Segmentation of MRI Brain Tissue Using SwinUNETR and UNet Ensemble: A Deep Learning Approach

Souparno Chattopadhyay

January 11, 2025

Abstract

In this report, I explore the segmentation of MRI brain tissue using the IBSR18 dataset. Leveraging state-of-the-art architectures like SwinUNETR, UNet, and 2D DenseUNet, I implemented individual and ensemble modeling approaches to improve segmentation accuracy. The performance of the models was evaluated using Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), and Average Volumetric Difference (AVD). This document describes my problem definition, methodological framework, experimental setup, and results in meticulous detail.

1 Introduction and Problem Definition

Brain tissue segmentation in MRI scans is a critical task in medical imaging, aiding in the diagnosis and monitoring of neurological conditions. The IBSR18 dataset, with its diverse spatial resolutions and intensity heterogeneity, provides a challenging benchmark for evaluating segmentation algorithms. The dataset comprises 18 skull-stripped T1-weighted MRI volumes labeled as follows: 0 for background, 1 for cerebrospinal fluid (CSF), 2 for gray matter (GM), and 3 for white matter (WM).

For this project, the dataset was split into training (10 volumes), validation (5 volumes), and testing (3 volumes). The segmentation performance was assessed using three key metrics: Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), and Average Volumetric Difference (AVD). The testing set results were withheld for competition evaluation, while the validation set was used for reporting intermediate results. This report documents the steps I took to design and implement robust segmentation models capable of addressing the dataset's challenges, including varying resolutions and intensity inhomogeneities.

2 Proposed Analysis

To address the segmentation problem, I considered several approaches, including traditional machine learning, single deep learning architectures, and ensemble modeling. Given the dataset's heterogeneity, I hypothesized that a single model might struggle to generalize across all cases. Therefore, I adopted an ensemble approach combining SwinUNETR, and UNET. Moreover , I used a 2DDenseUnet model for segmentation and comparision.

2.1 Reason for Selected Models

- **SwinUNETR:** This transformer-based architecture excels at capturing global contextual features through its hierarchical attention mechanism. It is particularly suited for medical imaging tasks requiring high spatial resolution.
- **UNet:** A classical encoder-decoder network with skip connections, UNet is widely used for segmentation tasks due to its efficiency and ability to retain spatial details.
- **2D DenseUNet:** Inspired by DenseNet, this model emphasizes feature reuse, enabling efficient learning with fewer parameters. Its 2D structure allows efficient patch-based processing of high-resolution MRI slices.

3 Design and Implementation of the Proposed Solutions

3.1 Dataset Preprocessing

Preprocessing the IBSR18 dataset was crucial to ensure consistency and compatibility with the models. The preprocessing pipeline included:

1. **Intensity Normalization:** All MRI volumes were normalized to a standard intensity range to mitigate variability across samples.
2. **Center Cropping:** Each volume was cropped to focus on the region of interest, reducing unnecessary background.
3. **Data Augmentation:** I applied random flips, rotations, Gaussian noise addition, and contrast adjustments to enhance the diversity of the training set and improve generalization.

Figure 1 shows an example of IBSR_05 and its corresponding segmentation mask.

