Impact of random forest classifier parameters on evaluation metrics

Fabien Jacquat
Faculty of Biomedical Engineering
University of Bern

Pierre Louis Treyer
Faculty of Biomedical Engineering
University of Bern

Guillaume Cheng
Faculty of Biomedical Engineering
University of Bern

***Abstract*—Medical image segmentation of the brain is a critical aspect of modern healthcare, aiding in precise diagnosis and treatment planning for neurological conditions. Various segmentation algorithms, from classical methods to contemporary deep learning approaches, have been employed to delineate brain structures. The evaluation of these segmentation techniques is essential for assessing their accuracy and reliability. Commonly used metrics, including the Dice coefficient, Jaccard index and Hausdorff distance play a crucial role in quantifying the performance of these algorithms. Our test results show that the number of estimators and class weights have an insignificant impact on the evaluation metrics. Upon comparing Random Forest (RF) and Extremely Randomized Trees classifiers (ERT), RF remains stable across various number of estimators and ERT shows errors with low number of estimators, performing worse for the 5% percentile of Hausdorff distance outliers compared to RF. Understanding the nuances and trade-offs associated with these metrics is essential for researchers and practitioners striving to enhance the accuracy and clinical applicability of brain image segmentation methods.**

***Index Terms*—Medical imaging, Brain segmentation, Evaluation metrics, Machine learning, Image processing**

I. Introduction

As the world grapples with an aging population, increased life expectancy, and a rise in the incidence of neurological conditions, the demand for neurological surgery has witnessed a substantial increase. To enhance the implementation of these therapeutic interventions, effective treatment planning is imperative. Magnetic Resonance Imaging (MRI) serves as a prevalent tool for evaluating and determining the optimal treatment course.

Image segmentation is a critical process in medical image analysis, essential for tasks such as tumor detection in brain MRI scans [1]. It involves the partitioning of an image into meaningful regions. After preprocessing to enhance image quality, segmentation algorithms identify regions with distinct intensity characteristics. Binary manual segmentation, a two-class approach, is considered a gold standard by clinicians. However, its drawback lies in its time-consuming nature, requiring clinicians to meticulously analyze each MRI slice. Additionally, manual segmentation is prone to inter-observer variability and subjectivity. Intensity variations, especially at tumor borders, pose challenges, impacting the precision of tumor removal in clinical applications. If the tumor is not adequately removed, there is a risk of residual cancer cells remaining in the patient's body or removing tissue from critical regions.

Furthermore, evaluation metrics are essential for meaningful assessments in the complex landscape of image segmentation and medical image analysis. It serves as quantitative measures to assess the reliability of these segmentation algorithms. However, its application comes with inherent challenges. The clinical relevance of segmentation results may not be fully captured by geometric metrics alone. The ambiguity and heterogeneity often present at tumor borders pose challenges in establishing a universally accepted ground truth. Striking a balance between accuracy and computational efficiency is crucial, as computationally intensive metrics may hinder real-time or large-scale clinical application.

In this study, the Random Forest and Extremely Randomized Tree machine learning algorithms were employed as classifiers for the automated segmentation and labeling of distinct anatomical structures within MRI images of the brain. The segmentation task involved categorizing the brain into five specific labels: amygdala, thalamus, hippocampus, white matter, and grey matter, each corresponding to a distinct anatomical region. To establish a baseline for our experiments, a model was constructed, and subsequent post-processing techniques were applied to refine the segmentation outcomes.

The primary objective of our investigation was to test the hypothesize that alterations in algorithmic parameters, such as maximum depth, number of estimators, and the assignment of label weights, could lead to improvements in specific evaluation metrics.

Modifications to algorithmic parameters were made systematically, with the aim of enhancing metrics relevant to the segmentation task. These parameter adjustments were executed within a well-defined experimental framework, allowing for the systematic exploration of their impact on segmentation accuracy.

