# Enhancing Brain Tissue Analysis: A 2D and 3D CNN Ensemble Approach for MRI Segmentation

X. Beltran Urbano[1] and F. Hartmann[1]

[1] *University of Girona, Erasmus Mundus Joint Master Degree in Medical Imaging and Applications (MAIA)*

**Abstract**—Brain tissue segmentation is crucial in medical imaging for accurately distinguishing different brain areas and is essential for diagnosing and treating neurological conditions. It also significantly aids neuroscientific research by allowing for in-depth study of brain structure and function, advancing our knowledge of the brain. In this project, we present a comprehensive ensemble approach using 2D and 3D convolutional neural networks to segment brain tissues in MR images. We employ the IBSR18 dataset to train and validate our models, focusing on the segmentation of gray matter, white matter, and cerebrospinal fluid. Our method leverages the strengths of different network architectures and planes, integrating them into an effective ensemble framework. We demonstrate that this approach not only improves the accuracy and robustness of segmentation but also provides insightful implications for medical imaging analysis.

**Keywords**—Brain Tissue Segmentation, IBSR18, Convolutional Neural Networks, MR Images, Ensemble Learning, Gray Matter, White Matter, U-Net, Cerebrospinal Fluid, Deep Learning

## I. INTRODUCTION

Segmentation of medical images has been utilized for decades and is an indispensable method for enhancing patient diagnosis and treatment. By means of this procedure, an image is partitioned into distinct subregions according to common attributes such as similarity, texture, and intensity. Common practice in brain imaging is the segmentation of images into distinct regions, which facilitates tissue extraction. Particularly, segmenting magnetic resonance (MR) images to distinguish gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF) is a popular technique in quantitative brain analysis. The aforementioned methodology is of considerable importance in the identification and management of neurological disorders, including Alzheimer's and Parkinson's disease.

In this project, we have developed an innovative approach to performing brain tissue segmentation using deep learning techniques.

## II. Materials and Methods

### I. Dataset

The dataset used to develop this approach contains 18 T1-w scans of normal subjects from the Internet Brain Segmentation Repository (IBSR) [1], available from the Center for Morphometric Analysis at Massachusetts General Hospital. Also known as ISBR18, this dataset is formed by scans with 1.5 mm slice thickness (256 128 256), which have been previously preprocessed with the Autoseg bias field correction routines from the Center for Morphometric Analysis. The dataset is also supplied with the ground truth corresponding to white matter (WM), grey matter (GM), and cerebrospinal fluid (CSF) [2]. In addition, this dataset was acquired by three different laboratories. To implement our approach, the dataset has been split as follows:

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

For the test set, we were not provided with the ground truth since those cases will be the ones used to evaluate this project.

### II. Preprocessing

In order to better segment the tissues of the brain, some preprocessing techniques have been applied to the raw data. They are the following:

- **Normalisation:** Due to the fact that the scans have been acquired in different laboratories, their intensity distribution is different. For this reason, a normalization step has been performed in order to deal with this inconvenience. To perform this task, a robust z-normalization technique has been used. We selected this normalization approach due to the non-Gaussian distribution observed in Group 2's data, characterized by a significant presence of outliers towards higher values, as shown in Figure 1. The equation used for normalization is displayed in Equation 1.

$$X_{\text{noramlized}} = \frac{X - \text{mean}_Q(X)}{\text{std}_Q(X)}; \quad \forall X > 0 \tag{1}$$

  Where:

  - $X$ are the intensity values.

