Hypothesis 5: It is favourable to predict multiple labels with the same model

Marta Colmenar Herrera
Medical Image Analysis Laboratory
University of Bern
marta.colmenarherrerastudents.unibe.ch

Filip Krähenbühl
Medical Image Analysis Laboratory
University of Bern
filip.kraehenbuehlstudents.unibe.ch

Mena Lerf
Medical Image Analysis Laboratory
University of Bern
mena.lerfstudents.unibe.ch

***Abstract*—In medical diagnostics, precise brain region delineation from Magnetic Resonance images is vital. We employ a Random Forest classifier for segmentation, comparing binary and multi-label methods. We scrutinise the Dice Similarity Coefficient and Hausdorff Distance metrics using the Human Connectome Project dataset. Additionally, we assess an Overlapping Subtraction technique's impact on segmentation outcomes. We assess visually and quantitatively, emphasising tailored classification strategies for specific brain structures.**

***Index Terms*—image segmentation, binary-label, machine learning**

I. Introduction

In medical diagnostics, brain imaging plays an important role, particularly in identifying debilitating conditions such as Parkinson's disease, Alzheimer's syndrome, and brain tumours. The intricate anatomy of the brain adds complexity to this diagnostic process, with distinct regions playing unique roles in higher-order functions.

The neocortex, consisting of layers in the mammalian cerebral cortex, is crucial for sensory perception, cognition, motor command generation, spatial reasoning, and language [1]. Figure 1 illustrates the neocortex, with the grey area corresponding to this vital region comprising white (WM) and grey matter (GM), where the former constitutes the bulk and the latter envelops it [2].

Additionally, the brain houses smaller yet equally significant structures, such as the amygdala (AM), thalamus (TH), and hippocampus (HP), collectively forming the limbic system [2]. This system primarily processes and regulates emotions, memories, and learning. As depicted in Figure 1, these structures are relatively smaller than the neocortex, emphasising the challenge of precisely identifying the brain's architecture [3].

Magnetic Resonance (MR) imaging is crucial in this scenario, producing high-quality images without subjecting patients to harmful radiation and skull artefacts. Medical doctors conventionally rely on MR images to visually assess and pinpoint the location, shape, and type of brain tumours [4].

Consistent segmentation of brain regions from MR images is imperative for clinical evaluations, surgical planning, and post-surgical assessments [5]. Given that manual segmentation, especially for small brain structures, is laborious and time-consuming, various automated solutions have been explored [6]. Traditional machine learning (ML), specifically classification approaches, leverages multidimensional feature spaces from diverse MR modalities, facilitating the assignment of class labels to differentiating brain regions [5].

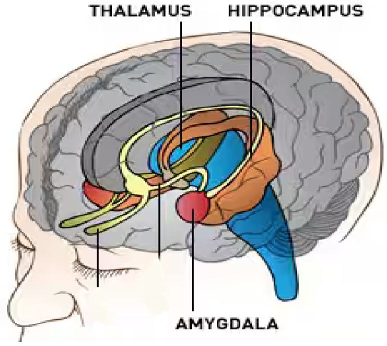


Fig. 1: **Depiction of the individual brain tissues.** In grey: Neo cortex, that consist out of grey and white matter.

In recent work related to brain imaging, MR images feature a Random Forest (RF) classifier, as shown in [7]. This classifier categorises each super-voxel into distinct categories, such as a core tumour, oedema, or healthy tissue. The approach integrates the unsupervised SLIC algorithm for initial tumour localisation with the supervised RF method for subsequent tumour classification. Despite success in classifying larger regions, challenges arise in handling intricate structural boundaries, such as smaller tumour cores, due to the inherent nature of super-voxels encompassing voxels from various tissue types.

When addressing classification challenges, the distinction between single-label and multi-label classification is key in image information retrieval [8]. Our hypothesis posits that predicting multiple labels with distinct binary models proves more favourable than relying on a single multi-label model. This choice is validated through the performance assessment of both binary class and multi-label classification problems.

II. Methodology

A. Data Acquisition

The dataset used for this study is sourced from the Human Connectome Project (HCP) dataset [9]. We selected 30 unrelated healthy subjects from the HCP dataset, dividing them into a training set (20 subjects) and a test set (10 subjects). Three atlas references from the HCP provided labels for brain regions, including white matter, grey matter, amygdala, thalamus, and hippocampus, derived from T1 and T2 MR images.

