Comprehensive Analysis of Brain Tissue Segmentation Using Atlas-Based and Deep Learning Models

Quang-Huy Tran
University of Girona
Medical Imaging and Applications

Sumeet Dash
University of Girona
Medical Imaging and Applications

***Abstract*—Accurate brain segmentation from MRI images is a fundamental task for medical diagnostics and neuroimaging research. This study evaluates and compares traditional atlas-based methods with advanced deep-learning approaches for segmenting key brain tissues, including white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF). Using the IBSR18 dataset, comprehensive preprocessing steps such as intensity normalization, Z-score standardization, and weighted sampling were employed to address class imbalance and enhance segmentation accuracy.**

Deep learning models, including 2D U-Net, 3D U-Net, and hybrid patch-based architectures, were trained using different loss functions such as Dice Focal Loss and Dice Cross-Entropy Loss, and their performance was compared to atlas-based predictions. Key metrics—Dice Score, Hausdorff Distance (HD), and Average Volumetric Difference (AVD)—were used to evaluate segmentation quality. The ensemble approach combining axial, coronal, and sagittal views yielded the highest Dice Score (0.9430), the lowest HD (7.8276), and a competitive AVD (0.0130), outperforming both individual slice-based and atlas-based methods.

The results highlight that deep learning methods significantly outperform traditional atlas-based segmentation, especially in capturing intricate boundaries and reducing volume discrepancies. The findings demonstrate the potential of combining preprocessing strategies with modern architectures to develop robust, efficient, and accurate solutions for brain tissue segmentation.

***Index Terms*—Brain MRI Segmentation, Multi-Atlas Method, Deep Learning, 3D U-Net, Patch-Based Segmentation, Slice-Based Segmentation, Medical Image Analysis**

I. Introduction

Brain segmentation is a critical task in medical image analysis, serving as a foundation for various applications in neuroimaging research and clinical diagnostics. Accurate segmentation of brain regions, such as white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF), is essential for understanding structural changes, diagnosing neurological disorders, and evaluating treatment outcomes. This project explores advanced segmentation techniques applied to brain MRI scans, leveraging both traditional and modern approaches.

The study combines the strengths of multi-atlas-based segmentation, a well-established traditional technique, with deep learning models. Multi-atlas methods rely on registering multiple labeled atlases to a target image, followed by label fusion, which is computationally intensive but robust. Complementing this, we utilize deep learning architectures, including 3D U-Net, to enhance segmentation accuracy and efficiency. Patch-based and slice-based strategies are also employed to manage memory requirements and focus on regions of interest, particularly in MRI datasets.

II. Datasets

The dataset utilized in this study comprises 18 T1-weighted (T1-w) MRI scans of healthy subjects sourced from the Internet Brain Segmentation Repository (IBSR) [1]. This dataset, referred to as IBSR18, was made available by the Center for Morphometric Analysis at Massachusetts General Hospital. The scans feature a voxel resolution of $256 \times 128 \times 256$ and a slice thickness of 1.5 mm. They have undergone preprocessing, including bias field correction using the Autoseg routines provided by the Center for Morphometric Analysis.

IBSR18 includes manually annotated ground truth segmentations for white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF) [2]. Notably, the dataset was acquired across three different laboratories, ensuring variability in imaging conditions.

For this study, the dataset was divided into three subsets (see Fig. 1):

- **Training Set:** Cases 1, 3, 4, 5, 6, 7, 8, 9, 16, and 18
- **Validation Set:** Cases 11, 12, 13, 14, and 17
- **Test Set:** Cases 2, 10, and 15

Ground truth segmentations are provided for the training and validation sets to facilitate model development and optimization. However, the ground truth for the test set is withheld, as these cases are reserved for independent evaluation of the proposed approach. This split ensures a robust assessment of the model's generalization performance.

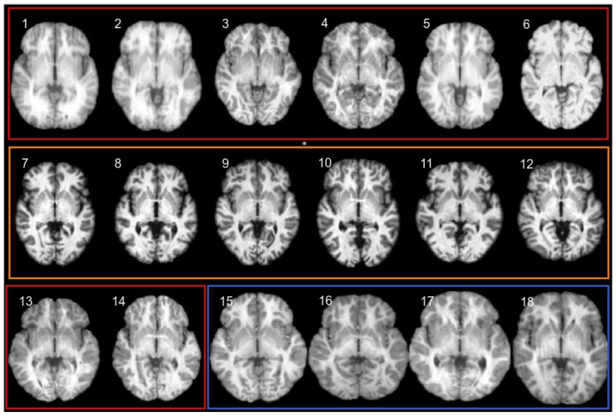


Fig. 1: IBSR18 Dataset.

III. Preprocessing

In this study, we perform basic preprocessing steps tailored to the requirements of both traditional and deep learning-based segmentation methods, refer to section IV. Preprocessing is an essential step to standardize the input images, reduce variability, and enhance the robustness of the segmentation algorithms.

A. *Preprocessing for Traditional Methods*

For traditional methods, we apply the following steps:

1) **Bias Field Correction:** The images are corrected for intensity inhomogeneities using the bias field correction tool in *SPM* (Statistical Parametric Mapping) [3]. This step ensures uniform intensities across the brain regions. Notably, the input images are already skull-stripped, so the correction focuses only on the brain tissue.
2) **Min-Max Scaling:** To normalize the intensity range, we apply min-max scaling to the voxel intensities, excluding the background voxels and voxels above 99.99 percentile (empirical observation, these voxels are then clipped to 255). This normalization scales the intensities to the range [0, 255], facilitating consistency across images.

