Is less more? Optimising feature sets for enhanced medical image segmentation

Alessia Bruzzo
Biomedical Engineering
University of Bern
Bern, Switzerland
alessia.bruzzo@students.unibe.ch

Tina Salvisberg
Biomedical Engineering
University of Bern
Bern, Switzerland
tina.salvisberg@students.unibe.ch

Sofia Villar
Biomedical Engineering
University of Bern
Bern, Switzerland
sofia.villar@students.unibe.ch

***Abstract*—This project explores the impact of feature extraction on brain MRI segmentation using a random forest model. Experiments on T1- and T2-weighted MR images tested whether fewer radiomic features could improve segmentation performance.**

The results showed minimal differences between models with reduced and expanded feature sets. Larger brain regions, such as the the grey/white matter, achieved better segmentation, while most structures like the hippocampus were overestimated, limiting the clinical usefulness of the outputs.

The findings suggest that class imbalance and the quality of ground truth data may have affected performance. Future work should focus on improving data quality, addressing class imbalance, and exploring alternative models like U-Net to enhance segmentation accuracy.

***Index Terms*—Magnetic resonance imaging (MRI), Brain segmentation, Feature extraction, Radiomics, Random forest**

I. Introduction

Medical image analysis (MIA) is in the center when diagnosing diseases, planning treatments, and monitoring conditions, particularly in neuroimaging and oncology. A central task in MIA is segmentation, the process of identifying and isolating regions of interest in medical images. Segmentation pipelines typically include pre-processing, feature extraction, and classification steps. Feature extraction transforms image data into numerical descriptors (features) that aid in diagnosis and prediction. Radiomic features, which provide quantitative insights into tissues and lesions, are especially valuable for addressing clinical questions [1].

Texture-based radiomic features have shown effectiveness in MRI analysis [2]. However, using a large number of features can increase model complexity and computational cost, potentially limiting clinical applicability. This raises the question: can a smaller, carefully selected set of features enhance MR brain image segmentation compared to larger, unfiltered feature sets?

To investigate this, advanced radiomic features are integrated into an MIA pipeline using PyRadiomics, a tool for extracting quantitative features from medical images [3]. Key texture features (GLCM contrast, difference entropy, and GLRLM run length non-uniformity) are selected based on recommendations for MR images [2]. The pipeline processes T1-weighted (T1w) and T2-weighted (T2w) MRI scans along with brain region segmentations to train a classification model. Segmentation performance is evaluated using the Dice Similarity Coefficient (DSC) for accuracy and the 95th Percentile Hausdorff Distance (HD95) for boundary alignment. This report outlines the methodology, feature extraction process, and experiments conducted to assess segmentation performance.

II. Methods

A. Data Description

The dataset for this project consists of 3T head MR images from 30 healthy subjects in the Human Connectome Project (HCP) [4]. For each subject, the dataset includes T1-weighted (T1w) and T2-weighted (T2w) volumes, defaced for anonymisation with bias field correction [5]. Available data per subject:

- T1w and T2w volumes (defaced, not skull-stripped, in native space).
- Ground truth label maps and brain masks (FreeSurfer 5.3, no manual annotations) [6].
- Affine transformations aligning images to MNI152 atlas (stored in `./data`).

Non-linear registration (FNIRT) is used for atlas alignment.

B. Pipeline Overview and implementation details

Pre-processing: T1w and T2w images are z-score normalised, and skull stripping is applied using brain masks.

Registration: Nearest-neighbour interpolation is used for ground truth, and linear interpolation for other images.

Feature Extraction: We extended the pipeline with PyRadiomics, now listing:

- **First-Order Features:** Voxel intensity stats (mean, variance) [7].
- **Atlas Coordinates:** Voxel location in the brain [7].
- **Neighborhood Features:** Local texture from neighbouring voxels [7].
- **New Radiomic Features:** GLCM (contrast, entropy) and GLRLM (non-uniformity) [2].