II. Materials and methods

A. Medical Image Analysis pipeline

The image analysis pipeline consists of registration of T1- and T2- weighted images, pre-processing including skull stripping, bias field correction, and intensity correction, feature

extraction, classification using algorithms, post-processing for refinement, and a final evaluation step. This structured approach ensures the quality and reliability of image analysis in diverse applications, such as medical diagnostics or scientific research. In the course of our investigation, we exclusively instantiated the classifier, undertook post-processing procedures, and employed evaluation metrics.

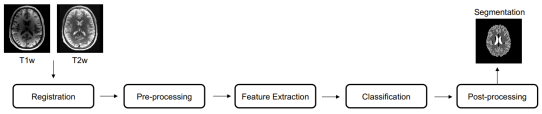

Fig. 1. Pipeline

B. Medical Background

In this paper, the segmentation focuses on five anatomical brain regions: thalamus, white matter, gray matter, amygdala, and hippocampus.

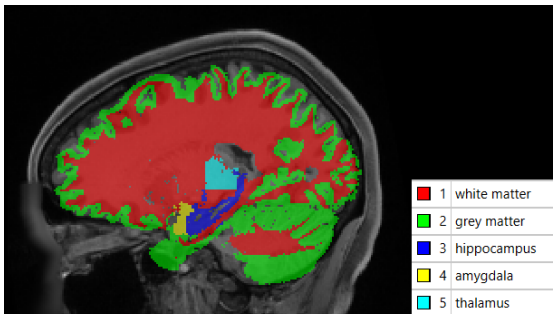

Fig. 2. Brain segmentation.

C. Data

In our image analysis pipeline, we utilized a dataset comprising 30 unrelated healthy subjects from the Human Connectome Project, each equipped with 3T MR T1- and T2-weighted images accompanied by ground truth annotations. For anonymity, images with skulls underwent defacement through blurred face and ear regions. For model training, 20 images were utilized, while 10 were reserved for testing. Each patient's data included a ground truth image, a brain mask, and both T1w and T2w images. The imaging was conducted using a 3T MRI machine. Ground truth labels were assigned using the silver standard and FreeSurfer software.

D. Classifiers

Random Forest and Extremely Randomized Trees (ERT) are learning algorithms commonly employed in machine learning for both classification and regression tasks [3]. Random Forest constructs a multitude of decision trees during training and outputs the average prediction of the individual trees. It introduces randomness by selecting a random subset of features at each node for tree construction, enhancing robustness and mitigating overfitting. On the other hand, ERT takes this concept further by introducing additional randomness in the selection of split points. In ERT, the decision tree nodes are split using random threshold values for each feature, without the need for an exhaustive search. This increased level of randomness often leads to faster training times and can be particularly beneficial when dealing with high-dimensional data.

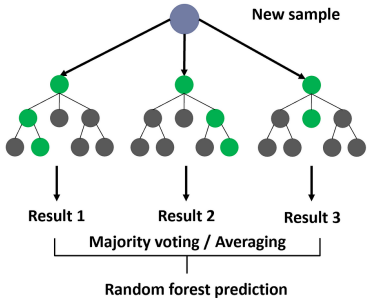

Fig. 3. Structure of Random Forest [3].

E. Post-processing

In the post-processing phase of our study, we employed a connected component algorithm. This algorithm serves to identify and label distinct connected regions within a binary image, where pixels are categorized as meaningful regions [5]. The process involves grouping adjacent pixels with the same label into connected sets, each assigned a unique identifier. This step refines structure, by isolating individual objects and mitigating the impact of spurious noise or small structures in the segmented images.

