IBSR18 Brain Tissue Segmentation Using nnUnet and Multi-atlas Approaches

Abdelrahman HABIB
ESCOLA POLITÈCNICA SUPERIOR
Universitat de Girona
Girona, Spain
u1985258@campus.udg.edu

Edwing, ULIN
ESCOLA POLITÈCNICA SUPERIOR
Universitat de Girona
Girona, Spain
u1984919@campus.udg.edu

Carmen Colin Tenorio
ESCOLA POLITÈCNICA SUPERIOR
Universitat de Girona
Girona, Spain
u1984944@campus.udg.edu

***Abstract*—In this project, we employed two segmentation approaches for three main brain tissues—cerebrospinal fluid (CSF), gray matter (GM), and white matter (WM)—using the IBSR 18 dataset. The first approach involved a multi-atlas technique, while the second utilized deep learning, specifically the nnUnet neural network. Following extensive experiments, the deep learning approach demonstrated superior performance with Dice scores of 0.92 for CSF, 0.88 for GM, and 0.94 for WM. These impressive results highlight the effectiveness of the nnUnet neural network in accurately segmenting brain tissues.**

***Index Terms*—Brain Segmentation Multi-Atlas nnUnet**

I. Introduction

Medical image segmentation is a critical aspect of modern diagnostic and treatment planning in neuroimaging. Among the diverse methodologies employed for the segmentation of brain tissues, we can find traditional multi-atlas techniques and state-of-the-art deep learning models.

Atlas-based segmentation is a commonly used technique to segment image data. In atlas-based segmentation, an intensity template is registered non-rigidly to a target image and the resulting transformation is used to propagate the tissue class or anatomical structure labels of the template into the space of the target image. Multi-atlas aproach is when several atlases from different subjects are registered to target data. The label that the majority of all warped labels predict for each voxel is used for the final segmentation of the target image.

The nnU-Net (no-new-Net) segmentation method represents a cutting-edge approach in the field of medical image segmentation, particularly for brain tissue analysis [**?**]. no-new-Net (nnU-Net), a segmentation method that includes a formalism for automatic adaptation to new datasets. Based on an automated analysis of the dataset, nnU-Net automatically designs and executes a network training pipeline. Being wrapped around the standard U-Net architecture. without any manual fine-tuning, the method achieves state-of-the-art performance on several well-known medical segmentation benchmarks. [**?**]

II. Dataset

The MRI dataset used in this project, IBSR 18, consisted of 18 volumes, which were divided into training, validation, and testing sets.

Dataset Split	Number of Images
Training	10
Validation	5
Test	3

TABLE I: Dataset Split and Number of Images

The images provided had different spatial resolutions and intensity distributions, with varying pixel sizes. The dataset encompassed three distinct pixel sizes, as detailed in Table II.

The diversity in intensity distributions among the images was taken into consideration, as it holds significance for the segmentation process.

Number	Pixel Size
1	0.9375, 1.5, 0.9375
2	1.0, 1.5, 1.0
3	0.8371, 1.5, 0.8371

TABLE II: Different Pixel Sizes

III. Preprocessing

All images underwent skull-stripping, rendering this step unnecessary. This preprocessing pipeline was only necessary for the second aproach MultiAtlas.

A. Bias-Field Correction

The initial and essential preprocessing step involved bias-field correction using the `N4BiasFieldCorrectionImageFilter()` method from SimpleITK. This filter effectively corrected intensity inhomogeneities, providing enhanced images. The correction process was applied uniformly to all images in the dataset.

B. Normalization

Given the diverse intensity distributions across images and the distinct characteristics of the target image, a normalization step was neccesary. We employed the `HistogramMatchingImageFilter()` method from SimpleITK to match each image's histogram to a specific target.

In our experiments, we performed two normalization approaches. Initially, we chose one training image (CASE 04) as the target, aligning all other images with its distribution. Additionally, we explored a more intricate method using individual testing images as targets. However, this approach required separate preprocessing for each test image, making the algorithm more complex.