B. *Preprocessing for Deep Learning Methods*

For deep learning-based methods, a simpler preprocessing step is employed:

1) **Normalization to [0, 1]**: In addition to z-score normalization, the voxel intensities are rescaled to the range [0, 1]. This ensures that the data lies within a consistent range, which can further aid in stabilizing the training process and improving convergence for deep learning models.
2) **Z-Score Normalization:** The voxel intensities are normalized using z-score normalization, which involves subtracting the mean intensity and dividing by the standard deviation. This standardization ensures that the input data has a mean of 0 and a standard deviation of 1, which is beneficial for training deep learning models.

By applying these preprocessing steps, we aim to provide standardized and optimized inputs for the respective segmentation approaches, ensuring that the methods operate effectively on consistent data.

IV. Methodology

A. *Traditional Approach*

Multi-atlas-based segmentation has emerged as a robust approach due to its ability to incorporate anatomical variations across a population.

1) ***Registration:*** The foundation of multi-atlas-based segmentation lies in accurate image registration. In this framework, non-rigid registration is performed to align multiple atlas images to a target (test) image. The registration process uses elastix, a widely-used tool for medical image registration, to apply a B-Spline-based transformation. By minimizing mutual information between the fixed (test) and moving (atlas) images, the registration ensures spatial correspondence between anatomical structures across datasets. We use an existing parameter file from the elastix model zoo, Parameter0009 [4], which includes an affine transform followed by a non-rigid B-Spline transform. This step generates deformation fields that map atlas images and their corresponding labels to the test image space.

The transformed labels obtained from this process form the input to the label fusion stage. To improve registration quality, we optimized the registration parameters, such as grid spacing, optimizer settings, and multi-resolution levels, ensuring better alignment even in regions with significant anatomical variability.

$$S(x) = \underset{c}{\operatorname{argmax}} \sum_{i=1}^{P} w_i(x) \cdot f\left(\pi_i^L(\hat{\tau}_i(x)), c\right) \tag{1}$$

$$f\left(\pi_i^L(\hat{\tau}_i(x)), c\right) = \begin{cases} 1 & : \pi_i^L(\hat{\tau}_i(x)) = c \\ 0 & : \pi_i^L(\hat{\tau}_i(x)) \neq c \end{cases}$$

where $S(x)$ represents the segmentation label, $w_i(x)$ are the weights, and f is a binary function evaluating the agreement.

2) ***Majority Voting:*** A conventional method for label fusion is majority voting, which assigns the most frequently occurring label among the aligned atlas labels to each voxel in the test image space. This method assumes equal reliability for all atlases and is straightforward to implement. However, its performance can

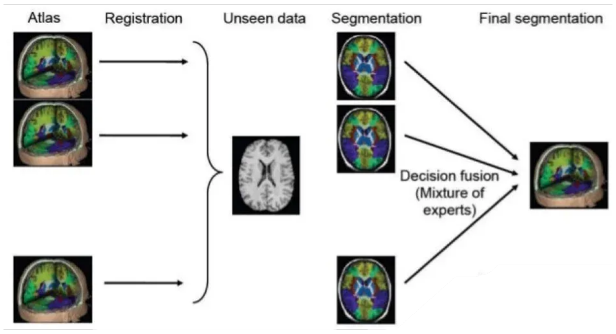


Fig. 2: Multi Atlas Based Segmentation Pipeline. [5]

be compromised in regions where atlas labels exhibit significant disagreement, such as boundaries or areas with inconsistent registration.

Despite its limitations, majority voting serves as a baseline in this framework due to its simplicity and computational efficiency. The results obtained from this method help benchmark the performance of more advanced fusion strategies.

*3) **Window Based Fusion**:* To address the limitations of majority voting, a window-based fusion (3D Patch) method is introduced. Instead of considering the entire image for label assignment, the test image is divided into smaller overlapping windows. Each window is treated as an independent region for fusion, allowing localized refinement of the segmentation process. This approach ensures that local context is preserved and enables the incorporation of spatially varying weights based on the reliability of the atlas labels within each window.

Window-based fusion enhances segmentation accuracy, particularly in regions with variable atlas agreement. Additionally, it provides a flexible framework for integrating advanced similarity measures and adaptive weighting schemes.

*4) **Similarity Measures for Adaptive Votes**:* To improve upon conventional weighted voting, this framework incorporates multiple similarity measures to compute adaptive weights for atlas labels. These measures are used to quantify the alignment and reliability of each elastically registered atlas image with respect to the fixed image (test image). Specifically, the framework uses:

1) **Normalized Cross-Correlation (NCC):** Assesses similarity by accounting for intensity scale differences, ensuring robustness against variations in intensity between the fixed image and the registered.

$$NCC(X,Y) = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2\sum_{i=1}^{N}(y_i-\bar{y})^2}}$$

where x_i, y_i are intensity values in the fixed and registered atlas, $\bar{x}, \bar{y}$ are their means, and N is the number of voxels.

2) **Mean Square Error (MSE):** A pixel-wise measure that calculates the average squared intensity difference between the fixed image and the registered atlas. Lower MSE values signify better alignment and greater reliability.

$$MSE(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i-y_i)^2$$

where x_i, y_i are intensity values in the fixed and registered atlas, and N is the number of voxels.

3) **Entropy:** Quantifies the uncertainty in the registered labels. Atlases with lower entropy are considered more reliable, leading to higher weights in the voting process.

$$H(X) = -\sum_{x\in X} p(x)\log p(x)$$

where $p(x)$ is the probability of intensity x in the registered atlas.

4) **Mutual Information (MI):** Mutual Information (MI): Measures the amount of shared information between the fixed image and each registered atlas. Mutual Information is robust to intensity variations and captures both linear and non-linear relationships, making it an effective similarity measure for image registration.

$$MI(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

where $p(x,y)$ is the joint probability distribution, and $p(x)$ and $p(y)$ are the marginal probabilities.