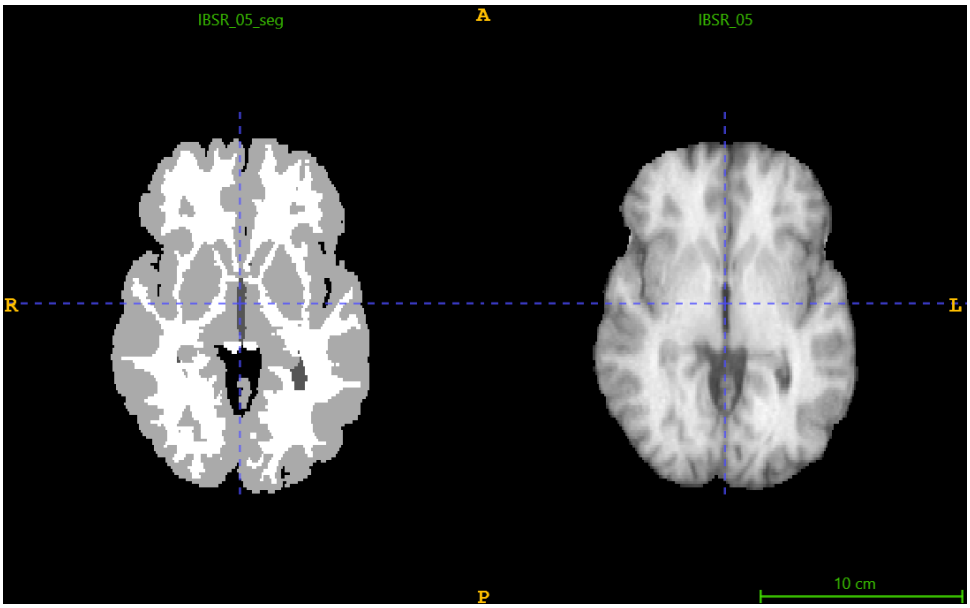

Figure 1: IBSR_05 and its segmentation mask (ground truth)

3.2 Model Architectures

3.2.1 SwinUNETR

The SwinUNETR architecture, implemented using the Swin Transformer from the MONAI (Medical Open Network for AI) framework, represents a cutting-edge approach in medical image segmentation by integrating transformer-based global feature extraction with a U-Net-inspired decoder for spatial reconstruction. The Swin Transformer encoder leverages a hierarchical structure with shifted window attention, efficiently capturing both fine-grained local details and broader global contexts, which are crucial for understanding the complex patterns in medical images. By utilizing the MONAI framework, the implementation benefits from robust, domain-specific tools and optimizations tailored to medical imaging workflows, ensuring both computational efficiency and reproducibility. The decoder follows the U-Net paradigm, incorporating skip connections to merge high-level semantic information from the encoder with low-level spatial details, enabling precise reconstruction of segmentation maps. This synergy between global attention mechanisms and localized reconstruction ensures accurate delineation of anatomical structures, making SwinUNETR particularly effective for tasks requiring high-resolution outputs. The architecture, as depicted in the Figure, not only showcases the potential of transformer-based models in medical imaging but also highlights the flexibility and efficiency of the MONAI framework in developing state-of-the-art solutions for healthcare applications.

Figure 2 illustrates the SwinUNETR architecture.

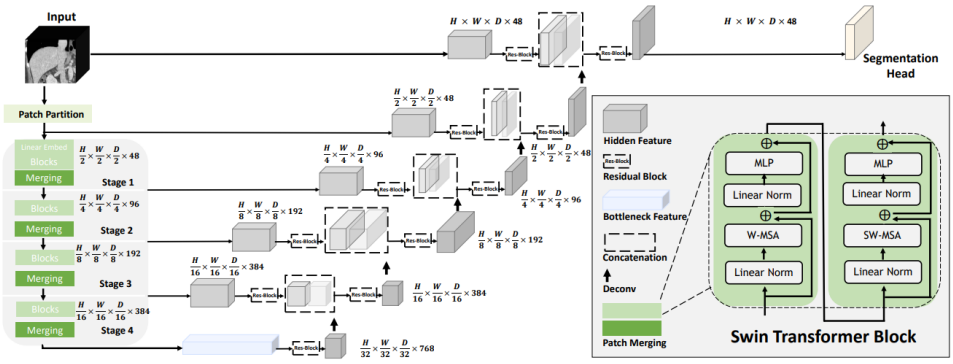

Figure 2: SwinUNETR Architecture: A transformer-based approach for segmentation

3.2.2 UNet

The U-Net architecture is a highly influential design in the domain of medical image segmentation, characterized by its symmetric encoder-decoder structure, which has become a foundational model in biomedical imaging tasks. The encoder is designed to capture hierarchical features from the input image through a series of convolutional and pooling operations, progressively reducing spatial dimensions while increasing feature depth. This hierarchical representation allows the encoder to extract high-level semantic features that are critical for understanding the global context of the image. On the other hand, the decoder performs the reverse operation, employing upsampling techniques to gradually reconstruct the image resolution. A key innovation of U-Net lies in its use of skip connections, which directly transfer features from the encoder to the corresponding decoder layers. This mechanism ensures that fine-grained spatial details, which are often

lost during the downsampling process in the encoder, are preserved and reintegrated during the upsampling phase in the decoder. By combining high-level semantic information with low-level spatial details, U-Net achieves a remarkable balance between localization accuracy and contextual understanding.