IV. Multi-atlas Segmentation

A. Registration and Label Propagation

For us to create a multi-atlas approach, we had to register the intensity images to a fixed reference frame in order to fuse the labels and create a deterministic segmentation for the three tissues. To do so, we followed two registration strategies.
The first registration strategy was to register all of the training intensity volumes to all of the test intensity volumes (we use the validation data to evaluate as we were given its labels), and then fuse the atlases using the different techniques that will be discussed in the next sub-section IV-B of label fusion. The second registration strategy was to register the volumes that have similar voxel sizes in millimeters together, making less number of registration to each test volume space. In the given dataset, there were three unique voxel sizes, which are $(0.9375, 1.5, 0.9375)$, $(1.0, 1.5, 1.0)$ and $(0.8370536, 1.5, 0.8370536)$. Each train volume from each voxel size was registered to the test of its similar size for this strategy experiment.
For Elastix and Transformix, we used version $elastix\_windows32\_v4.2$ and $Par0010$ affine and b-spline parameter files. We also did an additional step before propagating the labels, which is to change the $FinalBSplineInterpolationOrder$ value inside the transformation parameters file generated by Elastix from 3 to 0. This is to ensure that the labels volumes has integer intensity values and not floats.

B. Label Fusion

The label fusion is considered the most important technique to generate hard segmentation in a multi-atlas approach, and it is where most of the improvements on the baseline without any pre-processing can be done. In our implementation, we developed three label fusion techniques to fuse all the labels generated from the different registration strategies. Those techniques are: Majority Voting, Weighted Voting, and STAPLE.

1) Majority Voting: Majority voting is implemented in a way that all the propagated labels volumes are used in a majority voting technique, where based on the majority of the labels of each pixel, the final label will be determined on that majority for the output intensity segmentation.

2) Weighted Voting: Weighted voting is quite similar to majority voting, where the main difference is that we give a weight to each label based on a metric. We used a Mutual Information (MI) similarity metric, where we compared the registered intensity to the test intensity, and created weights for each label that is multiplied by that label. As the weight is a float, and when multiplied by the propagated label volume it won't have any longer integer labels. We obtained the mean for each class label for all label volumes, and computed the argmax among all tissue classes as well as the background.

3) STAPLE: To implement STAPLE, we used $STAPLE$ from $SimpleITK$ library that takes a list of binary masks for a single label class, and creates a probabilistic segmentation as mentioned in the documentation. Thus, we created a list for each class including the background, fused using the built-in STAPLE and then threshold-ed using a threshold of 0.6 to create a final hard segmentation of the brain volume.

V. Deep Learning Segmentation

In this project we implemented an application of the nnUNet framework, with a customize trainer to achieve the $IBSR18$ data set segmentation. The $nnUNET$, was selected due to it's self-configuring capabilities in medical image segmentation, automated configuration, adaptability and robustness. $nnUNet$ intelligently adapts the $UNet$ architecture, determining the optimal depth, width, number of convolutional filters, and pooling operations for each specific dataset. This adaptability was vital in selecting the model to the unique aspects of brain MRI images in the $IBSR18$ dataset.The implementation used the provide documentation from the Division of Medical Image Computing, German Cancer Research Center (DKFZ). [1]

A. Prepocessing

In this implementation we used the $nnUNET$ prepossessing approach, since it is design to be a benchmark method to compared your deep learning segmentations. The steps used to achieve the preprocessing consists. The used of resampling since the voxel spacing in the dataset is not consisting. In the nnUNET a pipeline of this methods is used to achieve the prepocessing; use of the median spacing to determine the targeting spacing to resample, the used of third order spline interpolation to preserve the image quality, then image are padded or cropped to a fixed size and at last a correction of the intensity.

B. Training and Validation

In the training a custom trainer with a less epochs was chosen since the normal model is trained until the 1000 epochs are achieved. The used of the nnUNET with our alterations can be seen in the project GitHub. [2] In the training we opted to used the 5-fold cross-validation too ensured a comprehensive learning and maximized the model's generalization. The steps that are follow in the nnUNET from DKFZ [1] can be summarized as follow:

1) Model Configuration: $nnUNet$ automatically configures its network architecture and hyperparameters based on the dataset's characteristics, like image size, spacing, and modality.

