3D U-Net for Brain Tumor Segmentation: A Cross-Validation Study on the BraTS Dataset

Anirudh Narasimha Bharadwaj

Abstract

This report presents a comprehensive implementation and evaluation of a 3D U-Net model for brain tumor segmentation on the BraTS dataset. The model segments three tumor regions—Enhancing Tumor (ET), Tumor Core (TC), and Whole Tumor (WT) using a 5-fold cross-validation approach on 484 Volumetric MRI samples. Implemented in PyTorch with MONAI, the model achieves an average WT Dice score of 0.6181, with fold-wise scores ranging from 0.5824 to 0.6434. Training loss decreases steadily across epochs, but Dice scores exhibit variability, indicating potential overfitting and sensitivity to data distribution. Visualizations reveal qualitative alignment with ground truth, though HD metrics highlight spatial discrepancies. The report discusses training dynamics, metric trends, generalizability challenges, and proposes improvements such as hyperparameter tuning, advanced augmentation, and label standardization to enhance performance and robustness for clinical applications.

Contents

1 Introduction

Brain tumor segmentation is a critical task in medical imaging, enabling precise diagnosis and treatment planning. The BraTS dataset provides a benchmark for segmenting tumor regions—Enhancing Tumor (ET), Tumor Core (TC), and Whole Tumor (WT)—from multi-modal MRI scans (FLAIR, T1w, T1gd, T2w). This assignment implements a 3D U-Net model using PyTorch and MONAI to address this task, employing a 5-fold cross-validation strategy to ensure robust evaluation.

The 3D U-Net architecture is well-suited for volumetric medical image segmentation due to its encoder-decoder structure and skip connections, which preserve spatial information. This report details the methodology, including data preprocessing, model architecture, training setup, and evaluation metrics (Dice Coefficient and Hausdorff Distance). Results are analyzed through quantitative metrics, training dynamics, and qualitative visualizations across all folds. The study also explores generalizability challenges and proposes strategies to improve performance, aiming to contribute to reliable tumor segmentation in clinical settings.

2 Methodology

2.1 Dataset

The Task01_BrainTumour dataset comprises 484 samples, each containing:

- **Input**: 4-channel 3D MRI volumes (FLAIR, T1w, T1gd, T2w) with dimensions 240×240×155 voxels.
- **Labels**: 3D segmentation masks with labels 0 (background), 1 (edema), 2 (non-enhancing tumor), and 3 (enhancing tumor).

Data preprocessing includes resizing to 128×128×128 voxels and applying augmentations: rotations (range 0.3 radians), flips (probability 0.2), additive noise (probability 0.1), and contrast adjustments (gamma range 0.9–1.1). Tumor regions are defined as ET (label 3), TC (labels 2+3), and WT (labels 1+2+3).

2.2 Model Architecture

The 3D U-Net, implemented via MONAI, has the following configuration:

- **Spatial Dimensions**: 3D for volumetric data.
- **Input Channels**: 4 (one per MRI modality).
- **Output Channels**: 4 (for labels 0, 1, 2, 3).
- **Channels**: (16, 32, 64, 128, 256), defining a five-level encoder-decoder structure.
- **Strides**: (2, 2, 2, 2) for downsampling/upsampling.
- **Residual Units**: 2 per block for improved gradient flow.

Inference uses `sliding_window_inference` with a window size of 128×128×128 and a batch size of 4.

2.3 Training Setup

- **Framework**: PyTorch with MONAI, executed on Newton Cluster's Tesla H100 using 2 CUDA-enabled GPU's.
- **Cross-Validation**: 5-fold with random shuffling (seed 42).
- **Optimizer**: Adam with a CyclicLR scheduler:
  - Initial learning rate: 0.001.
  - Maximum learning rate: 0.005.
  - Step size up: 32 iterations.
- **Batch Size**: 8, with 62 workers for data loading.
- **Epochs**: 16 per fold.
- **Loss**: Dice Loss.
- **Validation**: Performed every epoch.

2.4 Evaluation Metrics

- **Dice Coefficient**: Measures overlap between predicted and ground truth masks:

  $$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|}$$

  where P and G are the predicted and ground truth masks, respectively.
- **Hausdorff Distance**: Quantifies the maximum surface distances between predicted and ground truth boundaries.