F. Evaluation metrics

The Dice coefficient, also known as the Sørensen-Dice coefficient, is a similarity metric commonly used in image segmentation and pattern recognition [4]. It ranges from 0 to 1, where 0 indicates no overlap and 1 represents perfect agreement. It is particularly useful in evaluating the accuracy of segmentation algorithms, where A and B might represent the segmented region and the ground truth, respectively.

$$DICE(A,B) = \frac{2 \times |A \cap B|}{|A| + |B|}$$

The Jaccard Index, known as Intersection over Union (IoU), is also a measure of similarity between two sets [4]. It also ranges from 0 to 1, with 0 indicating no overlap and 1 indicating perfect agreement.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

The Hausdorff distance is a metric that quantifies the dissimilarity between two sets by measuring the greatest distance from a point in one set to the nearest point in the other set, and vice versa [4]. It is really sensible to noise, this is why the Hausdorff distance 95th percentile was implemented as well.

It removes the 5th percentile and therefore removes outliers. It is given by the following formula:

$$H(A,B) = \max\left(\sup_{a\in A}\inf_{b\in B} d(a,b), \sup_{b\in B}\inf_{a\in A} d(a,b)\right)$$

G. Experiments

Three experiments which consisted of changing algorithmic parameters were performed:

- Experiment 1: Analysis of class weights
  In order to improve the segmentation of smaller volumes of brain images, their weight was set to 1.5, while bigger volumes were set to 1. In a second trial, a normalization process was performed to return weights between 1 and 1.5 according to the frequency of the labels in the brain image, with the biggest volume having weight 1 and the smallest 1.5. For the last trial, the biggest volume was set to 1 while the other labels were set to their volume ratio with the biggest volume.
- Experiment 2: Analysis of number of estimators
  The number of estimators was initially set to 5 and increased to 300 stepwise.
- Experiment 3: Analysis of max depth
  The max depth was initially set to 5 and increased to 100 stepwise.

If not specified in the experiments, the class weights of all labels were set to 1, the number of estimators to 100 and the max depth to 25. Every test was performed with both RF and ERT, with and without post-processing.

III. Results

Figure 4 compares the impact of RF and ERT in Hausdorff distance per label. Figure 5 shows the impact of postprocessing on a test done with Random Forest, number of estimators = 300, max depth = 25, and no class weights. Figure 6 and 9 show the impact on the average score of the Dice and Jaccard coefficient and Hausdorff distance 100th and 95th percentile given a set of class weights. Figures 7 and 10 shows the impact of the number of estimators and figures 8 and 11 the impact of the max depth. These figures can be found in the appendix.

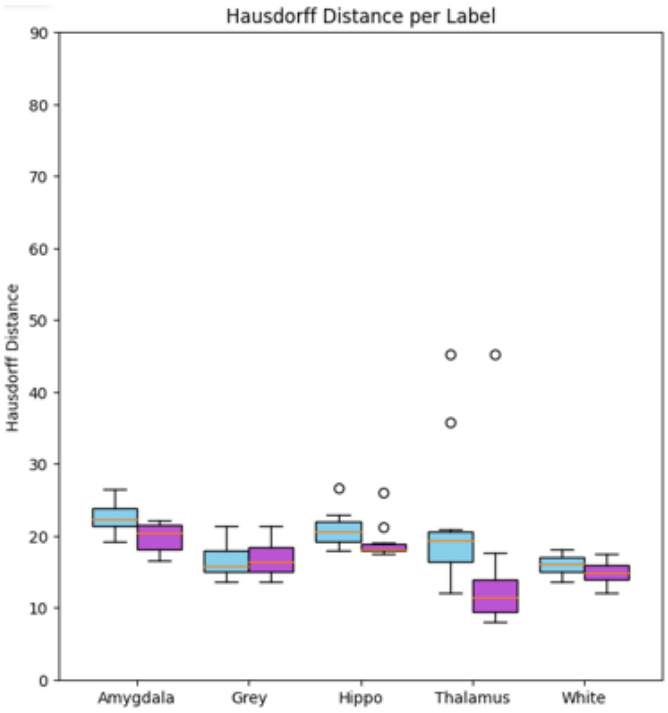

Fig. 4. Comparison RF and ERT (skyblue box : RF, violet box : ERT, number of estimators = 500, max depth = 150, no class weights).