Feature extraction requires a mask to define the region of interest (ROI). The default `label` value is used, with `voxelBased` set to `false`, indicating that features are extracted for entire regions rather than at the voxel level. For each feature and image type, feature images are generated and

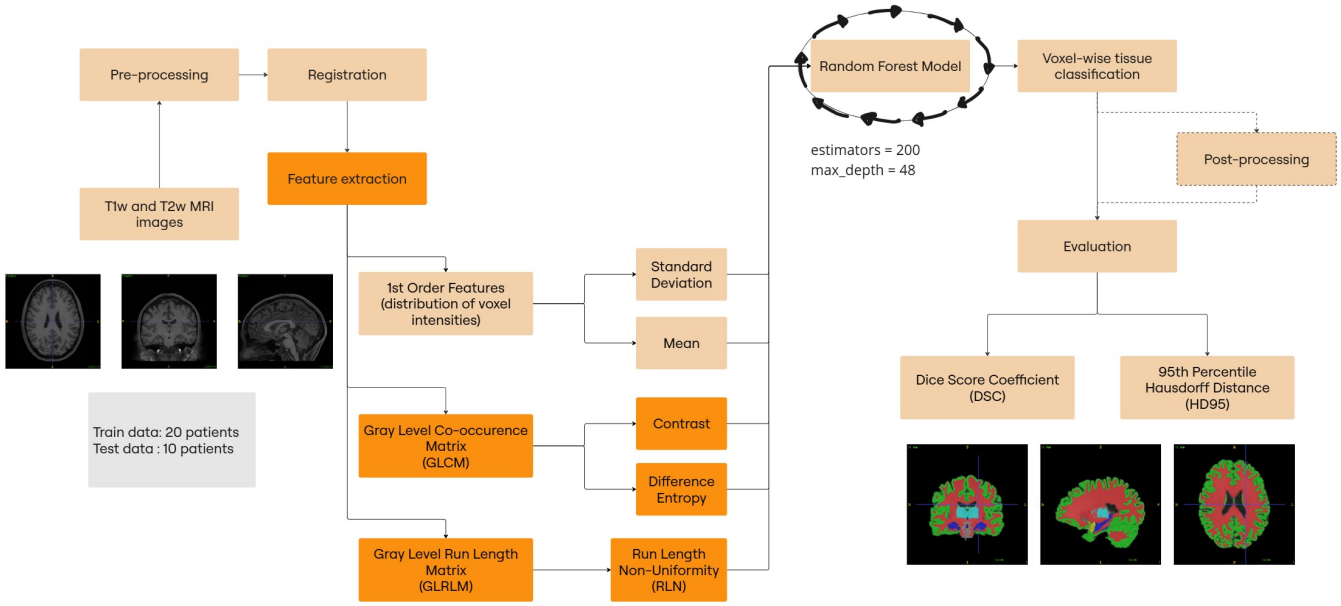

Fig. 1. The pipeline performs brain tissue segmentation through pre-processing, registration, feature extraction, classification, and evaluation (Figure 1). The brain regions of interest are white/grey matter, hippocampus, amygdala, and thalamus, adapted from MIA lab source code [7].

compiled into a matrix for Random Forest classification. These feature images are created for each modality (T1w and T2w) and combined for classification.

Classification: A Random Forest classifier is trained with extracted features and ground truth masks.

Evaluation: Dice Similarity Coefficient (DSC) and 95th Percentile Hausdorff Distance (HD95) measure the segmentation accuracy.

C. Model Training and Testing

A grid search with 5-fold cross-validation is conducted to optimise the Random Forest model by exploring the following parameters:

- **Number of estimators** (`n_estimators`): 50, 100, 150, 200
- **Maximum depth** (`max_depth`): 12, 24, 36, 48
- **Maximum features** (`max_features`): All features, `sqrt`, `log2`

The best-performing combination is: (`max_features=sqrt, n_estimators=200, max_depth=48`).

D. Challenges Addressed

We faced several challenges during the project:

- **Image Size Mismatch**: To be able to visualise the predicted labels as an overlay over the original T1w and T2w images, the images were saved after the registration process. This fixed the size mismatch between the original images and the predictions.
- **Missing Brain Regions**: At first, for some patients, there were no predicted labels for the amygdala and hippocampus. This was solved by changing Random Forest settings of the original pipeline after the grid search.
- **Per-Label Extraction**: It was possible to extract features for each brain label as opposed to the whole brain. Integrating these features into the pipeline after the training process in a meaningful way was outside the project's scope.
- **Combining Features**: We extracted features from T1 and T2 images separately and combined them for classification. Evaluating the effectiveness of this approach could be explored in future work.