2) Training Procedure: The network use the random sampled patches from the full-resolution image and a suitable loss function is selected normally a combination of Dice loss and cross-entropy loss.

3) During training: Periodic validation is performed on a separate set of data to monitor the model's performance and avoid overfitting. $nnUNet$ might adjust its strategies based on the observed performance, such as changing the patch size or sampling strategy.

4) After training: $nnUNet$ often employs ensembling and test-time augmentation to improve segmentation accuracy on the test data. The best-performing models during validation are chosen for the final segmentation task.

C. Post processing

The $nnUNet$ includes a prepossessing steps that help with the data cleaning and obtaining higher results when evaluating the model. This steps include:

1) Thresholding and Cleaning: Applying a threshold to convert the network's probabilistic output into a binary segmentation, followed by the removal of small, isolated segments that are likely to be noise or false positives.

2) Connected Component Analysis: This involves identifying and possibly discarding small connected components based on predefined criteria, to enhance the accuracy and relevance of the segmentation.

3) Region of Interest (ROI) Adjustment: Adjusting the segmentation to fit within the expected ROI, which can involve cropping or expanding the segmented regions based on anatomical knowledge or predefined constraints.

4) Mapping Back to Original Space: If the data was resampled during preprocessing, the segmented images are resampled back to their original space to align with the original imaging data.

VI. Results and Discussion

A. Preprocessing

Given the diverse histogram distributions across all images, a crucial step in standardization involved selecting a singular reference. Volume 4 from the training set was strategically chosen as our target histogram. Figure 3 illustrates the results of the preprocessing step, focusing on histogram matching as a key technique. In this figure volume N° 14 it is matched to the target histogram.

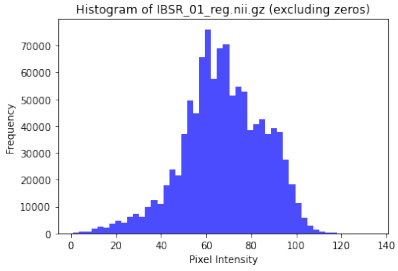
B. Multi-Atlas Results

In this subsection, we present quantitative and qualitative results of the segmentation performance achieved by our Multi-Atlas approach.

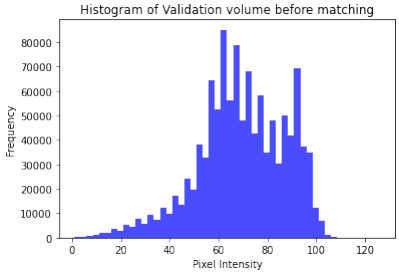
1) Quantitative Results: In this section, we present the segmentation results obtained through various fusion techniques in the context of multi-atlas segmentation. In the case of the weighted fusion, Table III, presented very good results accross all the patients, with a mean values avobe 80, for this case the most difficult tissue to segmenet was CSF. The best Dice score was for GM class, with values above .86.

The quantitative results for the majority voting, are presented in Table IV. For this tecnique we values for GM around .87 in patient IBSR_13. In the case of CSF the values are still the lower results. The maximum mean for the three classes was .842.

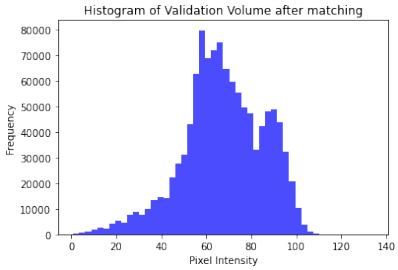
For the staple fusion technique, as presented in Table V, the mean Dice score was below 0.80, indicating comparatively lower overall segmentation performance. This technique faced challenges, especially in CSF segmentation, resulting in the