These similarity measures collectively ensure that weights are dynamically assigned to each atlas based on its alignment quality. The weights are normalized to sum to one, maintaining consistency in the voting process.

*5) **Weighted Voting**:* Using the computed similarity-based weights, we perform weighted voting for label fusion. In this method, the label probabilities at each voxel are scaled by the corresponding atlas weights. The label with the highest weighted probability is assigned to the voxel. This approach addresses the limitations of equal-weight methods like majority voting by dynamically adjusting the contribution of each atlas based on its similarity to the test image.

Weighted voting, especially when combined with similarity measures such as correlation, enhances segmentation robustness and accuracy. By prioritizing reliable atlases and down-weighting less reliable ones, this method ensures improved performance even in challenging anatomical regions.

B. Deep Learning Approach

This study employs three deep learning-based methods for brain MRI segmentation:

1) **Slice-Based Segmentation:** Processes 2D slices extracted from 3D volumes, offering computational efficiency and high accuracy in 2D planes.
2) **Full Volume-Based Segmentation:** Utilizes entire 3D volumes to capture spatial context holistically, improving segmentation performance.
3) **Patch-Based Segmentation:** Focuses on smaller 3D patches, emphasizing critical regions with weighted sampling for improved detail and memory efficiency.

These methods ensure flexibility and robustness across diverse segmentation tasks.

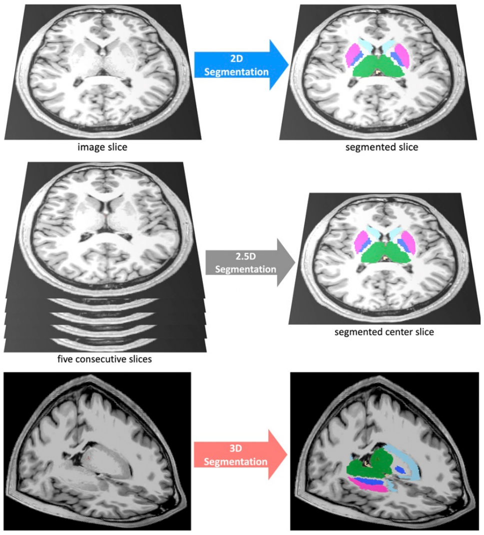


Fig. 3: Brain Tissue Segmentation Methods. [6]

*1) **Loss**:* To optimize segmentation performance, various loss functions were employed to address class imbalance and improve the delineation of brain structures. These include Dice Loss, Dice Cross-Entropy Loss, and Dice Focal Loss.

- **Dice Loss:** Dice Loss directly optimizes the Dice Similarity Coefficient (DSC) to maximize the overlap between predicted and ground truth masks. The formula is given as:

$$\text{Dice Loss} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i + \epsilon}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2 + \epsilon}$$

where p_i and g_i are the predicted and ground truth values, respectively, and ϵ is a small constant to avoid division by zero.

- **Dice Cross-Entropy Loss:** This combines Dice Loss with Cross-Entropy Loss to balance structure overlap and pixel-wise classification. The Cross-Entropy Loss component is defined as:

$$\text{Cross-Entropy Loss} = -\frac{1}{N}\sum_{i=1}^{N} g_i \log(p_i)$$

The overall Dice Cross-Entropy Loss is a weighted sum:

$$\text{Dice CE Loss} = \alpha \cdot \text{Dice Loss} + (1-\alpha) \cdot \text{Cross-Entropy Loss}$$

- **Dice Focal Loss:** This extends Dice Loss by incorporating Focal Loss to focus on hard-to-classify regions. Focal Loss is defined as:

$$\text{Focal Loss} = -\frac{1}{N}\sum_{i=1}^{N} \alpha(1-p_i)^{\gamma} g_i \log(p_i)$$

where α is a weighting factor, and $\gamma > 0$ controls the focus on hard-to-classify examples. The combined Dice Focal Loss is:

$$\text{Dice Focal Loss} = \beta \cdot \text{Dice Loss} + (1-\beta) \cdot \text{Focal Loss}$$

These loss functions were employed across slice-based, full-volume, and patch-based segmentation tasks, ensuring robust training and accurate segmentation of white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF).

*2) **Data Sampler**:* A **weighted sampling strategy** was employed to address class imbalance by assigning probabilities to each class based on their importance in the segmentation task. For instance, with label probabilities such as $\{0:1, 1:3, 2:2, 3:2\}$, the sampler focused more on class 1 while maintaining sufficient representation of other classes. This approach reduced the dominance of the background (class 0) and ensured the model trained effectively on critical foreground regions, thereby enhancing segmentation accuracy.

*3) **Training Configuration**:* The training setup was designed to ensure efficient and effective brain MRI segmentation while addressing challenges such as data imbalance and memory constraints.

- **Batch Size**: 1, to handle 3D volumes efficiently and 16 for 2D based segmentation
- **Learning Rate**: 0.01, decayed using a Cosine Annealing Scheduler for stable convergence.
- **Epochs**: Up to 300 with early stopping based on validation performance.
- **Optimizer**: Adam optimizer, known for adaptability to sparse gradients.
- **Augmentation Strategy**: Applied intensity rescaling, Z-score normalization, elastic deformations, and random flips to enhance model robustness. Performance was tested with and without augmentation.

- **Cross-Validation**: 5-fold cross-validation was used for a thorough evaluation, mitigating biases from train-validation splits.
- **Dropout**: A dropout rate of 10% was applied to prevent overfitting by introducing randomness in the training process.

*4) **Slice-Based Segmentation**:* Slice-based segmentation leverages 2D slices extracted from 3D brain MRI volumes for training and inference. Each volume is decomposed along a specific axis (axial, coronal, or sagittal) into individual slices, which are then processed independently. This approach simplifies the segmentation task by reducing the dimensionality of the input, enabling faster computation and requiring less memory.