Despite these issues, PyRadiomics worked, and the features we needed were extracted.

III. Experiments

We evaluate segmentation performance by testing seven feature combinations, with the original pipeline features always enabled.

A. Experimental Setup

Table I summarises the experimental design, indicating whether radiomic features are enabled (`Y`) or disabled (`N`).

TABLE I
FEATURE COMBINATIONS TESTED IN EACH EXPERIMENT.

Experiment	GLCM Contrast	GLCM Entropy	GLRLM RLNU
Base	N	N	N
Exp. 2	Y	N	N
Exp. 3	N	Y	N
Exp. 4	N	N	Y
Exp. 5	Y	Y	N
Exp. 6	Y	N	Y
Exp. 7	N	Y	Y
Exp. 1	Y	Y	Y

B. Reproducibility Assessment

To evaluate the reproducibility of the segmentation pipeline, we conducted two independent runs using the same experimental setup. The results showed consistent Dice scores across all brain structures, with minimal deviations in certain low-contrast regions. Full results and visual comparisons are provided in Appendix A.

C. Quantitative Results

The segmentation performance is evaluated using Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD95) across different brain regions. The results for all experiments are summarised in the heatmaps shown in Figure 2.

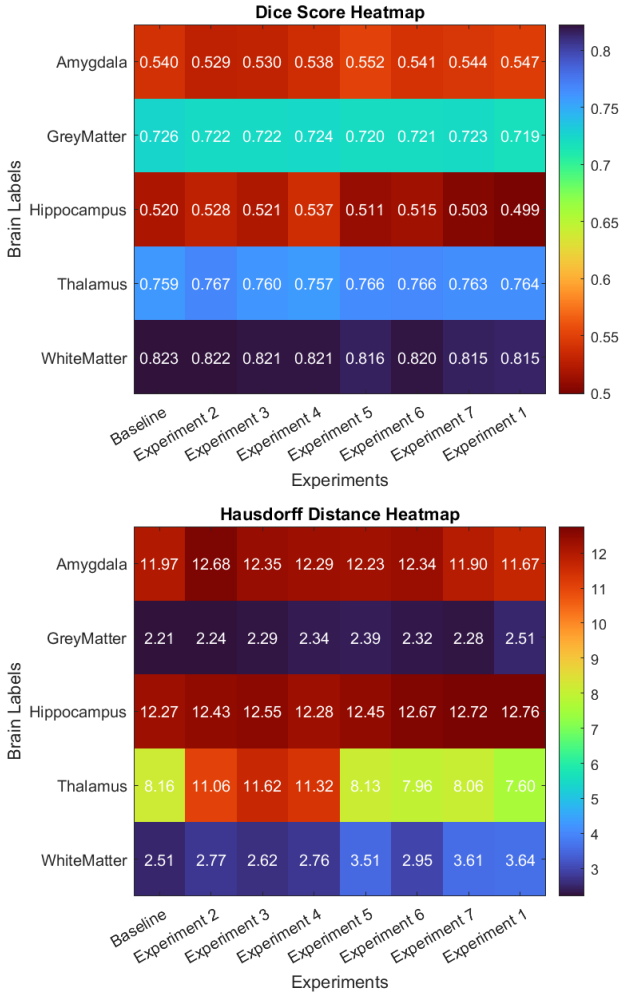

Fig. 2. Quantitative results for all experiments. Top: Dice score heatmap. Bottom: Hausdorff distance heatmap. Colours towards blue, meaning higher Dice scores and lower Hausdorff distances, indicate better performance.

The heatmaps show that higher Dice scores (blue regions) are observed for larger brain structures, such as white and grey matter, and also thalamus. Smaller structures like the hippocampus and amygdala exhibit lower Dice scores. Similarly, lower Hausdorff distances are achieved for larger structures.

For a more detailed breakdown of Dice scores and Hausdorff distances by brain region, box plots for Experiment 3 (Baseline + GLCM entropy) and Experiment 1 (All features) are presented in Appendix A.

D. Qualitative Results

To visually assess segmentation quality, we compare ground truth labels with the outputs of Experiment 1 (all features) and Experiment 3 (baseline + GLCM entropy). Figures 3, 4, and 5 show axial, coronal, and sagittal slices, respectively, for each experiment.