(a) Target Histogram



(b) Histogram before matching



(c) Histogram after matching

(d) Example of histogram matching

Volume	WM	GM	CSF	Mean
IBSR_11	0.794	0.822	0.785	0.801
IBSR_12	0.799	0.822	0.790	0.803
IBSR_13	0.783	0.866	0.770	0.806
IBSR_14	0.807	0.865	0.815	0.830
IBSR_17	0.783	0.869	0.865	0.838

TABLE III: Weighted voting fusion

Volume	WM	GM	CSF	Mean
IBSR_11	0.787	0.826	0.775	0.796
IBSR_12	0.801	0.830	0.815	0.815
IBSR_13	0.787	0.870	0.765	0.807
IBSR_14	0.805	0.866	0.809	0.827
IBSR_17	0.782	0.869	0.876	0.842

TABLE IV: Majority voting fusion

lowest scores among the three tissue classes.

Volume	WM	GM	CSF	Mean
IBSR_11	**0.802**	0.769	0.757	0.776
IBSR_12	0.755	0.740	0.684	0.726
IBSR_13	0.710	0.782	0.729	0.740
IBSR_14	0.778	**0.802**	**0.780**	**0.787**
IBSR_17	0.738	0.798	0.771	0.769

TABLE V: Staple fusion

To facilitate a comprehensive comparison of different fusion techniques, we present the mean Dice scores for each method in Table VI. In this case, majority voting shows better results among the others with a mean dice score of 0.8174, closely followed by weighted voting fusion with 0.8158.

Staple fusion obtains the lowest mean dice score at 0.7597, primarily influenced by its performance in CSF segmentation, where it scores 0.7442. Notably, majority voting excels in GM and CSF, while weighted voting fusion outperforms in WM.

Fusion Technique	WM	GM	CSF	Mean
Staple fusion	0.7565	0.7785	0.7442	0.7597
Majority voting fusion	0.7923	**0.8521**	**0.8078**	**0.8174**
Weighted voting fusion	**0.7933**	0.8489	0.8052	0.8158

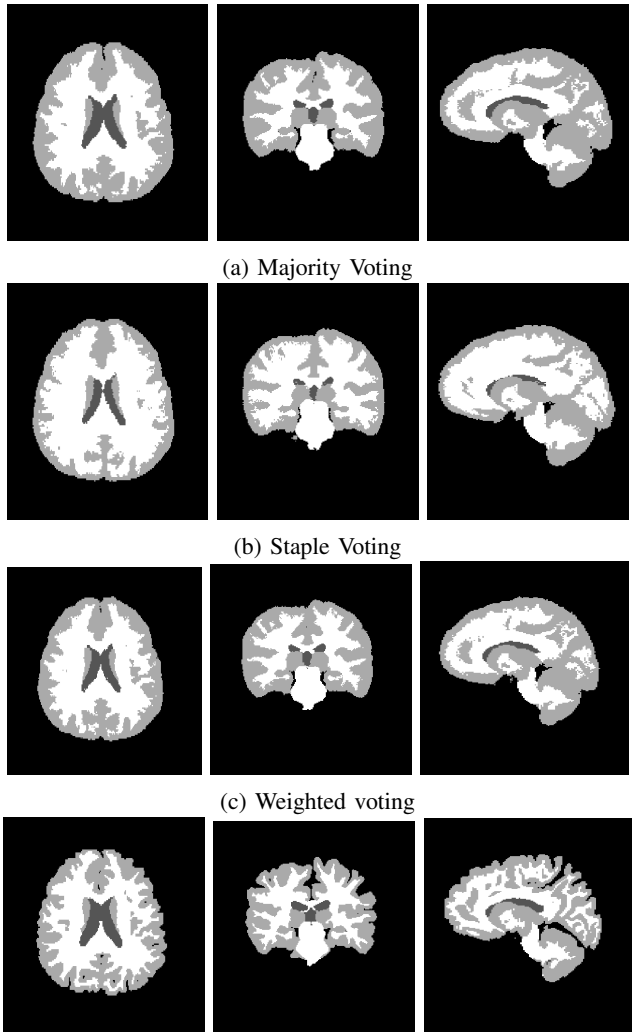
TABLE VI: Dice Scores for Fusion Techniques