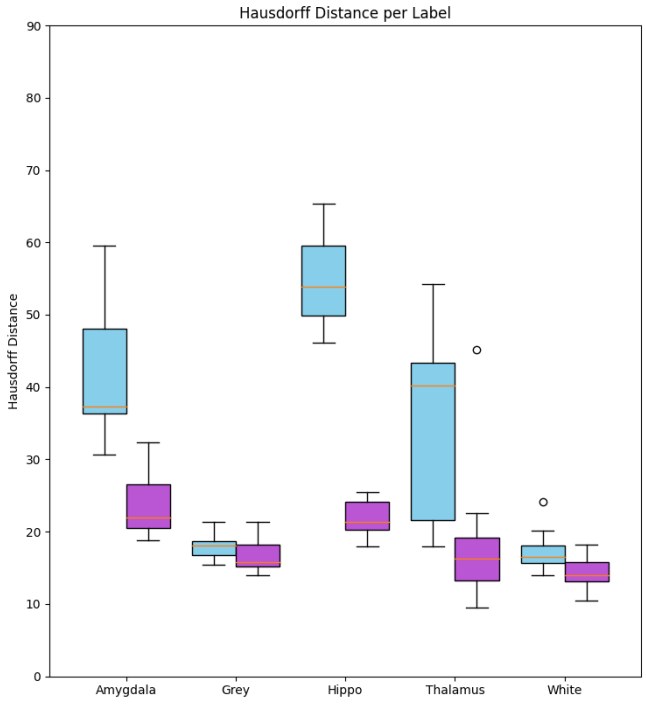

Fig. 5. Visualisation of post-processing impact (skyblue box : without postprocessing, violet box : with postprocessing, number of estimators = 300, max depth = 25, no class weights, Random Forest classifier).

IV. Discussion

In figure 4, ERT has a slightly better performance than RF with Hausdorff distance for smaller volumes, such as the amygdala, hippocampus and thalamus. It is due to the fact that ERT introduces more randomness into individual trees, resulting in lower bias on smaller volumes.

Post-processing with connected components was implemented to eliminate outliers, resulting in a notable improvement, as shown in figure 5. However, we noticed that this only impacts the 5% worst percentile in Hausdorff distance and when analyzing the 95% percentile in Hausdorff distance, the impact of post-processing is removed. This proves that the Hausdorff distances are very sensible to noise introduced in the segmentation.

Upon comparing RF and ERT, the number of estimators (figures 7 and 10) and class weights (figures 6 and 9) were found to have an insignificant impact on the evaluation metrics. For Dice and IoU, there was almost no change in every analysis performed. The Dice score was around 0.7-0.8 for the white, gray matter and thalamus and around 0.4 for the amygdala and hippocampus. The IoU score was around 0.5-0.7 for the white, gray matter and thalamus and around 0.25 for the amygdala and hippocampus.

Additionally, in the analysis of the max depth values for RF and ERT, there is a risk of false segmentation when the max depth values are below 25 (figures 8 and 11).

In figure 7, we observed that RF remains stable across various number of estimators but has a poorer performance for the 5% percentile of Hausdorff distance outliers, for low and high numbers of estimators. In the case of small volumes, ERT shows errors with low numbers of estimators and performs worse for the 5% percentile of Hausdorff distance outliers compared to RF, as shown in figure 10.

For low max depth values, RF demonstrates instability with errors for the 5% percentile of Hausdorff distance outliers, shown in figure 8. In contrast, ERT shows no errors and an overall superior performance compared to RF, as shown in figure 11.

A last test was conducted to assess whether our model exhibits overfitting or not, aiming to explore its limitations in terms of generalization. We experimented with significantly large values for the number of estimators and max depth, but observed no substantial changes. Due to the time-consuming nature of the test, further experiments were not pursued.