B. Pymia Framework & MIALab Pipeline

We implemented our ML pipeline using Pymia, an open-source Python package for medical image analysis [10]. The pipeline adheres to MIALab standards and is accessible on the MIALab GitHub repository [11]. It encompasses fundamental medical image segmentation steps, including pre-processing, feature extraction, voxel-wise tissue classification, post-processing, and evaluation.

C. Pre-processing

Pre-processing involves min-max and z-score normalisation to standardise pixel values across MR images. Additionally, we applied skull stripping to isolate relevant brain structures and image registration for spatial alignment using a specified atlas and transformation.

D. Feature Extraction

The feature extraction module enhances input images with informative features after initial pre-processing.

1) Atlas Coordinates Feature Extraction: This module generates a vector image with three components representing physical x, y, z coordinates in millimetres, contributing to spatial analysis and alignment.

2) Neighborhood Feature Extraction: Statistical measures, such as mean and variance, provide a local analysis of image characteristics for detailed information capture.

E. Model

We implemented an RF model from the scikit-learn library [12]. RF is a robust ensemble method that mitigates overfitting and enhances predictive accuracy by aggregating outputs from multiple decision tree classifiers. The classification challenge involves choosing between single-label and multi-label classification, where single-label instances belong to one of two classes, and multi-label instances can be associated with multiple labels [8]. We converted the multi-label prediction model to a binary prediction model by simply masking the pre-processing data for comparative analyses.

F. Parameter Optimization and Model Configuration

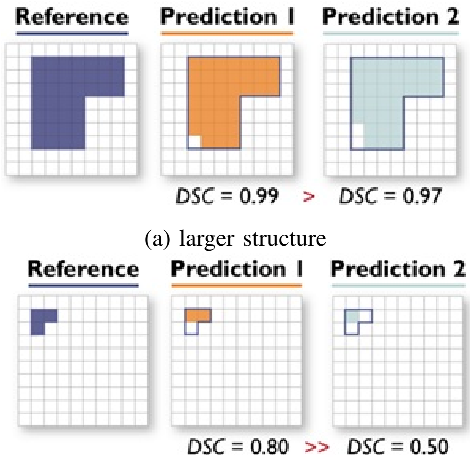
To enhance model performance, we conducted a thorough search for optimal parameters, focusing on the number of estimators and the maximum depth of the tree. Utilising the grid search module [13], we explored parameter combinations, including values such as 50, 100, and 150 for the number of estimators and None, 10, 20, and 30 for the maximum depth. The models subjected to comparison include the baseline multilabel and the binary model multilabel, offering insights into potential performance enhancements. Notably, excluding grid search for multilabel models in the comparison is intentional, as it involves distinct model parameters, making direct comparisons impractical. The focus remains on discerning performance disparities between baseline and binary classifiers.

G. Evaluation Metrics

1) Dice Similarity Coefficient: The Dice Similarity Coefficient (DSC) measures the similarity between two sets commonly employed in medical image analysis [14]. A higher DSC means a better agreement between structures. It is calculated using the formula:

$$DSC = \frac{2 \times \text{Intersection Volume}}{\text{Total Predicted Volume} + \text{Total Reference Volume}}$$

Where the Intersection Volume represents the overlapping volume between predicted and reference segmentations. In our study, significant variations in structure sizes emphasise the importance of careful interpretation, as shown in Figure 2.



(a) larger structure

(b) smaller structure

Fig. 2: **Effect of the structure size on the Dice Similarity Coefficient (DSC).** The predictions of two algorithms (Prediction 1/2) differ in only a single pixel. In the case of a small structure (b), this has a substantial effect on the associated metric value [15].

2) Hausdorff Distance: Hausdorff Distance (HD) is a metric that measures dissimilarity between two sets in medical image analysis. It quantifies the maximum distance between any point in one set and the nearest point in the other [16]. A smaller HD indicates better agreement between segmentations. The formula for HD is:

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

H. Overlapping subtraction

We systematically addressed potential overlaps between different brain structures in binary label predictions through straightforward subtraction; we call this Overlap Subtraction (OS). We sequentially processed images based on voxel count, subtracting overlapping regions from larger structures, ensuring the preservation of smaller structures like the AM, TH, or HP.

RESULTS

In this section, we present the results of our experiments comparing different classifiers and methods for brain image segmentation.

Table I compares the DSC and HD for multilabel classification between the model trained with default parameters and with grid search. Results are shown for different brain regions, including GM, WM, HP, TH and AM. The values are presented as the mean along with the standard deviation.

TABLE I: **Comparison of Dice Score Coefficient (DSC) and Hausdorff Distance (HD) for Multilabel Classification.** ”GS” indicates the results with grid search. Values are presented as mean (standard deviation). GM: Grey Matter, WM: White Matter, HP: Hippocampus, TH: Thalamus, AM: Amygdala.