2) Qualitative Results: In this subsection, we present qualitative results showcasing the segmentation outcomes using different fusion techniques compared to the ground truth. Figure 2 provides visual representations of the segmentation results for a validation volume across axial, coronal, and sagital views. In this figure it can be shown that the CSF class is the more challenging to segment. For majority voting, 2 a), we can see there are some difficulties specially in the border of the bran when training to find fine borders or curves. Also for majority voting it was difficult to differentiate in the borders of GM and GM and is labeling as GM. For staple fusion, 2 b), the segmentation tend to label as WM in the borders between GM and WM. In general the three techniques had difficulties in the border of the brain, covering areas that do not correspond to matter. For weighted voting, 2 c), this one tends to label as GM in the areas between GM and WM. Additionally, CSF was challenging to segment comparing to GT , Fig 2 d),

C. Deep Learning Results

In this subsection, we present quantitative and qualitative results of the segmentation performance achieved by our deep learning approach.

1) Quantitative Results: For the deep learning approach using the nnUnet, the obtained class-specific Dice scores and the mean Dice score for the validation set are summarized in Table VII. The mean Dice score provides an overall assessment of the model's performance across all classes, with a value of 0.9293 indicating a high level of segmentation accuracy for the validation set.



(a) Majority Voting

(b) Staple Voting

(c) Weighted voting

(d) GT

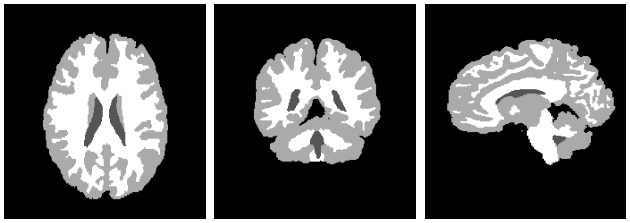
Fig. 2: Example Segmentation Result vs Ground Truth in a validation volume

WM	GM	CSF	Mean
0.8998	0.9485	0.9396	0.9293

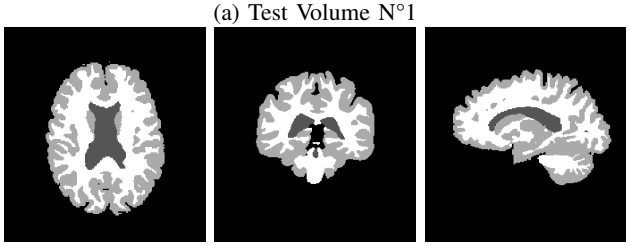
TABLE VII: Class Dice Scores and mean for Validation set

2) Quantitative Results: Figure 3 showcases examples of segmentation results for three different volumes from the testing set. Each row corresponds to a different test volume (1, 2, and 3), and each column represents a distinct anatomical view (axial, coronal, and sagital).

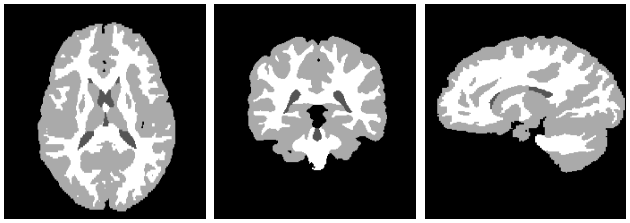
Figure 4 illustrates the segmentation results compared to the ground truth for two validation volumes (Volume N°11 and Volume N°12). Each row corresponds to a specific volume, with the left column displaying axial views, the center column showing coronal views, and the right column presenting sagittal views. As it is shown, the segmentation result is very similar to the ground truth, only with few differences. For the



(a) Test Volume N°1



(b) Test Volume N°2



(c) Test Volume N°3

Fig. 3: Example of segmentation result using the testing set: Volume N°1,2 & 3

three classes the result is very good, specially in the most complicated zones, specially with intricate details, such as borders and class separation. Despite minor differences, the segmented areas remain consistent with the ground truth, even in challenging regions. The sagittal view for CSF exemplifies the model's capability to segment small portions.