For this study, a **2D U-Net with an EfficientNet encoder** was employed. This model, implemented using the Segmentation Models PyTorch library [7], effectively combines the strength of U-Net's hierarchical structure with EfficientNet's advanced feature extraction capabilities.

To address the class imbalance, a **weighted sampling strategy** was adopted to ensure that slices containing underrepresented labels (e.g., cerebrospinal fluid) are sampled more frequently, promoting balanced training. The **Sampling Probability** column in Table I represents the likelihood of selecting a voxel as the center of a patch during training, expressed in relative percentages:

- **Background (0)**: Has a baseline sampling probability of $\frac{1}{9} \times 100 = 11.1\%$.
- **CSF (1)**: Assigned a higher priority with $\frac{4}{9} \times 100 = 44.4\%$, making it sampled much more frequently.
- **Gray Matter (GM, 2)** and **White Matter (WM, 3)**: Each has a sampling probability of $\frac{2}{9} \times 100 = 22.2\%$.

These percentages are derived from the total sampling weights (1 + 4 + 2 + 2 = 9). For example, CSF voxels (44.4%) are sampled 4 times more often than Background (11.1%), while GM and WM (22.2%) are sampled twice as often. This strategy ensures balanced training by prioritizing underrepresented classes.

TABLE I: Sampling Probabilities and Loss Weights for Weighted Sampler

Class	Sampling Probability	Loss Weight
0 (Background)	1	1.1329
1 (CSF)	4	7.3480
2 (Gray Matter - GM)	2	3.5245
3 (White Matter - WM)	2	4.1559

Additionally, a **class-weighted loss function** was used, where the weights were directly computed from the training dataset based on voxel counts for each class. However, due to the significant imbalance in the data, the computed weight for cerebrospinal fluid (CSF) was disproportionately high, potentially destabilizing training. To mitigate this, the weights were normalized using a logarithmic transformation to ensure stability while maintaining the benefits of class weighting (see column 2 in Table I)

Slice-based segmentation is particularly suitable for tasks where computational resources are constrained or when training data is limited, as it enables efficient processing of large volumes of data. Despite its reduced dimensionality, the model demonstrated strong performance, leveraging both robust architectures and carefully balanced training strategies.

*5) **Full Volume Based Segmentation**:* Full volume-based segmentation utilizes entire 3D MRI volumes for training and inference, preserving spatial information critical for accurate anatomical segmentation. This approach eliminates the need to decompose the volumes into slices or patches, making it well-suited for capturing global spatial relationships within the brain.

In this study, a **3D U-Net** architecture was employed, implemented using the MONAI library [8]. The model incorporates instance normalization and dropout layers to enhance stability and mitigate overfitting during training. The **Dice Focal Loss** was used to handle class imbalance effectively, combining the Dice coefficient for overlap accuracy and Focal Loss to address hard-to-classify regions, thereby improving sensitivity to underrepresented structures like cerebrospinal fluid (CSF).

Despite its advantages, this method is sensitive to dataset size. The limited training data in this study posed challenges in achieving robust generalization, with the model struggling to consistently capture fine-grained details. Full volume-based segmentation is promising for scenarios with ample data but may underperform when data is scarce, emphasizing the importance of leveraging larger datasets or advanced augmentation strategies for enhanced performance.

*6) **Patch-Based Segmentation**:* Patch-based segmentation processes 3D MRI volumes by dividing them into smaller, fixed-size sub-volumes (patches). This approach enables detailed learning of localized structures while effectively managing memory constraints. In this study, experiments were conducted with two patch sizes, **64x64x64** and **128x128x128**, to balance computational efficiency and the ability to capture structural details across varying scales.

The segmentation model utilized was the same **3D U-Net** employed for full-volume segmentation, but it was adapted to accommodate the smaller input sizes for patch-based processing. This adaptation allowed the model to focus on localized regions, improving its capacity to learn intricate features within each patch.

The **weighted sampling strategy** and **class-weighted Dice Focal Loss** from the slice-based method were applied here as well, with label probabilities and loss weights detailed in Table I. These strategies ensured balanced training by addressing the inherent class imbalance in the dataset.

During inference, patches were extracted with a **32-voxel overlap** for 64-patch models and a **64-voxel overlap** for 128-patch models, enabling an **ensemble prediction** approach. This overlap blended predictions from neighboring patches, improving segmentation consistency and accuracy across the entire volume. This methodology demonstrated the effectiveness of leveraging patch-based segmentation for capturing fine-grained details while maintaining computational feasibility.

V. Evaluation

A. Metrics

The **Dice Similarity Coefficient (DSC)** was used as the primary metric to evaluate the overlap between the predicted segmentation and the ground truth. The DSC is a widely recognized measure for assessing segmentation accuracy, defined as:

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

where A is the set of voxels in the predicted segmentation, and B is the set of voxels in the ground truth. Higher values indicate better agreement between the prediction and the ground truth.

In addition to DSC, the following metrics were also employed:

- **Hausdorff Distance (HD):** The HD measures the maximum distance between the predicted and ground truth surfaces, capturing the largest boundary discrepancy between the two segmentations. It is defined as:

$$\text{HD}(A, B) = \max\{\sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(b, a)\}$$

  where $d(a, b)$ represents the Euclidean distance between points a and b. Lower HD values indicate better boundary alignment.
- **Average Volumetric Difference (AVD):** The AVD measures the relative volume difference between the predicted and ground truth segmentations, quantifying over- or under-segmentation. It is computed as:

$$\text{AVD} = \frac{|V_{\text{pred}} - V_{\text{gt}}|}{V_{\text{gt}}}$$

  where V_{pred} and V_{gt} represent the predicted and ground truth volumes, respectively. Lower AVD values indicate closer volumetric agreement.

