2, 2, 2, 2, 2]).
    - **Kernel Size:** Uniform 3x3x3 kernels are utilized across all stages.
- **Downsampling Strategy (Strides):**
    - Stage 1: [1, 1, 1] (No downsampling)
    - Stages 2-5: [2, 2, 2] (Isotropic downsampling)
    - Stage 6: [2, 2, 1] (Anisotropic downsampling to preserve Z-axis resolution in deep layers).
- **Operational Components:**
    - **Convolution:** torch.nn.modules.conv.Conv3d
    - **Normalization:** torch.nn.modules.instancenorm.InstanceNorm3d (Affine=True, Epsilon=1e-05).
    - **Activation:** torch.nn.LeakyReLU (In-place=True).
- **Input Configuration:**
    - **Patch Size:** [128, 160, 112] voxels.
    - **Batch Size:** 2.
- **Training Strategy:**
    - **Cross-Validation:** A robust 5-fold cross-validation scheme was implemented.
    - **Loss Function:** A combination of Dice Loss and Cross-Entropy Loss was used to optimize segmentation overlap and pixel-wise classification accuracy.
    - **Deep Supervision:** Auxiliary loss layers were active during training to improve gradient flow in deeper network layers.

5. Comprehensive Performance Analysis

This section details the quantitative results obtained from the updated validation of the 3D full-resolution UNet model.

5.1. Global Performance Statistics

The aggregated metrics across the validation cohort indicate strong performance in precision, with moderate overlap scores, reflecting the challenge of small, fragmented sub-regions.

Metric	Value	Interpretation
Dice Similarity Coefficient (DSC)	**0.6459**	Indicates moderate overlap between prediction and ground truth.
Intersection over Union (IoU)	0.5690	Reflects the area of overlap divided by the area of union.
Precision	0.9129	High positive predictive value; predictions are highly reliable.
Recall	0.8725	Good sensitivity in detecting tumor presence.

5.2. Class-Specific Performance Breakdown

Performance stratification reveals significant variance across the defined tumor sub-regions.

Class	Region	Dice	IoU	Precision	Recall	Analysis
1	ET	**0.4616**	0.3946	0.8082	0.7546	**Improving.** While still the lowest, the Dice score has improved. High precision (0.80) suggests that when the model predicts ET, it is likely correct.
2	NETC	**0.8310**	0.7531	0.9247	0.8825	**Best Performer.** The necrotic core remains the most distinct and accurately segmented region.
3	SNFH	0.7167	0.6358	0.9370	0.8680	**Strong Precision.** The model is extremely precise (0.937) in identifying edema/infiltrative tissue, rarely generating false positives.
4	Resection Cavity (RC)	0.5745	0.4927	0.8314	0.8420	**Balanced.** The novel cavity class shows a balanced trade-off between precision and recall, though boundary delineation remains challenging.

5.3. Error Analysis and Pattern Discovery

Correlation analysis provides insight into the systematic behaviors of the model.

1. Sensitivity to Object Size (Volume Bias)

Class	Correlation (Vol vs Dice)	Interpretation
1 (ET)	**0.280**	**Size-Independent.** Unlike previous iterations, performance on Enhancing Tissue is now largely independent of lesion size, indicating improved robustness on small lesions.
2 (NETC)	0.401	**Positive Correlation.** The model struggles slightly more with smaller necrotic cores.
3 (SNFH)	0.390	**Positive Correlation.** Smaller regions of edema are harder to segment accurately.
4 (RC)	0.391	**Positive Correlation.** Smaller resection cavities are more prone to segmentation errors.

2. Inter-Class Dependency

- **SNFH Support (0.380):** A positive correlation exists between the volume of SNFH (Class 3) and the Dice

score of ET (Class 1). This suggests that larger regions of edema provide critical spatial context that aids the model in localizing the active enhancing tumor.
- **RC Interference (-0.312):** A negative correlation persists between the volume of the Resection Cavity (Class 4) and the Dice score of ET (Class 1). Large cavities degrade the segmentation of enhancing tissue, confirming the clinical difficulty of distinguishing post-surgical scarring (marginal enhancement) from true tumor recurrence.

3. False Positive Tendencies

- **Class 1 (ET):** Remains susceptible to noise. The average false-positive volume is approximately **17.9%** of the average object size, likely attributable to vascular mimics or blood products.
- **Class 4 (RC):** Similar noise profile, with false-positive volumes averaging **17.1%** of the object size, likely due to confusion with CSF in sulci or ventricles.

5.4. Extreme Case Analysis

To further characterize model performance, we examine the specific cases yielding the highest and lowest mean Dice scores. This analysis highlights potential failure modes and optimal operating conditions.

5.4.1. Top 20 Worst Performing Cases

These cases represent catastrophic failures or significant underperformance, often characterized by missing classes or severe under-segmentation. Note that a hyphen (-) indicates the class was not present in the ground truth.