V. Conclusion

The tuning of parameters for Random Forest (RF) and Extremely Randomized Trees (ERT) did not yield significant changes in the evaluation metrics. The implementation of post-processing, specifically connected components, resulted in an improvement in Hausdorff distance, although not significantly impacting other metrics. This observation may be attributed to the utilization of 100 estimators and a max depth of 25. To explore more substantial changes, the values of these parameters will be reduced for further observation and analysis.

References

[1] I. Despotović, B. Goossens, and W. Philips. "MRI Segmentation of the Human Brain: Challenges, Methods, and Applications". *National Center of Biotechnology Information*, 1st March 2015.

[2] Dr. R. Yoshua. "Random Forests". *Medium*, 25th March 2023. https://medium.com/@roiyeho/random-forests-98892261dc49.

[3] Emmanuella Budu. "Random Forest vs. Extremely". *Medium*, 15th March 2023. https://www.baeldung.com/cs/random-forest-vs-extremely-randomized-trees.

[4] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection , and tool, "BMC Med. Imaging, col 15, no. 1, 2015.

[5] "Connected Components". January 2024. https://www.sci.unich.it/ francesc/teaching/network/components.html

VI. Appendix : Plots of experiments

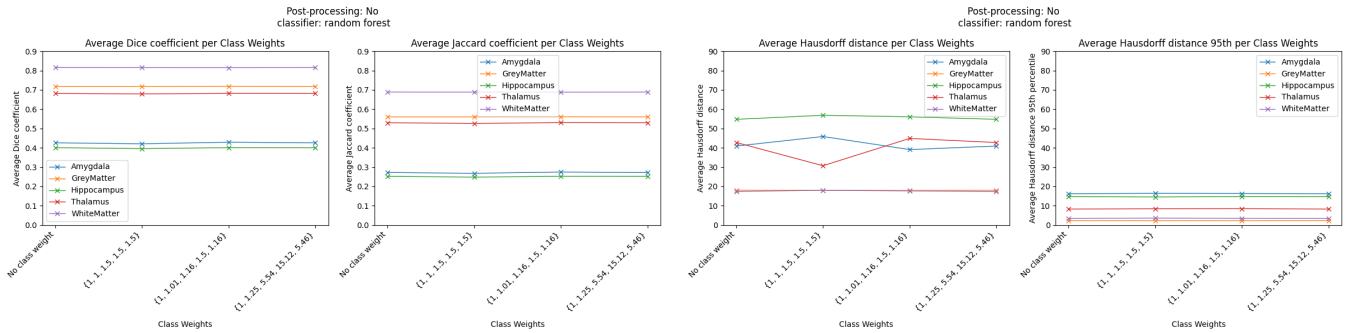

Fig. 6. **Analysis of class weights** without postprocessing and **Random Forest**, order of class weights is {White Matter, Grey Matter, Hippocampus, Amygdala, Thalamus}, from left to right: Average Dice score, Jaccard, Hausdorff distance and Hausdorff distance 95th percentile.