	DSC		**HD**	
	-	***GS***	**-**	***GS***
GM	0.42(0.02)	0.41(0.02)	3.65(0.29)	3.82(0.35)
WM	0.67(0.03)	0.67(0.03)	4.26(0.38)	4.12(0.42)
HP	0.38(0.02)	0.40(0.02)	19.41(0.80)	14.69(1.37)
TH	0.65(0.03)	0.68(0.03)	39.51(1.02)	19.01(0.98)
AM	0.42(0.04)	0.42(0.04)	15.39(2.76)	13.92(1.30)

Table II further compares the DSC and HD for Binary Classification. Similar to multi-label classification, the table includes GM, WM, HP, TH, and AM results, including results obtained for grid search.

TABLE II: **Comparison of Dice Score Coefficient (DSC) and Hausdorff Distance (HD) for Binary Classifier.** ”GS” indicates the results with grid search. Values are presented as mean (standard deviation). GM: Grey Matter, WM: White Matter, HP: Hippocampus, TH: Thalamus, AM: Amygdala.

	DSC		**HD**	
	-	***GS***	**-**	***GS***
GM	0.52(0.02)	0.51(0.02)	3.13(0.26)	3.54(0.38)
WM	0.75(0.02)	0.73(0.02)	5.67(0.38)	6.06(0.35)
HP	0.38(0.01)	0.37(0.02)	20.85(1.66)	21.81(3.74)
TH	0.66(0.02)	0.66(0.03)	31.17(7.94)	36.92(8.01)
AM	0.42(0.05)	0.41(0.02)	14.55(3.16)	28.13(4.74)

Table III compares the DSC and HD for Binary Classification with and without OS. The table includes GM, WM, HP, TH, and AM results.

Figures 3 and 4 illustrate DSC and HD values distribution for different brain regions using the binary classifier and the binary classifier with overlapping subtraction.

Figure 5 compares MR images’ sagittal views from Ground Truth, Binary Classification, and Binary Classification with Overlap Subtraction.

TABLE III: **Comparison of Dice Score Coefficient (DSC) and Hausdorff Distance (HD) for Binary Classification with and without Overlap Subtraction (OS).** Values are presented as mean (standard deviation). GM: Grey Matter, WM: White Matter, HP: Hippocampus, TH: Thalamus, AM: Amygdala.

	DSC		**HD**	
	-	***OS***	**-**	***OS***
GM	0.51(0.02)	0.34(0.01)	3.54(0.38)	4.78(0.32)
WM	0.73(0.02)	0.58(0.01)	6.06(0.35)	6.17(2.06)
HP	0.37(0.02)	0.33(0.01)	21.81(3.74)	23.02(4.21)
TH	0.66(0.03)	0.64(0.01)	36.92(8.01)	35.17(7.32)
AM	0.41(0.02)	0.38(0.01)	28.13(4.74)	27.72(4.57)

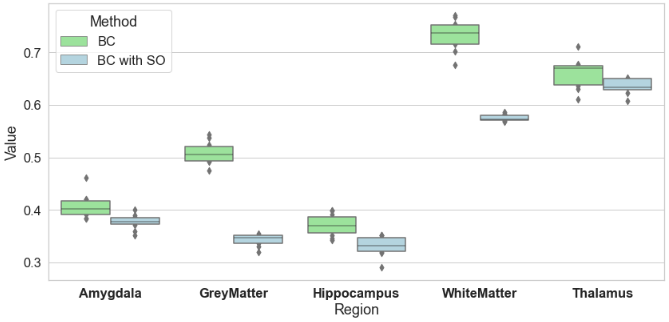


Fig. 3: **Comparison of Dice Score Coefficient (DSC) for Different Brain Regions:** Boxenplot illustrating the distribution of values for brain regions obtained using a binary classifier (BC) and a binary classifier with overlapping subtraction (BC with SO).