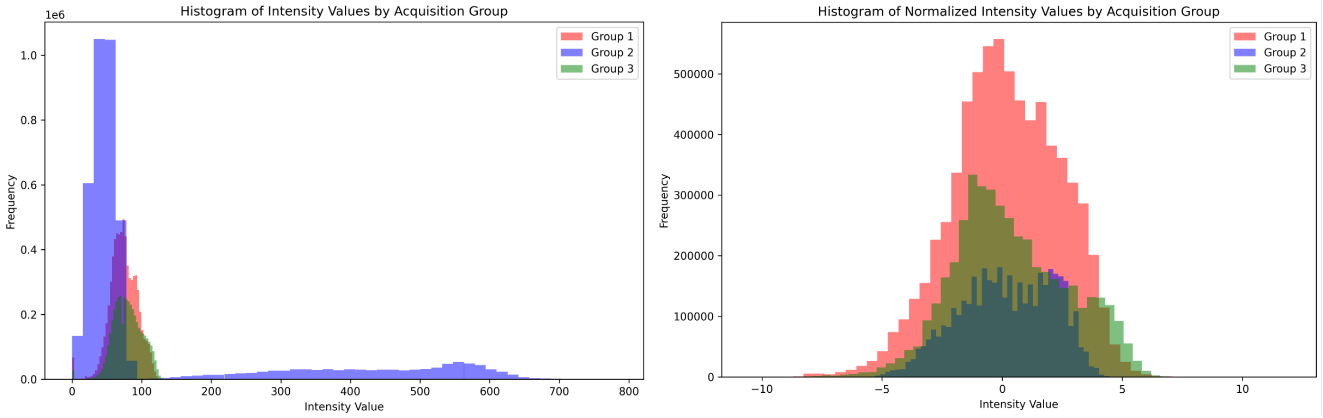

**Fig. 1:** Example of the ISBR18 dataset. Each of the colours (red, orange and blue) corresponds to a different laboratory where the images were acquired.

  - $\text{mean}_Q(X)$ is the mean computed using the 25th and 75th quantiles.
  - $\text{std}_Q(X)$ is the standard deviation computed over the same quantile range.

- **Data augmentation:** To improve the algorithm's reliability, random flips and rotations were performed on the original images as a part of the augmentation process.

- **Slice Selection:** Since we want to perform a multi-class segmentation approach where we have a clear class imbalance (the CSF label is the minority class), we have selected the images we used for the training in such a way that we used as many slices containing CSF as we could. In the event that we have fewer CSF slices than the selected minibatch, we randomly select the remaining slices, always taking into account that those slices do not contain only background. In the event that we have more CSF slices than we need, we just select them randomly among the ones we have and discard the remaining ones.

## *III. Models*

For this project, numerous model architectures have been developed from scratch, fine-tuned, and evaluated. These models include:

1. 2D U-Net
2. 2D Res-U-Net
3. 2D Multi-Resolution-U-Net
4. 2D Dense-U-Net
5. 3D U-Net

All U-Net architectures are adaptations of the original U-Net from Ronneberger et al. [3]. Some of these adaptations involve the use of residual and dense blocks. Besides the models implemented from scratch, other models have been used, such as the 2D and 3D SegResNet [4] which has been presented at MICCAI as the winner of the Brats2018 challenge. The code used is available through MONAI [5]. Since our project was built using TensorFlow, we chose to modify the PyTorch model code[1] to work with TensorFlow, ensuring consistency throughout the project. Furthermore, SynthSeg [6] was used for this project. In comparison to the techniques mentioned before, SynthSeg is contrast-agnostic, which means it is not dependent on intensity but focuses on other features such as shape. The key idea is to deform the mask and resample each class from a random distribution.

## *IV. Input*

In this project, multiple versions of inputs have been tested. Firstly, two- and three-dimensional input. For the 3D models, we've utilized the entire image as input. Although a patch-based approach with patch selection was implemented and tested for 3D, it's not discussed further in this report since it did not improve results and we had sufficient computational resources available. Secondly, and even more interestingly, we must take into account the orientation of the 2D slices. While slice selection, as previously mentioned, is crucial, it's equally important to acknowledge that distinct anatomical orientations contain different information. Although axial slices are commonly used in the literature, we made the deliberate choice to train multiple models on coronal slices as well. It is worth mentioning that the axial slices had to be zero-padded to ensure a square image. As the shape of the axial slices was 256x128, half of the images had to be added. However, all padded regions were removed during the reconstruction of the image.