Case ID	Mean Dice	ET Dice (C1)	NETC Dice (C2)	SNFH Dice (C3)	RC Dice (C4)
BraTS-GLI-02512-100	0.1307	0.0000	0.4435	0.0717	0.0077
BraTS-GLI-02567-100	0.2318	0.0000	0.4863	0.0529	0.3880
BraTS-GLI-02579-102	0.2335	0.0000	0.8801	0.0533	0.0007
BraTS-GLI-02618-101	0.2500	-	0.0000	0.0000	0.0000
BraTS-GLI-02510-102	0.2500	-	0.0000	0.0000	0.0000
BraTS-GLI-02640-101	0.2737	-	0.0950	0.0000	0.0000
BraTS-GLI-02577-100	0.2925	0.0000	0.9109	0.0000	0.2591
BraTS-GLI-02510-103	0.2977	0.0000	0.3353	0.8554	0.0000
BraTS-GLI-02539-100	0.3302	0.0000	0.8530	0.0162	0.4517
BraTS-GLI-02642-101	0.3579	-	0.4315	0.0000	0.0000
BraTS-GLI-02619-101	0.3621	0.0013	0.8384	0.6032	0.0056
BraTS-GLI-02479-100	0.3634	0.0000	0.9009	0.0003	0.5522
BraTS-GLI-02646-100	0.3815	0.0000	0.7316	0.4621	0.3322

BraTS-GLI-02578-100	0.3883	0.0000	0.9481	0.6036	0.0015
BraTS-GLI-02642-100	0.3966	-	0.1866	0.0073	0.3925
BraTS-GLI-02585-101	0.4190	-	0.1389	0.5370	0.0000
BraTS-GLI-02571-100	0.4222	0.0839	0.6776	0.5748	0.3527
BraTS-GLI-02568-101	0.4227	0.0000	0.8644	0.8263	0.0000
BraTS-GLI-02621-101	0.4244	-	0.1167	0.5809	0.0000
BraTS-GLI-02492-102	0.4308	-	0.6537	0.0696	0.0000

5.4.2. Top 20 Best Performing Cases

These cases demonstrate the model's upper performance bound, often achieving near-human inter-rater reliability. High scores in certain cases (e.g., BraTS-GLI-02492-104) may be partially attributed to the absence of difficult sub-regions.

Case ID	Mean Dice	ET Dice (C1)	NETC Dice (C2)	SNFH Dice (C3)	RC Dice (C4)
BraTS-GLI-02492-104	0.9866	-	0.9463	-	-
BraTS-GLI-02504-100	0.9776	0.9694	0.9687	0.9721	-
BraTS-GLI-02592-102	0.9775	-	0.9549	-	0.9553
BraTS-GLI-02639-100	0.9766	0.9641	0.9636	0.9787	-
BraTS-GLI-02417-100	0.9730	0.9659	0.9560	0.9700	-
BraTS-GLI-02412-100	0.9702	0.9428	0.9652	0.9730	-
BraTS-GLI-02574-103	0.9660	0.9372	0.9630	0.9638	-
BraTS-GLI-02408-100	0.9652	0.9023	0.9766	0.9818	-
BraTS-GLI-02406-100	0.9648	0.9588	0.9438	0.9567	-
BraTS-GLI-02405-101	0.9624	0.9572	0.9326	0.9599	-
BraTS-GLI-02574-102	0.9608	0.9182	0.9632	0.9616	-
BraTS-GLI-02490-100	0.9579	0.9396	0.9312	0.9608	-
BraTS-GLI-02489-100	0.9493	0.9578	0.9546	0.9591	0.9255
BraTS-GLI-02494-102	0.9492	0.9194	0.9763	0.9012	-
BraTS-GLI-02413-100	0.9474	0.9445	0.8746	0.9704	-

BraTS-GLI-02614-105	0.9439	-	0.9377	0.8587	0.9792
BraTS-GLI-02614-104	0.9437	-	0.9348	0.8723	0.9679
BraTS-GLI-02570-101	0.9408	0.9338	0.9491	0.9191	0.9610
BraTS-GLI-02508-104	0.9330	0.9640	0.9305	0.9311	0.9064
BraTS-GLI-02480-100	0.9311	0.8189	0.9742	0.9606	0.9707

6. Conclusion

The updated evaluation of the 3D full-resolution nnU-Net demonstrates improved efficacy in the segmentation of post-treatment glioma. The global Dice score has increased to **0.6459**, with notable stability in the detection of the necrotic core (**NETC, 83.10% Dice**) and high precision in peritumoral edema (**SNFH, 93.70% Precision**).

Key findings from this iteration include:

1. **Improved Robustness on Small Lesions:** The segmentation of Enhancing Tissue (ET) is no longer strongly correlated with lesion volume, suggesting the model has learned to detect small focal enhancements more effectively.
2. **Persistent Cavity Interference:** The distinction between the resection cavity and enhancing tumor remains the primary source of error, as indicated by the negative inter-class correlation.
3. **High Precision:** The model exhibits high precision across all classes (>0.80), indicating that it is conservative and reliable; it produces fewer false positives than false negatives.

Future Work:

- **Boundary Refinement:** Investigation of attention-based mechanisms to specifically address the ambiguous interface between the resection cavity (Class 4) and enhancing tissue (Class 1).
- **Hard Negative Mining:** Continued implementation of sampling strategies to reduce false positives in Class 1 and Class 4, which currently exhibit the highest relative noise ratios.

Appendix: Computational Configuration

Hardware Environment:

- **GPU:** 2 x NVIDIA H100 (160GB VRAM)
- **Compute Capability:** CUDA 11.x/12.x compatible