These metrics collectively provide a comprehensive evaluation of segmentation performance, addressing overlap accuracy, boundary precision, and volumetric consistency. They were calculated for each class (e.g., cerebrospinal fluid, gray matter, white matter) across the dataset to assess model performance.

B. Mutil-Atlas

All training images were registered to the validation images using the parameter set Par0009. The registration process aligned the training images to the validation images to ensure spatial correspondence.

Following the registration, the corresponding labels of the training images were transformed into the validation image space using the computed transformation parameters. This transformation ensured that the atlas labels accurately reflected the anatomical structures in the validation image space, forming the basis for label fusion strategies.

*1) **Majority Voting**:* The multi-atlas segmentation framework was first evaluated using majority voting for all validation images. This method assigns each voxel the label that occurs most frequently among the registered atlas labels.

Table II summarizes the results for each validation image.

*2) **Weighted Atlas Fusion with Window-Based Fusion**:* To improve upon majority voting, weighted atlas fusion was implemented using window-based fusion. Different similarity metrics, including Normalized Cross-Correlation (NCC), Mean Squared Error (MSE), Mutual Information (MI), and Entropy, were tested to compute weights dynamically for each atlas.

After extensive experimentation, the window size of $16 \times 16 \times 16$ and stride of $16 \times 16 \times 16$ produced the best results for weighted atlas fusion. Table III presents the evaluation metrics (Dice Score, HD, and AVD) for each similarity metric, demonstrating the performance of the weighted fusion strategy.

The weighted voting strategy slightly improved segmentation accuracy compared to majority voting, achieving an average Dice Score of 0.855, a reduced Hausdorff Distance of 11.0 mm, and an Average Volumetric Difference (AVD) of 0.067 using NCC as the similarity metric. Among all tested similarity metrics, NCC consistently outperformed MSE, Entropy, and MI.

C. Deep Learning

All experiments utilized the data sampler probabilities and augmentation strategies described in Table I and Section IV-B3.

1) Slice-Based Segmentation: To evaluate the effectiveness of slice-based segmentation, multiple experiments were conducted using different configurations and data strategies. The results are presented in Table IV.

Eight experiments with different planes and configurations were conducted as follows:

TABLE II: Majority Voting Evaluation

Experiments	**Dice Score**				**Hausdorff Distance (HD)**				**Average Volumetric Differences (AVD)**			
	CSF	**GM**	**WM**	**Avg**	**CSF**	**GM**	**WM**	**Avg**	**CSF**	**GM**	**WM**	**Avg**
Majority Voting	0.8405	0.8760	0.8357	0.8507	15.3358	9.1405	8.9351	11.1371	0.1071	0.0633	0.0700	0.08013

TABLE III: Weighted Voting-Based Segmentation Evaluation

Experiments	**Dice Score**				**Hausdorff Distance (HD)**				**Average Volumetric Differences (AVD)**			
	CSF	**GM**	**WM**	**Avg**	**CSF**	**GM**	**WM**	**Avg**	**CSF**	**GM**	**WM**	**Avg**
NCC	**0.8497**	**0.8759**	**0.8407**	**0.8554**	15.4168	9.0627	8.6024	**11.0273**	0.0864	**0.0520**	0.0653	0.0679
MSE	0.8454	0.8753	0.8362	0.8523	**15.1818**	9.2210	8.7728	11.0585	**0.0855**	0.0526	**0.0565**	**0.0649**
Entropy	0.8313	0.8668	0.8262	0.8414	17.7892	**8.7505**	8.7164	11.7521	0.1054	0.0615	0.0672	0.0780
MI	0.8408	0.8649	0.8305	0.8454	17.9933	9.2093	**8.3141**	11.8389	0.0918	0.0669	0.0956	0.0848

- **s1**: Coronal + Data Sampler + Data Augmentation
- **s2**: Coronal
- **s3**: Sagittal
- **s4**: Axial
- **s5**: Coronal + Data Sampler
- **s6**: Sagittal + Data Sampler
- **s7**: Axial + Data Sampler
- **s8**: Ensemble s3, s5, s7

The experiments showed that the coronal view with augmentation (s1) performed worse than its non-augmented counterpart (s5), indicating that augmentation did not benefit this dataset and may have introduced unnecessary variability.

Adding a data sampler (s5, s6, s7) significantly improved metrics across all views compared to configurations without samplers (s2, s3, s4). This underscores the sampler's effectiveness in mitigating class imbalance by emphasizing underrepresented classes during training.

Among the individual views, the axial view with a sampler (s7) achieved the highest Dice score (0.9193) and the lowest AVD (0.0094), while the sagittal view with a sampler (s6) obtained the lowest Hausdorff Distance (9.6448).

The ensemble approach (s8), combining predictions from axial, coronal, and sagittal views, delivered the best results across all metrics. It achieved the highest Dice score (0.9352), the lowest Hausdorff Distance (9.5503), and the lowest AVD (0.0048), demonstrating the strength of leveraging multiple complementary views for accurate and consistent segmentation.

2) Full Volume Based Segmentation: Whole-volume segmentation faces significant data imbalance due to the predominance of background voxels. To address this, experiments with different loss functions and configurations, as shown in Table V, were conducted.

- **v1**: Dice Focal Loss
- **v2**: Dice Cross Entropy Loss
- **v3**: Dice Cross Entropy Loss + Augmentation

Focal Loss (v1), designed to emphasize hard-to-classify voxels, was tested but underperformed compared to Dice Cross Entropy Loss (v2). The latter effectively balanced class overlap and voxel-wise classification, achieving superior Dice scores (0.9223) and the lowest Hausdorff Distance (10.9195).