3 Results

3.1 Quantitative Results

The model was evaluated across five folds, with the best WT Dice scores reported as:

- Fold 0: 0.6134
- Fold 1: 0.5824
- Fold 2: 0.6094
- Fold 3: 0.6434
- Fold 4: 0.6420

The average best WT Dice score across folds is 0.6181.

Figure 1 shows the Dice scores for ET, TC, and WT across all epochs and folds. The median Dice scores are approximately 0.6 (ET), 0.5 (TC), and 0.6 (WT), with WT showing the highest variability (outliers at 0.2). Figure 2 presents HD scores, with medians around

10 voxels for ET and TC, and 8 voxels for WT, but with outliers reaching 40 voxels, indicating occasional large spatial discrepancies.

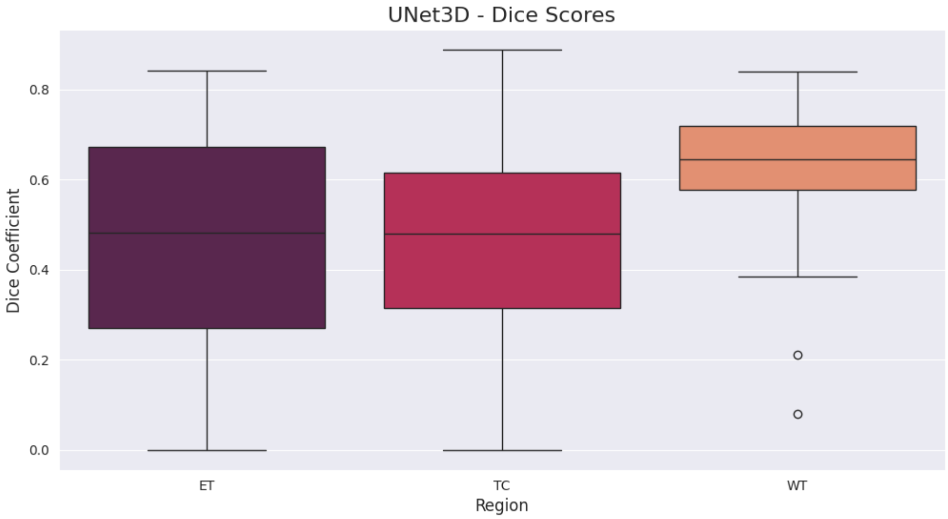

Figure 1: Dice scores for ET, TC, and WT regions across all folds and epochs.

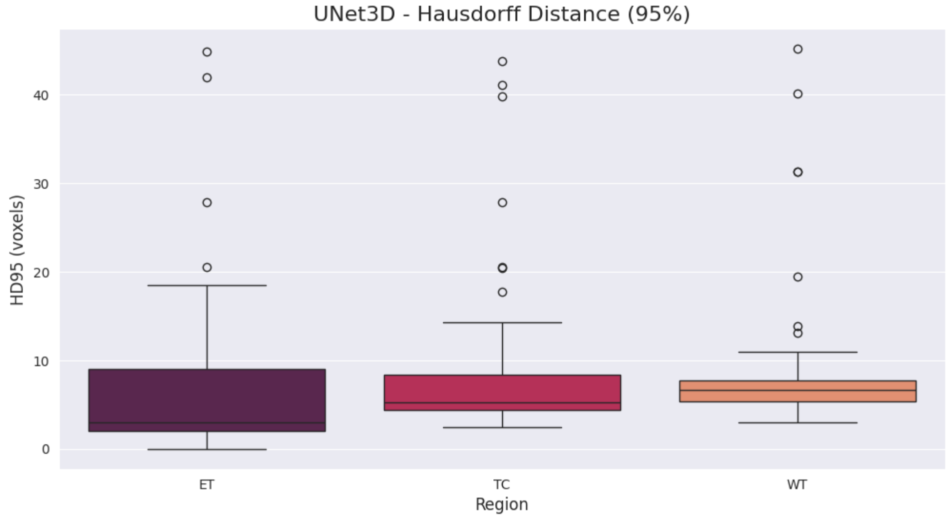

Figure 2: HD scores for ET, TC, and WT regions across all folds and epochs.

3.2 Qualitative Results

Figure 3 visualizes a sample MRI volume with its four modalities and ground truth mask, confirming data integrity and label distribution (labels 0, 1, 2, 3). Figures 4 to 8 compare predictions against ground truth for each fold, showing the FLAIR modality, ground truth, and predicted segmentation for the middle slice.