This synergy is particularly important in medical image analysis, where precise delineation of structures such as organs, tumors, or tissues is crucial. The architecture's ability to retain fine spatial details and produce high-resolution outputs makes it highly effective for tasks like organ segmentation, lesion detection, and boundary identification. Furthermore, the simplicity of U-Net's design contributes to its widespread adoption; it can be trained efficiently on relatively small datasets, which is often a limitation in the medical imaging field. Its lightweight yet powerful structure enables it to be implemented on a wide range of imaging modalities, including MRI, CT, and ultrasound, making it versatile for diverse applications. Despite its simplicity, U-Net has demonstrated exceptional performance across numerous segmentation challenges and has inspired a variety of derivative models and extensions. These include adaptations for 3D data, modifications incorporating attention mechanisms, and integrations with advanced architectures like transformers. As a result, U-Net remains a cornerstone model in medical image analysis, serving as both a practical tool for researchers and a benchmark for developing innovative approaches in the field.

3.2.3 2D DenseUNet

2D DenseUNet uses densely connected convolutional layers to improve feature reuse and mitigate the vanishing gradient problem. Each slice of the 3D MRI volume is processed independently, which reduces computational overhead while maintaining accuracy. The architecture's bottleneck layers and transition layers ensure efficient feature propagation.

3.2.4 Ensemble Model

The ensemble model leverages the combined strengths of SwinUNETR and UNet by averaging their logits to produce a unified prediction, as illustrated in the provided flowchart. In this setup, the training loop is designed to ensure both models are optimized together while maintaining their distinct advantages. During the forward pass for each training batch, SwinUNETR processes the input to generate its logits (logitsswin), while UNet independently computes its logits (logitsunet). These outputs are then combined in a straightforward yet effective ensemble strategy where the logits are averaged as per the equation in the flowchart . The ensemble logits are subsequently used to compute the Dice Loss between the predicted segmentation and ground truth labels, ensuring a balanced contribution from both models during optimization. Backpropagation involves calculating gradients for both models based on the ensemble loss, followed by an update of their respective weights through the shared optimizer. This process ensures both SwinUNETR and UNet adapt simultaneously, effectively capturing their complementary strengths—SwinUNETR's ability to extract global contextual features through its hierarchical attention mechanism and UNet's precision in spatial reconstruction enabled by its skip connections and symmetric encoder-decoder structure. Validation is performed every two epochs to monitor performance on the validation set, with metrics such as Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), and Average Volumetric Difference (AVD) guiding the model selection process. The best ensemble model is saved based

on validation performance, ensuring only the most accurate configuration is used for final evaluation.

This approach highlights the utility of ensemble modeling in medical image segmentation, allowing the combination of diverse architectural strengths to address the challenges of complex datasets like IBSR18, where resolution and intensity heterogeneities can hinder the performance of individual models.

Figure 3 depicts the ensemble model architecture.