Augmentation combined with Dice Cross Entropy Loss (v3) yielded the best AVD (0.0092), indicating improved volumetric alignment but failed to enhance overall segmentation accuracy.

In conclusion, Dice Cross Entropy Loss (v2) proved to be the most effective, delivering the best balance across all evaluation metrics.

3) Patch-Based Segmentation: Patch-based segmentation experiments were conducted with two patch sizes (64×64×64 and 128×128×128), each evaluated with and without data samplers, along with an ensemble approach combining predictions across configurations as follows:

- **p1**: Patch 64, Stride 32
- **p2**: Patch 64, Stride 32 + Data Sampler
- **p3**: Patch 128, Stride 64
- **p4**: Patch 128, Stride 64 + Data Sampler
- **p5**: Ensemble p1, p2, p3, p4

As can be seen from Table VI, the smaller patch size (64) with overlap (p1) performed well, achieving a mean Dice score of 0.9384 and an HD of 9.7486. Adding a data sampler (p2) improved HD to 7.9140 and achieved the lowest AVD (0.0112), indicating better boundary delineation and volumetric consistency.

The larger patch size (128) (p3) yielded slightly better Dice scores (0.9395) and the best AVD (0.0107). However, the introduction of a data sampler (p4) resulted in an inconsistent HD (27.0194), likely due to overfitting or misrepresentation of class boundaries.

The ensemble approach (p5) combining predictions from p1, p2, p3, and p4 demonstrated superior performance across all metrics. It achieved the highest Dice score (0.9430), the lowest HD (7.8276), and competitive AVD (0.0130), highlighting the benefits of leveraging predictions from multiple configurations.

VI. Comparison

The segmentation results for brain tissues using both atlas-based methods and deep learning-based approaches are shown in Figure 4, visualized across axial, coronal,

TABLE IV: Slice-Based Segmentation Evaluation

Experiments	**Dice Score**				**Hausdorff Distance (HD)**				**Average Volumetric Differences (AVD)**			
	CSF	GM	WM	Avg	CSF	GM	WM	Avg	CSF	GM	WM	Avg
Coronal + Sampler + Aug	0.8459	0.9090	0.9168	0.8906	20.0007	13.6774	14.9360	16.2047	0.0069	0.0470	0.0452	0.0330
Coronal	0.7293	0.9046	0.9067	0.8469	20.1004	17.9091	8.9364	15.6486	0.1580	0.0145	0.0054	0.0593
Sagittal	0.8978	0.9276	0.9195	0.9150	21.2354	**8.5844**	7.8491	12.5563	0.0241	**0.0014**	0.0163	0.0139
Axial	0.9014	0.9241	0.9229	0.9162	13.2919	13.3132	12.3656	12.9902	0.0277	0.0163	0.0119	0.0186
Coronal + Sampler	0.8671	0.9015	0.9151	0.8946	24.5062	23.8343	22.1061	23.4822	0.0233	0.0030	0.0489	0.0251
Sagittal + Sampler	0.8968	0.9239	0.9125	0.9111	12.8219	8.2409	**7.8716**	9.6448	0.0335	0.0046	0.0463	0.0281
Axial + Sampler	0.9082	0.9261	0.9237	0.9193	12.0497	13.1029	9.1734	11.4420	**0.0016**	0.0145	0.0121	0.0094
Ensemble	**0.9212**	**0.9435**	**0.9407**	**0.9352**	**11.6425**	9.0392	7.9693	**9.5503**	0.0091	0.0042	**0.0010**	**0.0048**

TABLE V: Volume-Based Segmentation Evaluation

Loss Function	**Dice Score**				**Hausdorff Distance (HD)**				**Average Volumetric Difference (AVD)**			
	CSF	GM	WM	Avg	CSF	GM	WM	Avg	CSF	GM	WM	Avg
Dice Focal Loss	0.8852	0.9236	0.9096	0.9061	28.1306	14.2657	11.7589	18.0518	0.0066	0.0652	0.0703	0.0474
Dice Cross Entropy Loss	**0.9042**	**0.9358**	**0.9269**	**0.9223**	**11.8027**	**10.4878**	**10.4681**	**10.9195**	0.0215	0.0023	0.0140	0.0126
Dice Cross Entropy Loss + Aug	0.8820	0.9206	0.9057	0.9028	29.9572	13.7154	10.6235	18.0987	**0.0009**	**0.0010**	**0.0259**	**0.0092**

TABLE VI: Patch-Based Segmentation Evaluation

Configuration	**Dice Score**				**Hausdorff Distance (HD)**				**Average Volumetric Difference (AVD)**			
	CSF	**GM**	**WM**	**Avg**	**CSF**	**GM**	**WM**	**Avg**	**CSF**	**GM**	**WM**	**Avg**
Patch 64	0.9246	0.9476	0.9431	0.9384	10.6180	8.3984	10.2294	9.7486	0.0208	0.0119	0.0078	0.0135
Patch 64 + Sampler	0.9238	0.9470	0.9430	0.9379	8.8792	**7.6025**	7.2603	7.9140	**0.0117**	0.0145	0.0074	0.0112
Patch 128	0.9252	0.9495	0.9438	0.9395	12.3265	8.4761	8.4534	9.7520	0.0188	**0.0077**	0.0058	**0.0107**
Patch 128 + Sampler	0.9277	0.9455	0.9381	0.9371	14.8209	58.5632	7.6742	27.0194	0.0194	0.0200	0.0096	0.0163
Ensemble	**0.9305**	**0.9517**	**0.9470**	**0.9430**	**8.5560**	7.9178	**7.0091**	**7.8276**	0.0213	0.0128	**0.0047**	0.0130

and sagittal planes for Case 11. Below is a detailed comparison of the two approaches:

A. Atlas-Based Segmentation

- Relies on pre-aligned anatomical templates and image registration for segmentation.
- Effectively captures general brain structures but struggles with precise boundary delineation, particularly in regions with high inter-subject variability or anatomical irregularities.
- Exhibits noticeable misclassification in areas with partial volume effects, especially at CSF-GM and GM-WM interfaces.