Fig. 3. Axial view comparison of ground truth (left), Experiment 3 (middle), and Experiment 1 (right).

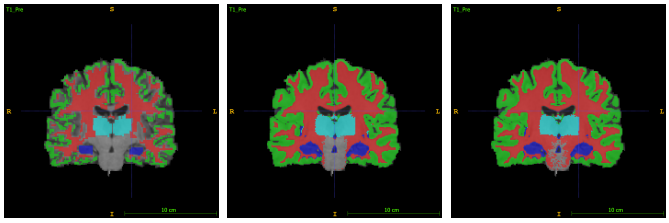

Fig. 4. Coronal view comparison of ground truth (left), Experiment 3 (middle), and Experiment 1 (right).

Fig. 5. Sagittal view comparison of ground truth (left), Experiment 3 (middle), and Experiment 1 (right).

3D Visualization of Errors: The hippocampus segmentation in Experiment 3 shows noticeable overestimation. This excessive segmentation beyond the actual anatomical boundaries is illustrated in Figure 6.

IV. Discussion

The experiments of this project are designed to prove or falsify the initial hypothesis that fewer features result in a better outcome. The results show large differences between the different regions of the brain, but only small differences when changing the number and type of extracted features. The thalamus segmentation has a lower and therefore better

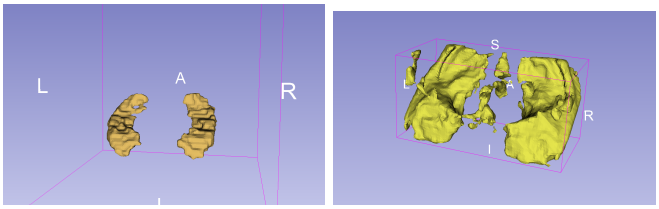

Fig. 6. 3D visualization of hippocampus segmentation: Ground truth (left) and Experiment 3 (right).

Hausdorff distance when more features are used to train the random forest model. This could be an indicator that the initial hypothesis is proven wrong. However, when looking at the individual test patient results on a box plot, this finding is relativised. The variance among the results is high and on an individual patient level, the random forest with more features does not always perform better.

Visual inspection of the results reveals that the models generally seem to overestimate the segmentation. The predicted segmentation volumes are bigger than the ground truth. For some labels such as the hippocampus this leads to unusable results. In other cases, the overestimation of the grey matter seems to compensate for certain shortcomings of the ground truth data, which is not of the highest quality.

The poorer results for smaller structures could be an indication of class imbalance. Another potential issue could be that the feature extraction is performed on the entire brain and not on the individual labels.

Overall, the segmentation results are not satisfactory from an engineering point of view. The clinical value of the segmentation results would in any case need to be discussed with clinical personnel.

V. Conclusion and Future Directions

The results of this project do not clearly answer the question of whether fewer features lead to better segmentation. The experiments show that different parts of the brain had varying results, but adding or removing features only makes small differences. For example, the thalamus segmentation improves slightly when more features are used, but this is not consistent across all patients. Some results have a lot of variation, which means that using more features does not always guarantee better results.

By looking at the segmentation images, we find that the models often overestimate the size of brain structures. This overestimation causes problems for smaller regions like the hippocampus, making the results unrealistic. In some cases, such as grey matter, the overestimation helps correct errors in the ground truth data, but the overall quality of the ground truth is not very high.

Smaller structures tend to have worse results, which may point to an issue of class imbalance. Additionally, extracting features from the whole brain, rather than focusing on specific regions, may be reducing the accuracy of the results.

From a technical perspective, the segmentation results are not as good as expected, and further work is needed to ensure they could be useful for clinical applications. In the future, these steps could improve the project:

- **Improve Ground Truth Data:** Collect better quality ground truth data to improve the training process.
- **Optimise the Model:** Use a wider grid search to find the best parameters for the random forest or switch to a different model, such as U-Net, which may perform better for medical image segmentation.
- **Fix Class Imbalance:** Adjust the loss function or apply methods that focus more on small brain regions to avoid poor results for underrepresented structures.
- **Targeted Feature Extraction:** Perform feature extraction on specific brain regions instead of the entire brain to improve accuracy.
- **Apply Post-Processing:** Use post-processing to refine the segmentation results and correct overestimation issues.
- **Collaborate with Clinicians:** Work closely with clinical experts to validate the segmentation results and ensure that they meet their needs.