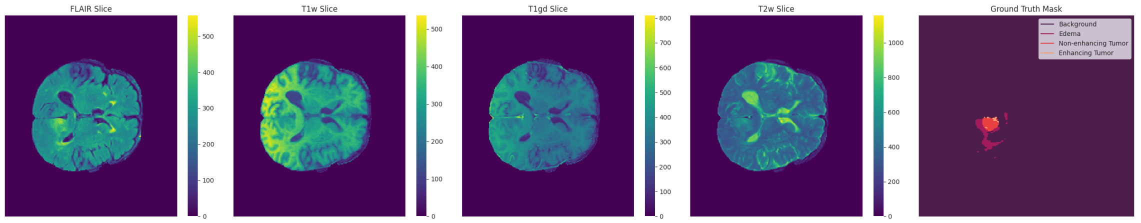

Figure 3: Sample visualization of MRI modalities (FLAIR, T1w, T1gd, T2w) and ground truth mask.

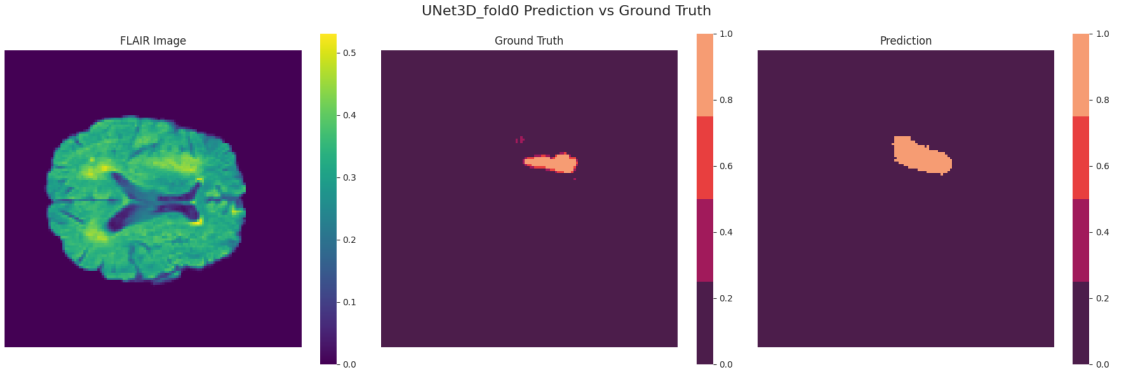

Figure 4: Prediction vs. ground truth for Fold 0

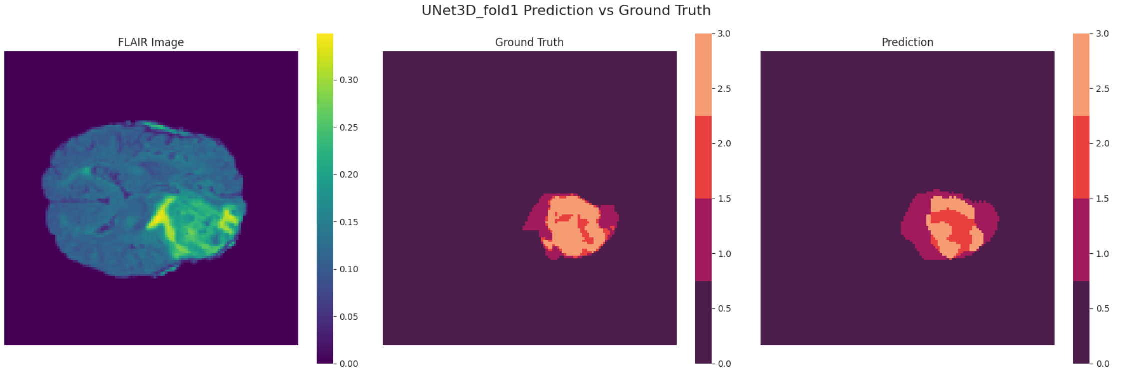

Figure 5: Prediction vs. ground truth for Fold 1

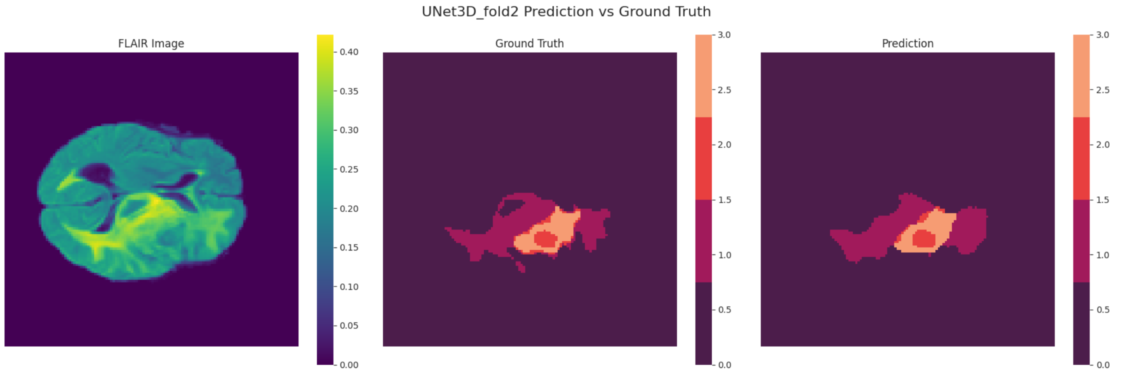

Figure 6: Prediction vs. ground truth for Fold 2

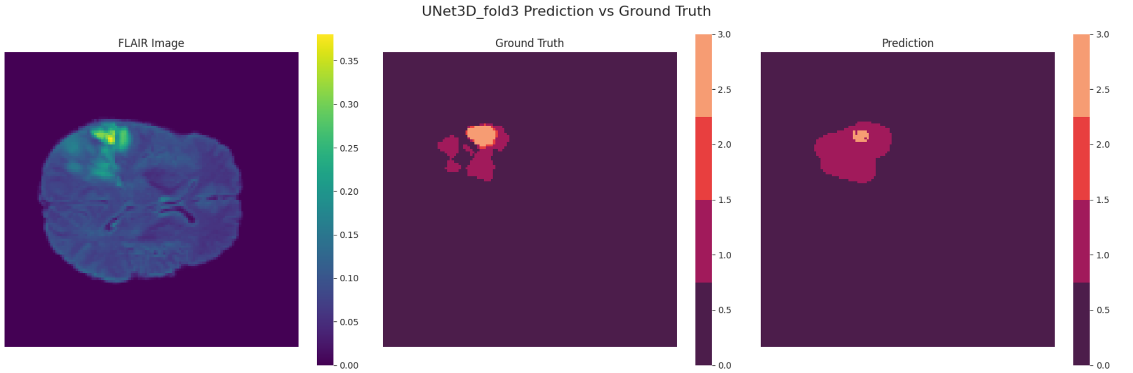

Figure 7: Prediction vs. ground truth for Fold 3