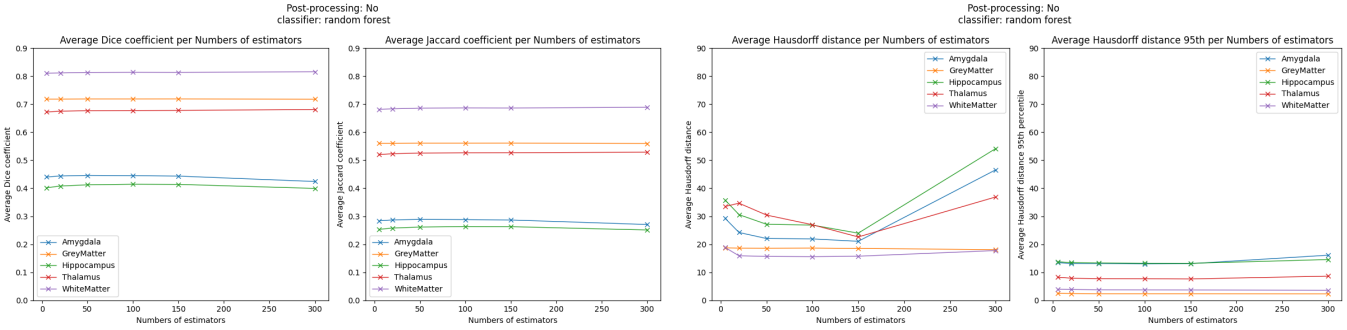

Fig. 7. **Analysis of number of estimators** without postprocessing and **Random Forest**, from left to right: Average Dice score, Jaccard, Hausdorff distance and Hausdorff distance 95th percentile.

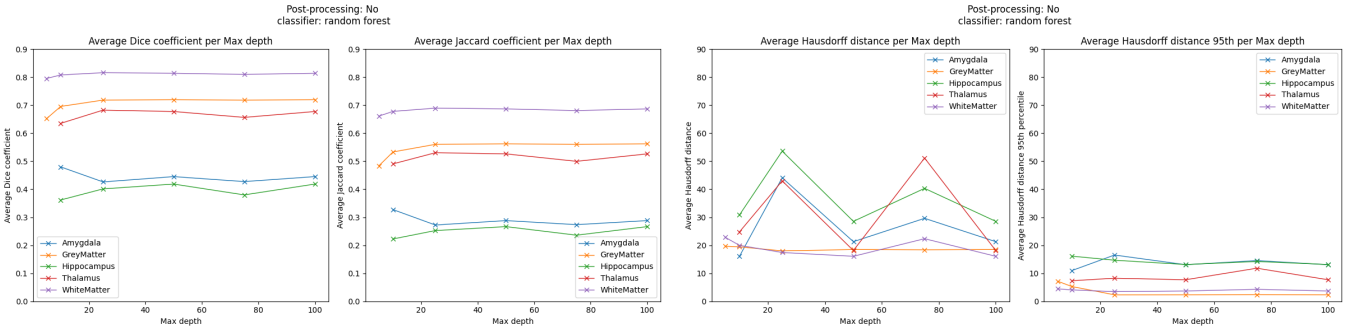

Fig. 8. **Analysis of max depth** without postprocessing and **Random Forest**, from left to right: Average Dice score, Jaccard, Hausdorff distance and Hausdorff distance 95th percentile.

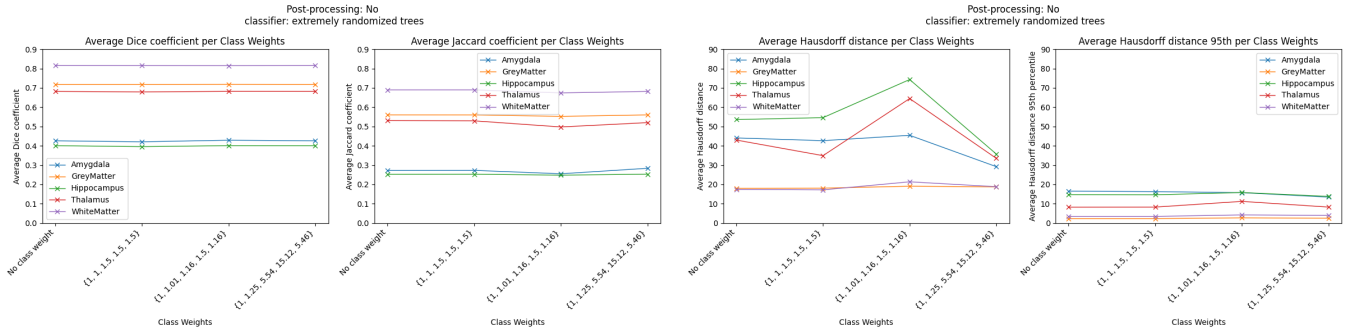

Fig. 9. **Analysis of class weights** without postprocessing and **Extremely Randomized Trees**, order of class weights is {White Matter, Grey Matter, Hippocampus, Amygdala, Thalamus}, from left to right: Average Dice score, Jaccard, Hausdorff distance and Hausdorff distance 95th percentile.