B. Deep Learning-Based Segmentation

- Leverages a data-driven approach, learning features from a labeled dataset, resulting in improved adaptability to individual anatomical variations.
- Provides superior performance in boundary delineation, capturing finer details of brain tissues with sharper transitions between CSF, GM, and WM.
- Achieves higher segmentation accuracy due to its ability to generalize better across the dataset, reducing over-segmentation and under-segmentation issues observed in atlas-based methods.

In summary, while atlas-based segmentation serves as a reliable baseline method, deep learning significantly outperforms it in terms of accuracy, precision, and adaptability. The findings validate the potential of deep learning as a superior approach for brain tissue segmentation tasks.

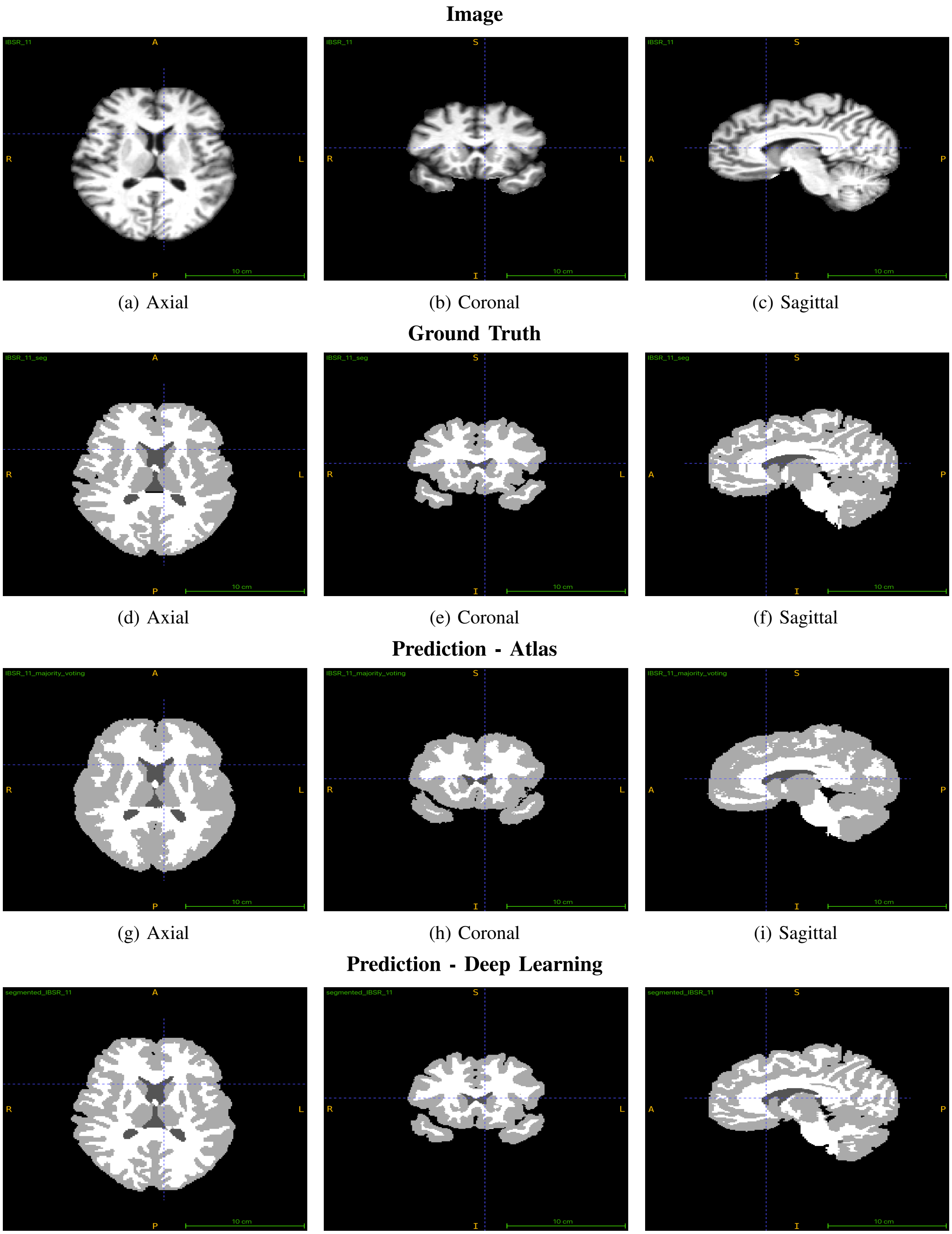
Finally, we used the ensemble approach (p5) for segmenting the test images. Figure 5 visualizes the segmentation of test case 10, showcasing the performance of the ensemble method in accurately segmenting brain tissues across CSF, GM, and WM regions.

VII. Discussion

This study highlights the advancements in brain tissue segmentation by leveraging both traditional multi-atlas-based methods and modern deep learning approaches. The results underline the complementary strengths of these techniques, paving the way for improved segmentation accuracy and robustness.

The traditional approach of multi-atlas-based segmentation demonstrated its reliability in capturing general brain structures. Using elastix for registration and label fusion through majority voting served as a robust baseline. However, the limitations of majority voting, such as its inability to handle inter-subject variability and boundary inconsistencies, were addressed through weighted atlas fusion with window-based strategies. By introducing similarity metrics like NCC, MSE, Entropy, and MI, the adaptive weighting mechanism enhanced segmentation accuracy, particularly in challenging anatomical regions. The incorporation of a 16×16×16 window size with stride ensured localized refinement, significantly improving the results.