## *V. Training*

In this section, the techniques used for training the model are described.

- **Loss function**: In this challenge, various loss functions, including dice loss and cross-entropy loss, were experimented with. However, to address the class imbalance issue, we ultimately opted for a weighted categorical cross-entropy loss. This allows us to give a higher importance, i.e., weight, to the minority class of cerebrospinal fluid. The formula can be found in Equation 2.

$$L = -\sum w \cdot (y \log(\hat{y})) \tag{2}$$

[1] Accessed on: January 12, 2024. URL: `https://docs.monai.io/en/latest/_modules/monai/networks/nets/segresnet.html`

Here,

- L is the weighted categorical cross-entropy loss.
- $w$ are the weights given to each class.
- $y$ is the ground truth.
- $\hat{y}$ are the predicted labels.

The weights for each class were empirically chosen based on the distribution and difficulty of segmentation. The weights can be found in Table 1.

TABLE 1: THE WEIGHTS ASSIGNED FOR EACH CLASS FOR THE WEIGHTED CATEGORICAL CROSS-ENTROPY LOSS

| Class | Weight $w$ |
|---|---|
| Background | 1 |
| Cerebrospinal Fluid | 10 |
| Gray Matter | 3 |
| White Matter | 3 |

- **Learning Rate:** Additionally, a learning rate scheduler was implemented in order to avoid early convergence on a plateau. If the loss is not decreasing after ten epochs, the learning rate is divided by ten. This allows for a refined gradient descent into the minimum. The initial learning rate was chosen as $5 \times 10^{-4}$.
- **Early Stopping:** Furthermore, early stopping with a patience of twenty epochs based on the validation loss is carried out, which means the stopping criteria is triggered after the validation loss does not improve for twenty epochs. This helps to avoid unnecessary computational costs.
- **Best Model:** In addition to early stopping, the selection of the best model is also done using the validation loss. The selection of the epoch with the lowest validation loss is performed regardless of whether early stopping is triggered or not.

### *VI. Ensemble*

Following the training of multiple ensembles, various strategies were employed to merge their outputs. These strategies are crucial in deciding the final label for each pixel. They include:

- **Majority Voting:** The most frequent label predicted by the models is chosen.
- **Mean of Probabilities:** Here, each model gives a probability for each label. The average of these probabilities is computed for every label, and the label with the highest average probability is selected.
- **Maximum Probability:** The label with the highest probability across all models is selected.

As over twenty different ensembles with different merging strategies were evaluated the listing, let alone the discussion would be out of the scope for this project. Instead, only the most representative ensembles are chosen. It is worth mentioning that only ensembles where each individual model had similar performances were tested.

- **The Coronal Ensemble Mean:** 2D U-Net, 2D Dense-U-Net, 2D Multi-Resolution-U-Net and 2D Res-U-Net trained only on coronal slices combined with mean of probabilities.
- **The Coronal Ensemble Maximum:** Same as the Coronal Ensemble, but combined with the maximum probability.
- **The Coronal Ensemble Majority:** Same as the Coronal Ensemble, but combined with majority voting.
- **The Axial Ensemble:** 2D U-Net, 2D Dense-U-Net, 2D Multi-Resolution-U-Net and 2D Res-U-Net trained only on axial slices combined with mean of Probabilities.
- **The Axial Ensemble Maximum:** Same as the Axial Ensemble, but combined with the maximum probability.
- **The Axial Ensemble Majority:** Same as the Axial Ensemble, but combined with majority voting.
- **The Coronal + Axial Ensemble Mean:** 2D U-Net, 2D Dense-U-Net, 2D Multi-Resolution-U-Net and 2D Res-U-Net trained on both axial and coronal (separately) slices combined with mean of Probabilities.
- **The Coronal + Axial Ensemble Maximum:** Same as the the Coronal + Axial Ensemble Mean, but combined with the maximum probability.
- **The Coronal + Axial Ensemble Majority:** Same as the Coronal + Axial Ensemble Mean, but combined with majority voting.
- **The Multidimensional Ensemble:** 2D Multi-Resolution-U-Net and 2D Dense-U-Net trained only on coronal slices and the 3D U-Net and the 3D SegResNet trained on the full images. All models were combined with the mean of probabilities.