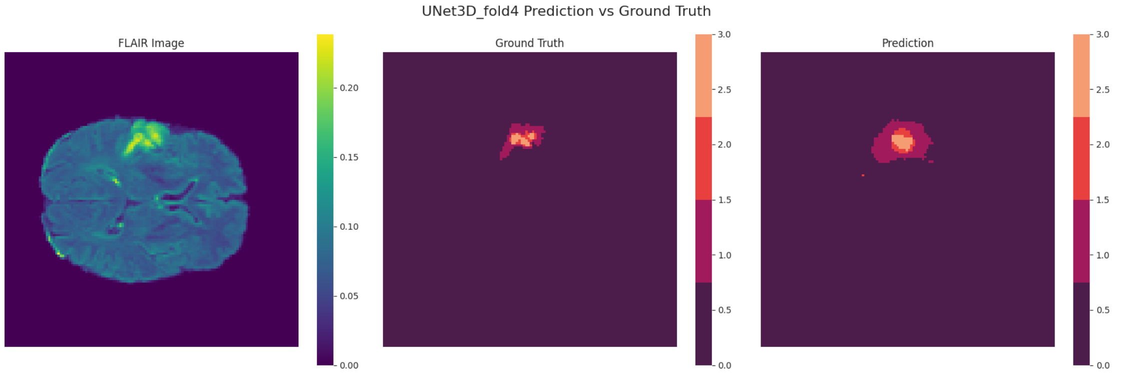

Figure 8: Prediction vs. ground truth for Fold 4

3.3 Training Dynamics

Training loss decreases consistently across folds, starting at approximately 0.77–0.79 and stabilizing around 0.35–0.37 by epoch 16. For example, in Fold 0, the loss drops from 0.7642 (epoch 1) to 0.3700 (epoch 16), a 52% reduction. This trend indicates effective optimization but suggests potential early convergence, as the loss plateaus after epoch 10 in most folds.