Deep learning-based segmentation exhibited superior performance due to its ability to learn complex features directly from data. Slice-based segmentation offered computational efficiency, while patch-based segmenta-



Image

(a) Axial

(b) Coronal

(c) Sagittal

Ground Truth

(d) Axial

(e) Coronal

(f) Sagittal

Prediction - Atlas

(g) Axial

(h) Coronal

(i) Sagittal

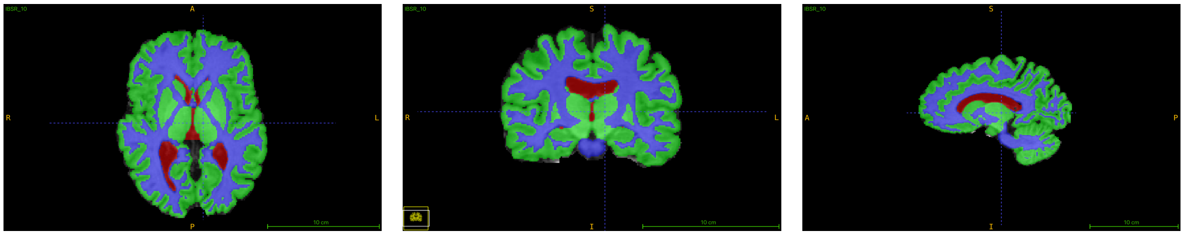
Prediction - Deep Learning

(j) Axial

(k) Coronal

(l) Sagittal

Fig. 4: Comparison of Brain Tissue Segmentation Between Atlas-Based and Deep Learning-Based Segmentation: Image, Ground Truth and Predictions Visualized Along Axial, Coronal, and Sagittal Axes for Case 11



(a) Axial (b) Coronal (c) Sagittal

Fig. 5: Test Prediction for Case 10 Visualized Along Separate Axes

tion provided a balance between memory constraints and detailed feature learning. Full-volume segmentation preserved spatial context, improving boundary delineation but requiring larger datasets for optimal generalization.

The experiments revealed that ensemble strategies combining predictions from multiple configurations provided the best results across metrics. This highlights the potential of integrating multiple views and methodologies to leverage complementary strengths.

While atlas-based methods are computationally intensive and rely heavily on accurate registration, they provide a reliable framework for segmentation in the absence of extensive labeled datasets. Conversely, deep learning approaches excel in adaptability and precision but require robust training datasets and computational resources. The findings validate the efficacy of combining traditional and modern approaches, with deep learning significantly outperforming traditional methods in terms of Dice scores, Hausdorff distance, and volumetric differences.

Despite achieving state-of-the-art results, this study acknowledges certain limitations. The traditional methods are computationally intensive, particularly during the registration process, while deep learning models require extensive training data and computational power. Future work could explore hybrid approaches, leveraging the strengths of both methodologies to further enhance segmentation accuracy and efficiency. Additionally, expanding the dataset size and diversity could improve generalization for deep learning models, addressing challenges in rare and complex anatomical regions.

This comparative analysis underscores the importance of method selection based on specific application needs, with both traditional and deep learning approaches offering unique advantages for brain tissue segmentation.

References

[1] NITRC, "IBSR: Tool/Resource Info," https://www.nitrc.org/projects/ibsr, accessed: [date of access].

[2] S. Valverde *et al.*, "Comparison of 10 brain tissue segmentation methods using revisited ibsr annotations," *Journal of Magnetic Resonance Imaging*, vol. 41, pp. 93–101, Jan. 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/jmri.24517

[3] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-B. Poline, C. D. Frith, and R. S. Frackowiak, "Statistical parametric maps in functional imaging: A general linear approach," *Human brain mapping*, vol. 2, no. 4, pp. 189–210, 1994.

[4] X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de Solorzano, "Combination strategies in multi-atlas image segmentation: application to brain mr data," *IEEE Transactions on Medical Imaging*, vol. 28, no. 8, pp. 1266–1277, 2009.

[5] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain mri segmentation combining label propagation and decision fusion," *NeuroImage*, vol. 33, no. 1, pp. 115–126, 2006.

[6] A. Avesta, S. Hossain, M. Lin, M. Aboian, H. M. Krumholz, and S. Aneja, "Comparing 3d, 2.5d, and 2d approaches to brain image auto-segmentation," *Bioengineering*, vol. 10, no. 2, 2023. [Online]. Available: https://www.mdpi.com/2306-5354/10/2/181

[7] P. Yakubovskiy, "Segmentation Models PyTorch," https://github.com/qubvel/segmentation_models.pytorch/, 2019.

[8] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, V. Nath, Y. He, Z. Xu, A. Hatamizadeh, A. Myronenko, W. Zhu, Y. Liu, M. Zheng, Y. Tang, I. Yang, M. Zephyr, B. Hashemian, S. Alle, M. Z. Darestani, C. Budd, M. Modat, T. Vercauteren, G. Wang, Y. Li, Y. Hu, Y. Fu, B. Gorman, H. Johnson, B. Genereaux, B. S. Erdal, V. Gupta, A. Diaz-Pinto, A. Dourson, L. Maier-Hein, P. F. Jaeger, M. Baumgartner, J. Kalpathy-Cramer, M. Flores, J. Kirby, L. A. D. Cooper, H. R. Roth, D. Xu, D. Bericat, R. Floca, S. K. Zhou, H. Shuaib, K. Farahani, K. H. Maier-Hein, S. Aylward, P. Dogra, S. Ourselin, and A. Feng, "Monai: An open-source framework for deep learning in healthcare," 2022. [Online]. Available: https://arxiv.org/abs/2211.02701