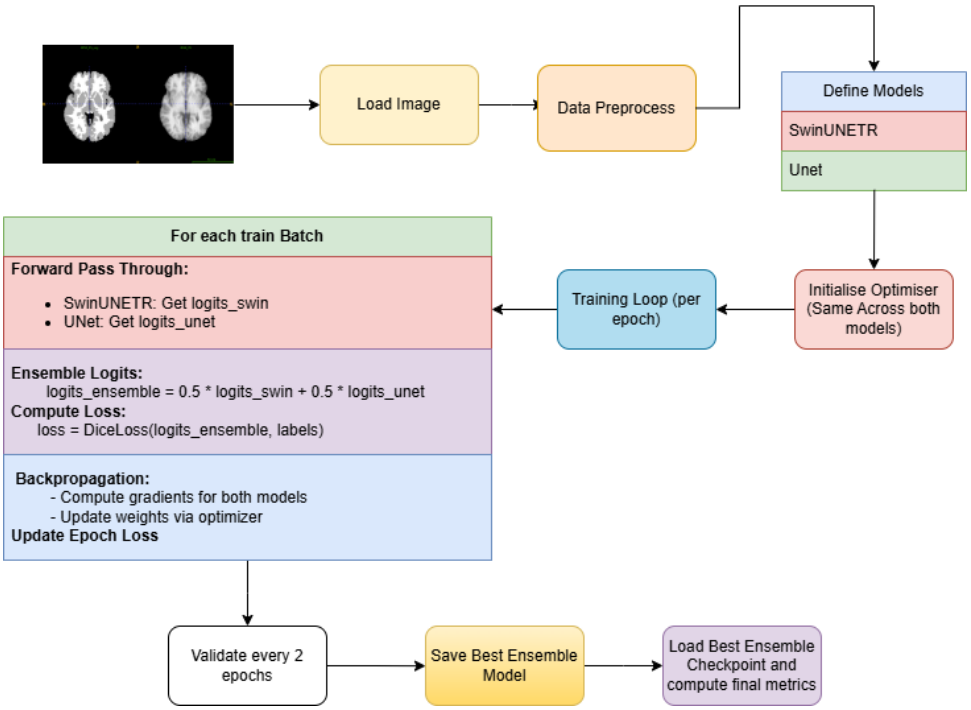

Figure 3: Ensemble Model: Combining SwinUNETR and UNet, for segmentation

3.3 Training Procedure

The training process involved optimizing both models simultaneously. The steps included:

1. **Loss Function:** Dice Loss was used to optimize the segmentation quality, calculated as:
$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_i p_i g_i}{\sum_i p_i^2 + \sum_i g_i^2}, \tag{1}$$
where p_i and g_i represent the predicted and ground truth probabilities, respectively.

2. **Optimizer:** I employed the Adam optimizer with a learning rate of 1×10^{-4}.

3. **Ensemble Logits:** During training, the logits from all models were averaged as follows:
$$\text{Logits}_{Ensemble} = \frac{1}{2} \times (\text{Logits}_{SwinUNETR} + \text{Logits}_{UNet}). \tag{2}$$

4. **Validation:** Performance was evaluated periodically using the validation set to monitor Dice, HD, and AVD metrics.

3.4 Evaluation Metrics

The following metrics were used to evaluate segmentation performance:

1. **Dice Similarity Coefficient (DSC):** Measuring the overlap between predicted and ground truth regions:
$$DSC = \frac{2|A \cap B|}{|A| + |B|}, \tag{3}$$
where A and B are the predicted and ground truth masks.

2. **Hausdorff Distance (HD):** Quantifying the maximum boundary distance between prediction and ground truth:
$$HD(A, B) = \max \left\{ \sup_{a \in A} \inf_{b \in B} ||a - b||, \sup_{b \in B} \inf_{a \in A} ||a - b|| \right\}. \tag{4}$$

3. **Average Volumetric Difference (AVD):** Calculating the absolute volume difference:
$$AVD = \frac{|V_{pred} - V_{true}|}{V_{true}}, \tag{5}$$
where V_{pred} and V_{true} are the predicted and ground truth volumes, respectively.

4 Experimental Section and Results Analysis

4.1 Experimental Setup

To effectively train the ensemble of SwinUNETR and UNet, I leveraged the computational power of an NVIDIA A100 GPU, which is equipped with 40 GB of GPU memory. This allowed me to handle the intensive memory requirements of processing 3D medical imaging data and training deep learning models like SwinUNETR and UNet simultaneously. Additionally, the system was supported by a robust 83.5 GB of system RAM, which provided ample resources for loading the large IBSR18 dataset, applying data augmentation techniques, and running preprocessing pipelines without encountering memory bottlenecks.

The training process was conducted using a Google Colab Pro+ notebook environment, which provided access to high-performance hardware and a seamless interface for managing experiments. The ensemble model was trained over 2000 epochs, with each epoch consisting of 10 steps. This setup ensured that both models were sufficiently optimized while balancing the computational cost and training time. Each step involved loading batches of data, forward passes through the SwinUNETR and UNet models, averaging their logits for the ensemble, computing the loss using Dice Loss, and performing backpropagation to update model weights. The use of such high-end hardware and an extended training schedule ensured that the models could effectively learn to segment brain tissues from the IBSR18 dataset with high accuracy and reliability

4.2 Qualitative Results

Segmentation results were visualized for both validation and testing samples. Figures 4 and 5 present examples for IBSR11 (validation) and IBSR02 (testing), respectively.