Validation Dice scores improve over epochs but exhibit fluctuations. For instance, in Fold 0, WT Dice increases from 0.0960 (epoch 1) to a peak of 0.6134 (epoch 11), then dips to 0.4602 (epoch 12) before recovering to 0.5929 (epoch 16). Similar patterns are observed across folds, with peaks at different epochs (e.g., Fold 3 peaks at 0.6434 in epoch 13). ET and TC Dice scores follow a similar trend but are generally lower, reflecting the challenge of segmenting smaller regions.

HD scores show high variability, with outliers up to 40 voxels, indicating that while the model captures tumor regions' general shape (per Dice), boundary precision is inconsistent. This is evident in visualizations (e.g. Figures 4 to 8), where predicted boundaries occasionally deviate from the ground truth, especially for TC.

4 Discussion

4.1 Analysis of Loss and Metrics

The steady decline in training loss across all folds demonstrates effective optimization, but the plateauing after 10 epochs suggests the model may be reaching a local minimum. The cyclic learning rate (0.001 to 0.005) aids early convergence by allowing the model to escape saddle points, but the high maximum learning rate (0.005) may contribute to the observed fluctuations in validation Dice scores, as seen in Fold 0 (e.g., WT Dice drops from 0.6134 to 0.4602 between epochs 11 and 12).

Dice scores indicate moderate performance, with WT achieving the highest average (0.6181), likely due to its larger region size reducing the impact of boundary errors. ET and TC scores are lower, reflecting the difficulty of segmenting smaller, less frequent regions (e.g., enhancing tumor). The variability in Dice scores across folds (e.g., WT ranges from 0.5824 to 0.6434) suggests sensitivity to data distribution, possibly due to class imbalance or fold-specific characteristics.

HD scores highlight a key limitation: while Dice captures overlap, it does not penalize spatial misalignment. The high HD outliers (up to 40 voxels) indicate that the model struggles with precise boundary delineation, particularly for ET and TC, where small errors in voxel prediction lead to large distance penalties. This is corroborated by visualizations, where predicted regions sometimes over- or under-segment compared to the ground truth (e.g., Fold 1, Figure 5).

4.2 Generalizability

The 5-fold cross-validation ensures robust evaluation, but the variability in fold-wise Dice scores (e.g., WT Dice from 0.5824 to 0.6434) suggests limited generalizability. This could stem from:

- **Data Distribution**: Differences in tumor size, location, or modality contrast across folds.
- **Overfitting**: The model may overfit to training folds, as evidenced by validation Dice drops after peaking (e.g., Fold 0, epoch 12).

Generalizability to unseen datasets (e.g., different scanners or patient populations) may be limited due to the model's sensitivity to training data distribution.

4.3 Improvements

To enhance performance and generalizability, the following strategies are proposed:

1. **Hyperparameter Tuning**:
   - Increase `max_epochs` (e.g., to 35) to allow further convergence, as the loss plateau suggests undertraining.
   - Reduce `max_lr` (e.g., to 0.003) to stabilize training and minimize Dice fluctuations.
   - Adjust `step_size_up` (e.g., to 50) for a slower LR cycle, preventing abrupt changes.
2. **Advanced Augmentation**:
   - Increase `rotation_range` (e.g., to 0.5) and `flip_prob` (e.g., to 0.5) to improve robustness to orientation variations.
   - Introduce elastic deformations to simulate anatomical variability.
3. **Model Architecture**:
   - Increase channel depth (e.g., to (32, 64, 128, 256, 512)) to capture more complex features.
   - Add dropout layers (e.g., 0.3 probability) and weight_deacy to reduce chances of overfitting.
4. **Loss Function**:
   - Combine Dice loss with Cross-Entropy to directly optimize for overlap, addressing the high HD values.
   - Apply class weighting to handle imbalance (e.g., higher weights for ET and TC).

5 Conclusion

The 3D U-Net model achieves an average WT Dice score of 0.6181 across 5 folds, with fold-wise scores ranging from 0.5824 to 0.6434, demonstrating moderate performance in brain tumor segmentation. Training loss decreases steadily, but validation Dice fluctuations and high HD outliers indicate challenges in stability and boundary precision. Qualitative visualizations show reasonable alignment with ground truth, though improvements are needed for smaller regions (ET, TC). Generalizability is limited by data distribution sensitivity. Future work should focus on hyperparameter optimization and advanced augmentatio to enhance clinical applicability.