## III. RESULTS

### *I. Metrics*

To evaluate the effectiveness of our approach and analyze the laboratory experiments, we assessed the various cases in our dataset using two key metrics: the Dice Score (DSC), as outlined in Equation 3, and the *Hausdorff Distance (HD)*, which is presented in Equation 4.

$$DSC = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3}$$

$$d_H(X,Y) = \max\{d_{XY}, d_{YX}\} = \max\left\{\max_{x \in X}\min_{y \in Y} d(x,y),\; \max_{y \in Y}\min_{x \in X} d(x,y)\right\} \tag{4}$$

### *II. Evaluation*

In this section, the results of some of the experiments carried out are presented. Table 2 shows the results of the single models, while Table 3 displays the results of the ensemble on the validation set. The qualitative result of the best model can be seen in Figure 2.

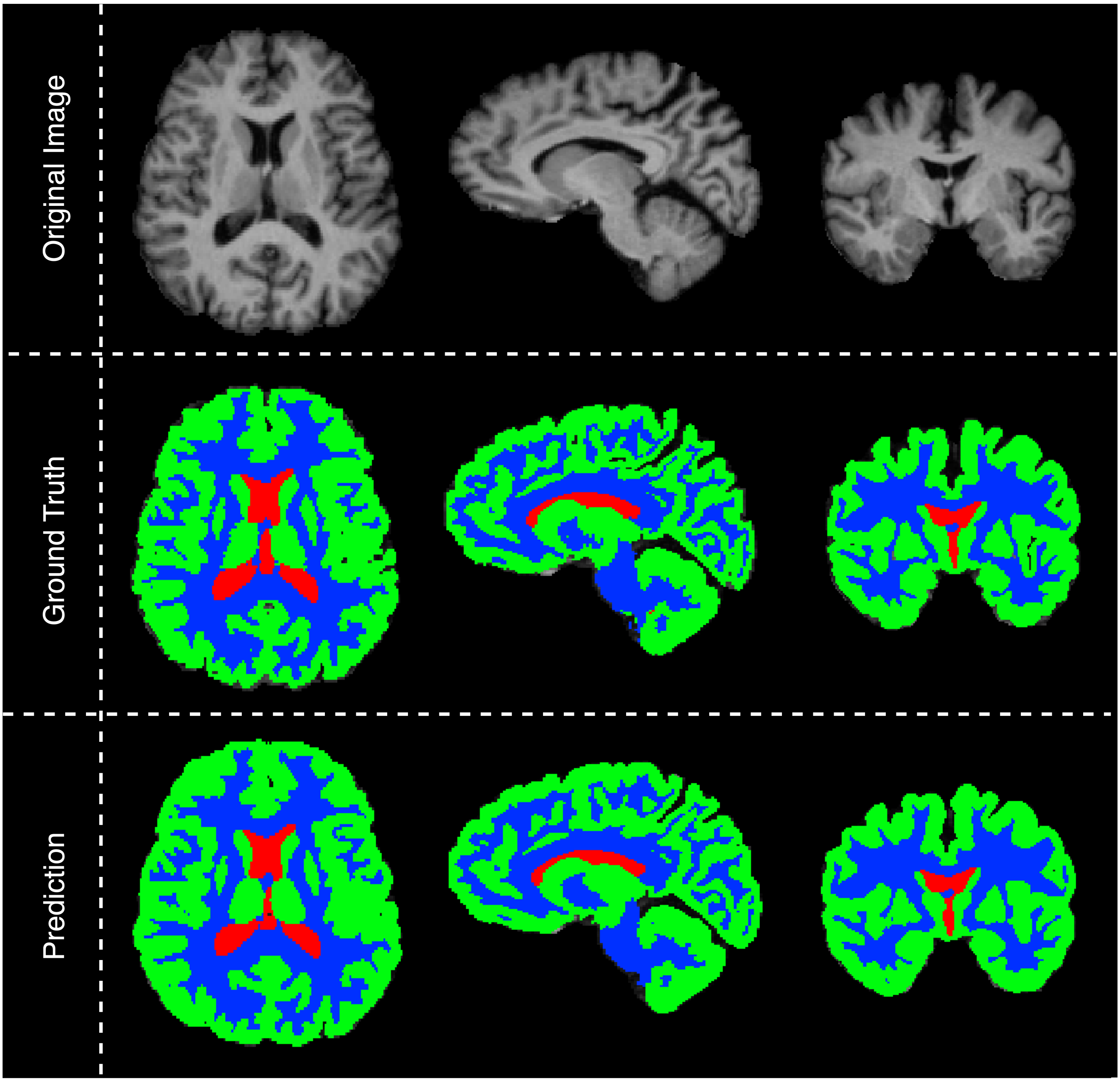

**Fig. 2:** Comparsion of segmentation results of the best performing ensemble: The Multidimensional Ensemble. Displayed are axial slices (left), sagittal slices (middle) and coronal slices (right).

## IV. Discussion

In the following sections, the results will be discussed critically based on quantitative and qualitative results.

### I. Comparison of Axial and Coronal

Here, the results achieved by training a model on coronal slices and by training a model on axial slices will be compared. From Table 2 it can be noted that training a 2D architecture on coronal slices instead of axial slices leads to an improvement across all metrics and tissues. This is especially interesting, as there are four identical model pairs that have been trained exactly the same. However, the differences are relatively small: The best axial model, the Multi-Resolution-U-Net, achieves a mean dice of 0.908 DSC while the best coronal model, the Dense-U-Net, achieves a mean dice of 0.925 DSC. This trend is confirmed in the results of the ensembles (see Table 3). While the coronal ensemble segments the tissues with a mean dice score of 0.925, the axial ensemble is slightly lower with a mean dice score of 0.914. A possible reason for this difference might be the different information present in the different orientations. Another origin of the difference might be that the axial slices were heavily padded, while the coronal slices were not. It would be very interesting to see such a comparison using identical image sizes.

### II. Comparison of 2D and 3D

Training the network using the 3D image leads to a higher dice score in gray and white matter (see Table 2). However, the segmentation results of the cerebrospinal fluid are constant at a dice score of around 0.88. This is rather interesting, as we expected the 3D model to segment the cerebrospinal