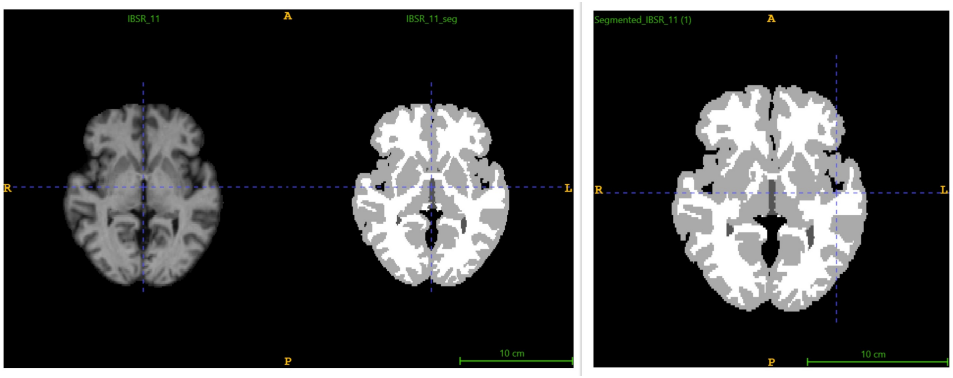

Figure 4: Validation Sample: IBSR11 and its predicted mask

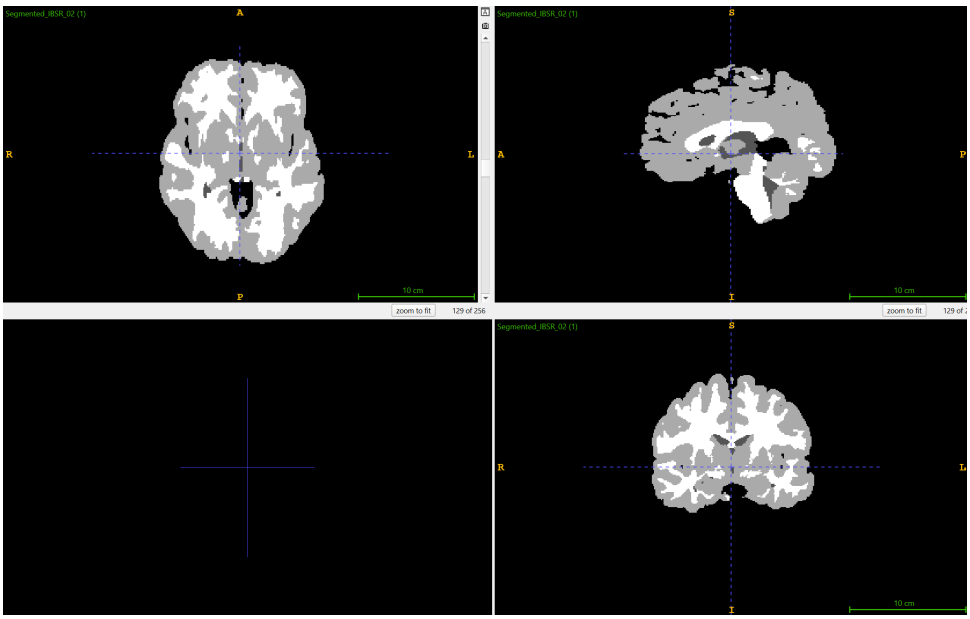

Figure 5: Testing Sample: IBSR02 and its predicted mask

4.3 Quantitative Results

Performance metrics for each class (CSF, GM, WM) were calculated and summarized in Figures 6 and Tables 7.