Overall, the current results show the potential of the pipeline but highlight the need for further improvements. The next steps should focus on improving the model and ground truth data before running more experiments.

Acknowledgment

We would like to thank our supervisors, Shelley and Amith, for their guidance and support throughout this project. The code and resources developed during this project are available at: https://github.com/tinasalvisberg/mialab.

APPENDIX

REPRODUCIBILITY RESULTS

To verify reproducibility, we compare Dice scores from two independent runs of the segmentation pipeline. Figures 7 and 8 present scatter plots of mean Dice scores for the baseline and Experiment 2.

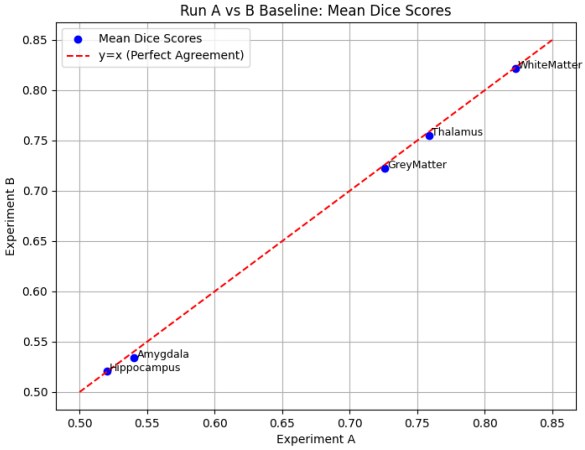

Fig. 7. Comparison of Dice scores between two runs for the baseline experiment. The dashed line represents perfect agreement.

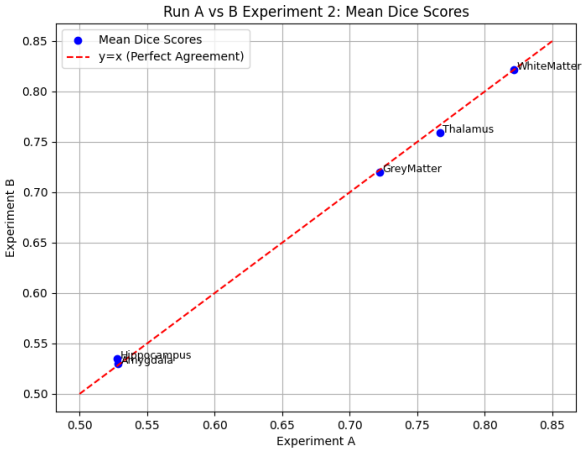

Fig. 8. Comparison of Dice scores between two runs for Experiment 2. The dashed line represents perfect agreement.

The results indicate strong reproducibility across both runs, with minor differences observed. Differences are expected in regions with lower contrast and smaller volumes, like the hippocampus and the amygdala.

DETAILED QUANTITATIVE RESULTS

Figures 9 and 10 illustrate Dice scores and Hausdorff distances by brain region for Experiment 3 (baseline + GLCM entropy) and Experiment 1 (all features).