**TABLE 2:** SINGLE MODEL RESULTS ON THE VALIDATION SET

| Model | Dice | | | | Hausdorff | | | |
|---|---|---|---|---|---|---|---|---|
| | CSF | GM | WM | Mean | CSF | GM | WM | Mean |
| 2D Coronal U-Net | 0.878 | 0.937 | 0.933 | 0.917 | 39.352 | 11.344 | 10.443 | 20.380 |
| 2D Coronal Dense U-Net | 0.899 | 0.937 | 0.938 | 0.925 | 17.168 | 12.199 | 8.149 | 12.502 |
| 2D Coronal Multi-U-Net | 0.890 | 0.935 | 0.936 | 0.920 | 26.234 | 13.391 | 8.422 | 16.016 |
| 2D Coronal Res-U-Net | 0.882 | 0.931 | 0.931 | 0.915 | 21.894 | 12.000 | 10.905 | 14.933 |
| 2D Axial U-Net | 0.868 | 0.929 | 0.922 | 0.906 | 26.598 | 9.876 | 9.887 | 15.454 |
| 2D Axial Dense-U-Net | 0.868 | 0.920 | 0.920 | 0.902 | 27.137 | 11.281 | 10.580 | 16.333 |
| 2D Axial Multi-U-Net | 0.876 | 0.923 | 0.926 | 0.908 | 30.938 | 10.546 | 9.872 | 17.119 |
| 2D Axial Res-U-Net | 0.866 | 0.925 | 0.921 | 0.904 | 23.733 | 21.277 | 10.113 | 18.375 |
| 2D Seg-Res-Net | 0.877 | 0.933 | 0.935 | 0.915 | **13.540** | 9.977 | **9.449** | **10.989** |
| 3D U-Net | 0.882 | **0.942** | **0.942** | **0.922** | 16.202 | 12.864 | 11.574 | 13.486 |
| 3D Seg-Res-Net | **0.888** | 0.935 | 0.937 | 0.921 | 15.198 | 10.367 | 9.541 | 11.702 |
| SynthSeg | 0.812 | 0.829 | 0.888 | 0.843 | 29.822 | **8.353** | 12.066 | 16.747 |

**TABLE 3:** ENSEMBLE RESULTS ON THE VALIDATION SET

| Model | Dice | | | | Hausdorff | | | |
|---|---|---|---|---|---|---|---|---|
| | CSF | GM | WM | Mean | CSF | GM | WM | Mean |
| The Coronal Ensemble Mean | 0.895 | 0.939 | 0.939 | 0.925 | 18.508 | 9.630 | 7.783 | 11.974 |
| The Coronal Ensemble Maximum | 0.893 | 0.939 | 0.939 | 0.923 | 23.860 | 9.811 | 8.843 | 14.171 |
| The Coronal Ensemble Majority | 0.890 | 0.939 | 0.937 | 0.922 | 19.123 | 11.465 | 7.564 | 12.717 |
| The Axial Ensemble Mean | 0.884 | 0.930 | 0.928 | 0.914 | 17.121 | 10.704 | 9.127 | 12.317 |
| The Axial Ensemble Maximum | 0.881 | 0.930 | 0.927 | 0.913 | 23.055 | 10.655 | 9.782 | 14.498 |
| The Axial Ensemble Majority | 0.877 | 0.930 | 0.925 | 0.911 | 22.114 | 10.946 | 9.277 | 14.112 |
| The Coronal + Axial Mean | 0.897 | 0.939 | 0.938 | 0.925 | 16.410 | 8.902 | 9.095 | 11.469 |
| The Coronal + Axial Maximum | 0.893 | 0.938 | 0.937 | 0.923 | 21.901 | 10.270 | 9.370 | 13.847 |
| The Coronal + Axial Majority | 0.894 | 0.940 | 0.938 | 0.924 | 16.611 | 9.811 | 8.653 | 11.692 |
| The Multidimensional Ensemble Mean | **0.904** | **0.945** | **0.948** | **0.932** | **11.918** | **8.730** | **7.660** | **9.436** |

fluid better. Instead, it seems to reduce errors at the border of gray and white matter. Most surprisingly, it seems to segment different areas of CSF with higher confidence than its 2D counterpart. This conclusion is drawn from the fact that the addition of a 3D model to the ensemble improves the dice scores of the CSF significantly (0.895 vs 0.904).