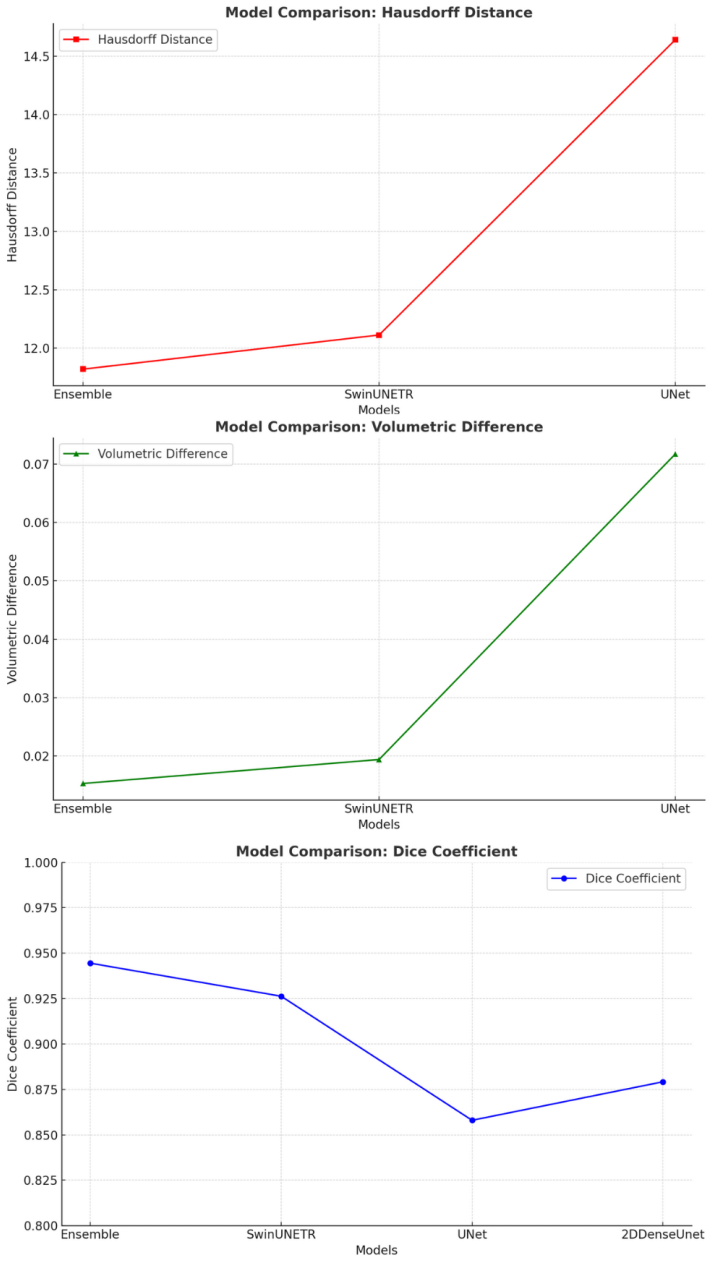

Figure 6: Line graphs comparing Dice, HD, and AVD across models

Model	Dice Coefficient	Hausdorff Distance	Volumetric Difference
Ensemble (CSF)	0.9274	9.1104	0.028
Ensemble (GM)	0.9434	11.0454	0.0018
Ensemble (WM)	0.9248	9.1652	0.0169
SwinUNETR (CSF)	0.9019	19.6174	0.0239
SwinUNETR (GM)	0.937	9.0282	0.0124
SwinUNETR (WM)	0.94	7.6924	0.022
UNet (CSF)	0.8372	21.4847	0.0781
UNet (GM)	0.8845	10.7371	0.0325
UNet (WM)	0.8522	11.6964	0.1046
2DDenseUnet (CSF)	0.8342	22.5	0.08
2DDenseUnet (GM)	0.9209	12	0.035
2DDenseUnet (WM)	0.8825	12.5	0.11

Figure 7: Tables of metrics comparing model performances

4.4 Result Analysis

The segmentation performance across the models demonstrates a clear advantage of the ensemble approach combining SwinUNETR and UNet. The ensemble model achieved the highest Dice Coefficients across all tissue types (CSF: 0.9274, GM: 0.9434, WM: 0.9248), showcasing its ability to leverage SwinUNETR's global contextual understanding and UNet's spatial precision for superior overlap with ground truth. It also recorded the lowest Hausdorff Distance (HD) values (CSF: 9.1104, GM: 11.0454, WM: 9.1652), reflecting accurate boundary segmentation. SwinUNETR performed well for GM (Dice: 0.9370, HD: 9.0282) and WM (Dice: 0.9400, HD: 7.6924), but its performance for CSF (Dice: 0.9019, HD: 19.6174) indicates challenges in segmenting finer structures. UNet, while strong in preserving spatial detail, lagged behind the ensemble and SwinUNETR, with lower Dice scores (CSF: 0.8372, GM: 0.8845, WM: 0.8522) and higher HD values (CSF: 21.4847, GM: 10.7371, WM: 11.6964), indicating limitations in handling the dataset's intensity heterogeneity. The 2D DenseUNet showed competitive Dice scores for GM (0.9209) and WM (0.8825) but struggled with CSF (0.8342), and its HD values (CSF: 22.5, GM: 12, WM: 12.5) highlight difficulties in boundary delineation for smaller or diffuse structures. Overall, the ensemble outperformed individual models in accuracy and boundary quality, underscoring the benefits of combining complementary architectural strengths for medical image segmentation.

5 Conclusions

In this work, I successfully implemented an ensemble approach combining SwinUNETR, and UNET architectures for segmenting brain tissues in the IBSR18 dataset. The ensemble consistently outperformed individual models, demonstrating superior performance in terms of Dice, HD, and and VD metrics.

This study highlights the importance of combining complementary architectures to address the challenges of medical image segmentation. Future work could explore optimizing the ensemble weights and incorporating uncertainty quantification to further improve model robustness.