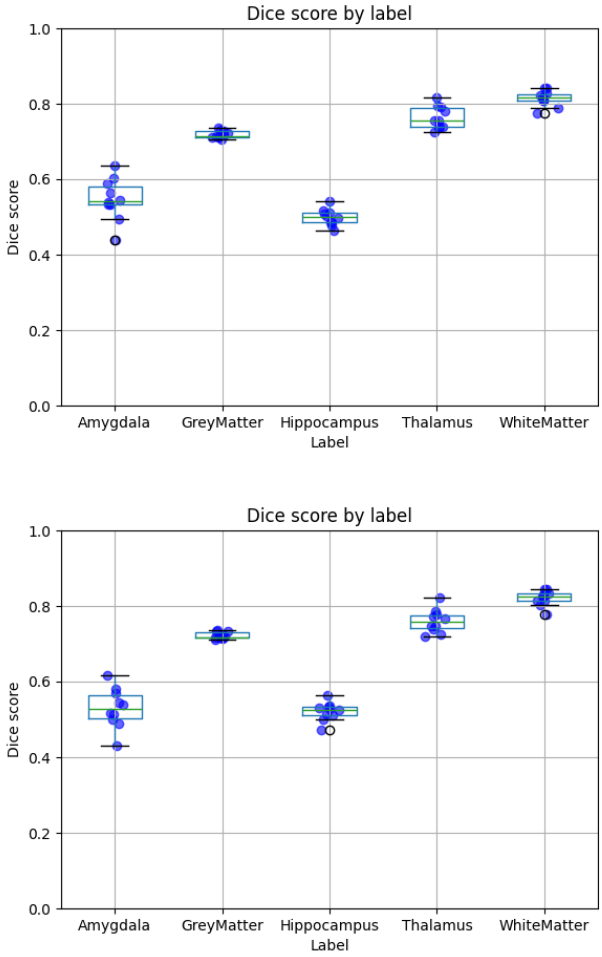

Fig. 9. Dice scores by brain region for Experiment 1 (top) and Experiment 3 (bottom).

REFERENCES

[1] M. E. Mayerhoefer, A. Materka, G. Langs, I. Häggström, P. Szczypiński, P. Gibbs, and G. Cook, "Introduction to radiomics," *Journal of Nuclear Medicine*, vol. 61, no. 4, pp. 488–495, 2020.

[2] J. E. van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, "Radiomics in medical imaging—"how-to" guide and critical reflection," vol. 11, no. 1.

[3] J. J. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts, "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Research*, vol. 77, pp. e104–e107, Oct. 2017.

[4] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, and K. Ugurbil, "The wu-minn human connectome project: An overview," *NeuroImage*, vol. 80, pp. 62–79, 2013. Mapping the Connectome.

[5] M. Milchenko and D. Marcus, "Obscuring surface anatomy in volumetric imaging data," *Neuroinformatics*, vol. 11, pp. 65–75, Sept. 2012.

[6] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A. M. Dale, "Whole brain segmentation," *Neuron*, vol. 33, pp. 341–355, Jan. 2002.

[7] A. Kamath, "Medical image analysis laboratory." original-date: 2024-07-12T07:46:45Z.

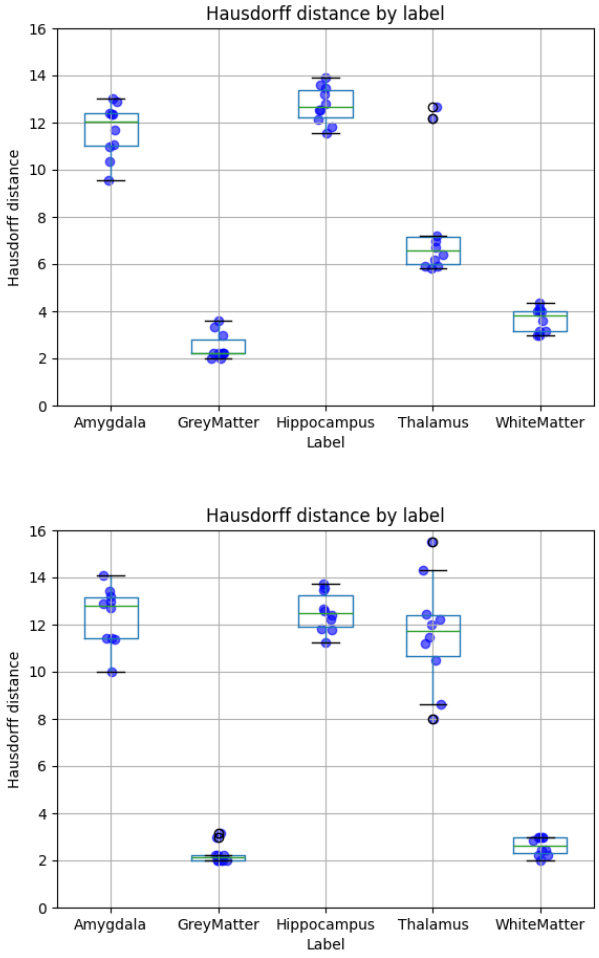

Fig. 10. Hausdorff distances by brain region for Experiment 1 (top) and Experiment 3 (bottom).