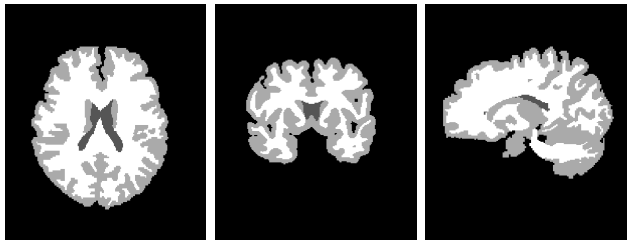
D. Multi-Atlas vs Deep Learning

The nnUnet method provided superior segmentation accuracy, especially in complex brain regions. However, it required significantly more computational resources compared to the multi-atlas approach, which offered faster processing times but with slightly lower accuracy. One of the major limitations of the multi-atlas method is its sensitivity to the anatomical variability among patients. On the other hand, the nnUnet's performance is highly dependent on the volume and quality of training data. In scenarios where computational resources are limited, the multi-atlas method may be more feasible, whereas nnUnet is preferable in settings where accuracy is paramount and resources are abundant.

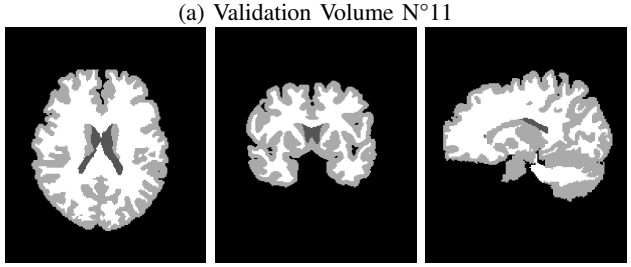
CONCLUSION

This project demonstrates the effectiveness of both the multi-atlas and nnUnet methods in brain tissue segmentation using the IBSR 18 dataset. With three different techniques for multi-atlas segmentation (Majority voting, Staple voting, and Weighted voting), we achieved a mean Dice score of 0.81 in the validation set, indicative of robust segmentation accuracy.

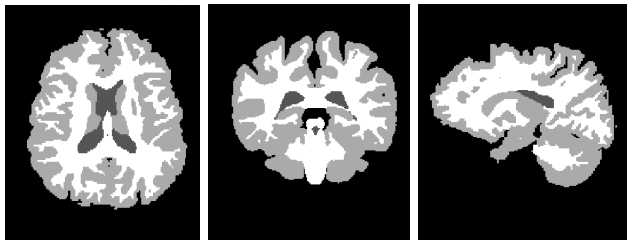
In the second approach, deep learning, we employed the nnUnet model, which proved to be particularly effective for brain segmentation. The implementation of the nnUnet model yielded promising results, obtaining a mean Dice score of 0.92 on the validation set.



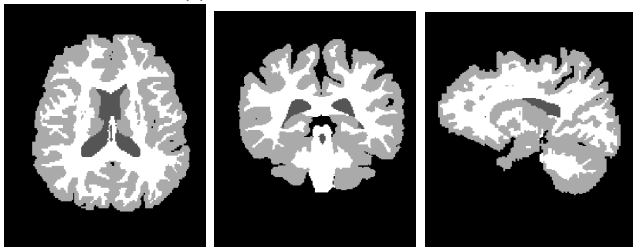
(a) Validation Volume N°11



(b) GT Validation Volume N°11



(c) Validation Volume N°12



(d) GT Validation Volume N°12

Fig. 4: Example Segmentation Result vs Ground Truth in a validation volume

REFERENCES

[1] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, pp. 1–9, 2020.

[2] A. Habib, E. Ulin, and C. Colin, "Ibsr18-brain-tissue-segmentation: Brain tissue (wm, gm, csf) segmentation using both multi-atlas and nnunet approaches," https://github.com/abdel-habib/IBSR18-brain-tissue-segmentation, 2020.