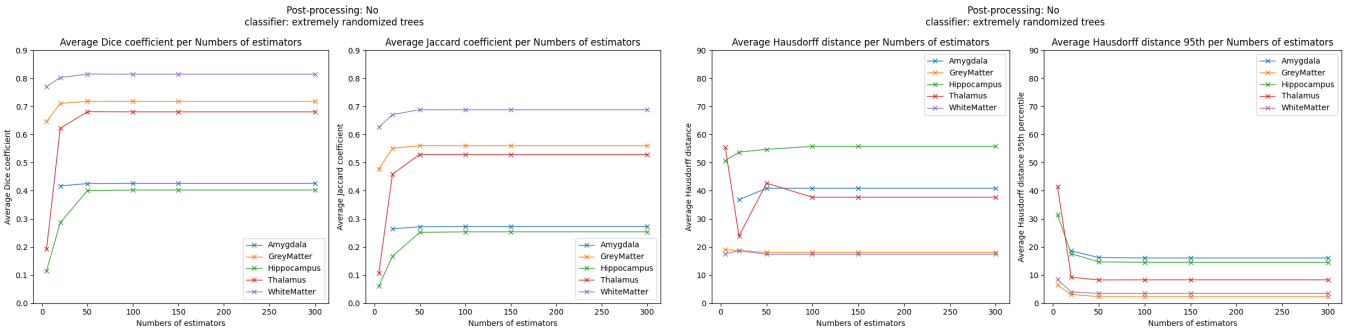

Fig. 10. **Analysis of number of estimators** without postprocessing and **Extremely Randomized Trees**, from left to right: Average Dice score, Jaccard, Hausdorff distance and Hausdorff distance 95th percentile.

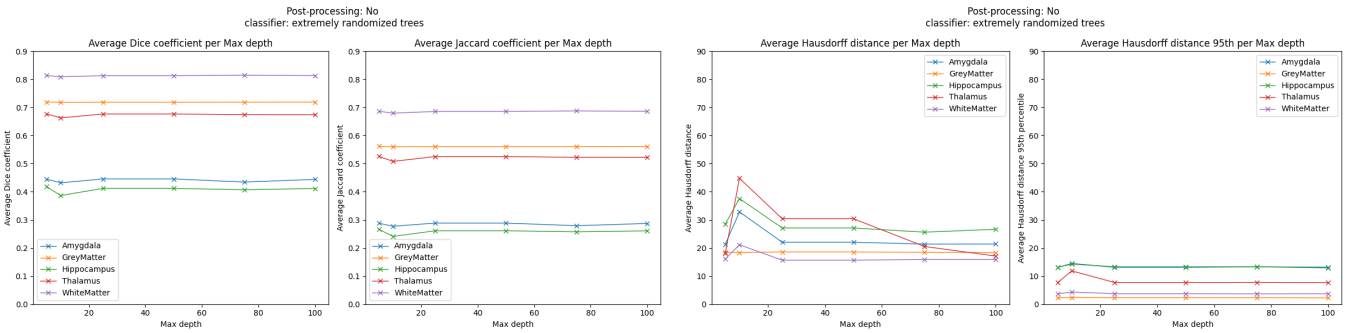

Fig. 11. **Analysis of max depth** without postprocessing and **Extremely Randomized Trees**, from left to right: Average Dice score, Jaccard, Hausdorff distance and Hausdorff distance 95th percentile.