### *III. Comparison of network Architectures*

Another key factor was selecting the right model architecture. In the experiments, a multitude of different networks were implemented and tested, as is obvious from Table 2. Despite all the different models, it is worth mentioning that all models had comparable results, with the exception of SynthSeg. On the one hand, one might say that the reason for this is the missing information about intensity. On the other hand, the model was only trained for 1500 epochs ( 18 hours) instead of the 100.000 epochs proposed in the original paper. It would be very interesting to see the segmentation results of SynthSeg after a larger number of epochs. Nevertheless, it is very interesting to see that, despite the significantly lower number of epochs, the model was still able to segment the tissue decently. Especially in terms of Hausdorff distance of the gray matter where it achieved lowest over all model with 8.353. This is despite the fact that a dice loss is used instead of a loss taking into account the class imbalance. Furthermore, it is worth mentioning that there is no clear best model architecture based on the network comparison of axial and coronal slices. While Multi-resolution U-Net works best for axial slices, Dense U-Net performs better for coronal slices.

### *IV. Comparison of Single Model and Ensemble*

Judging from both qualitative and quantitative results across all metrics, it can be said that using an ensemble of models can be preferred over a single model for our scenario. Especially, the merging of 2D and 3D models improves results significantly. As visible from Table 3 the combination of the information gained from coronal slices and the spatial information added from a 3D model leads to the best overall performance with a mean dice of 0.932. It is worth mentioning that this combination of 2D and 3D models is the only one to achieve a dice of over 0.9 for the minority class of cerebrospinal fluid.

### *V. Comparison of Merging Strategies*

In total, three different merging strategies for ensembles were tested: majority voting, mean of probabilities, and maximum probability. As visible from Table 3 taking the mean of the probabilities performs slightly better than the others (dice of 0.925 vs 0.913 and 0.924 in coronal ensembles and 0.914 vs 0.911 and 0.913 in axial ensembles). The comparison to majority voting, with a dice of $0,913$ for coronal and 0.911 for axial, shows that the inclusion of the probabilities helps to make the model more robust. However, solely relying on the maximum probability leads to the use of the most-confident model, which could be overconfident instead of providing better segmentation. The key seems to be the comparison between model confidence and the prediction of the majority of the models: the mean of probabilities.

## V. Conclusion

Summing it all up, multiple conclusions can be drawn. Firstly, the different orientations of the 2D slices matter. Secondly, the combination of 2D and 3D helps to segment different different parts more accurately. Thirdly, the ensemble of multiple models through the mean of probabilities leads to the most robust results. Overall, this was a very interesting project for both of us. We had the opportunity to test a multitude of networks, see the importance of preprocessing steps, adapt to entirely different ideas, such as SynthSeg, or see the importance of an ensemble. It is safe to say that we both enjoyed a lot working on this project.

## VI. Hardware Specifications

In this project, we utilized the NVIDIA RTX A6000 GPU, featuring 48 GB of GDDR6 memory and based on NVIDIA's Ampere architecture, for efficient deep learning computations. Complementing this, our general computational tasks were handled by an Intel(R) Xeon(R) Gold 5315Y CPU @ 3.20GHz, equipped with 8 cores and 96 MiB of L3 cache, ensuring robust overall data processing.

## References

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

[2] Sergi Valverde et al. "Comparison of 10 brain tissue segmentation methods using revisited IBSR annotations". In: *Journal of Magnetic Resonance Imaging* 41 (1 Jan. 2015), pp. 93–101. ISSN: 1522-2586. DOI: 10.1002/JMRI.24517. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/jmri.24517%20https://onlinelibrary.wiley.com/doi/abs/10.1002/jmri.24517%20https://onlinelibrary.wiley.com/doi/10.1002/jmri.24517.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV].

[4] Andriy Myronenko. *3D MRI brain tumor segmentation using autoencoder regularization*. 2018. arXiv: 1810.11654 [cs.CV].

[5] M. Jorge Cardoso et al. *MONAI: An open-source framework for deep learning in healthcare*. 2022. arXiv: 2211.02701 [cs.LG].

[6] Benjamin Billot et al. "SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining". In: *Medical Image Analysis* 86 (2023), p. 102789. ISSN: 1361-8415. DOI: 10.1016/j.media.2023.102789.