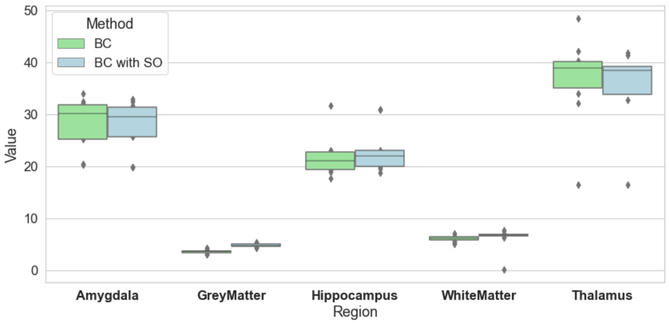
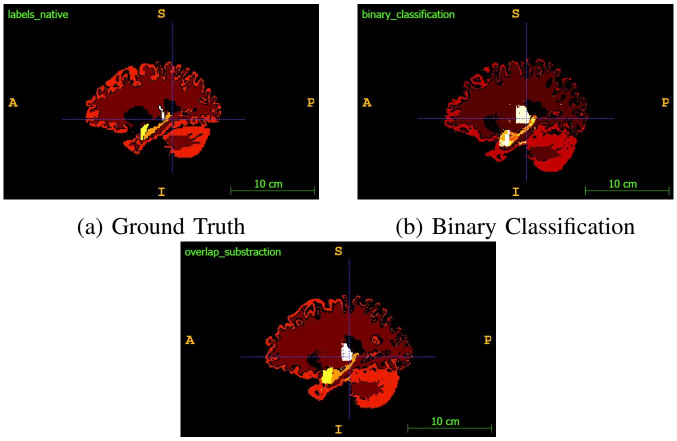


Fig. 4: **Comparison of Hausdorff Distance (HD) for Different Brain Regions:** Boxenplot illustrating the distribution of values for brain regions obtained using a binary classifier (BC) and a binary classifier with overlapping subtraction (BC with SO).



(a) Ground Truth

(b) Binary Classification

(c) Overlap Subtraction

Fig. 5: **Magnetic Resonance (MR) Images Sagittal Views Comparison:** Comparison of MR images sagittal views from Ground Truth (a), Binary Classification (b), and Binary Classification with Overlap Subtraction (c).

Discussion

The study explores the application of an RF classifier in segmenting brain regions from MR images. The results are presented through a comprehensive comparison of DSC and HD metrics for multi-label and binary classification, with and without OS. Additionally, visual comparisons of sagittal MR image views offer qualitative insights. The comparison between results obtained from the grid search optimisation for binary and multi-label classification models reveals performance differences. As hypothesised, the findings align with the notion that predicting multiple labels through separate binary models yields advantages over a unified multilabel model. Specifically, the binary model exhibits robust performance in capturing the broader structures, such as GM and WM, as reflected in comparable or even superior DSC values. Conversely, challenges persist in smaller structures like the HP and TH. The observed pattern underscores the significance of considering the unique characteristics of each brain region and tailoring classification strategies accordingly, supporting the hypothesis that distinct binary models provide a more favourable approach to achieving precise segmentation.

Table III and Figures 3 and 4 highlight the impact of overlap subtraction on segmentation outcomes. The binary classification with OS demonstrates improvements in DSC values for AM and TH, indicating enhanced accuracy in delineating smaller structures. However, trade-offs are observed in other regions, raising the need for tailored strategies based on the characteristics of the target structure.

The qualitative evaluation depicted in Figure 5 complements and visually substantiates the quantitative observations. Ground Truth is the baseline reference against which Binary Classification and Binary Classification with OS reveal their respective segmentation outputs. Notably, the absence of structure subtraction in Binary Classification results in a visual amalgamation, hindering the clear demarcation of structural boundaries. The introduction of OS addresses this issue, enhancing visual clarity and aligning with the quantitative metrics. This improved visual clarity aligns with the quantitative metrics, as evidenced by the reduction in variance between structures for both DSC and HD, as indicated in Figures 3 and 4.

While our study provides valuable insights, limitations exist, such as focusing on a specific dataset, and generalizability to diverse populations warrants further investigation. Future research could explore advanced techniques, such as deep learning architectures, to address complex structural boundaries and enhance segmentation accuracy.

In conclusion, our study contributes to the ongoing discourse on brain region segmentation, emphasising the interplay between classification models and the impact of post-processing techniques like overlap subtraction.

References

[1] H. Raju and P. Tadi, ”Neuroanatomy, Somatosensory Cortex,” *StatPearls*, Nov 7, 2022. https://www.ncbi.nlm.nih.gov/books/NBK555915/

[2] N. Alvarez Toledo, S. Munakomi, and C. J. Prestigiacomo, ”Neuroanatomy, Sylvian Fissure,” Aug 28, 2023. https://www.ncbi.nlm.nih.gov/books/NBK574552/

[3] L. Thau, V. Reddy, and P. Singh, ”Anatomy, Central Nervous System,” *StatPearls*, Oct 10, 2022. https://www.ncbi.nlm.nih.gov/books/NBK542179/

[4] S. Bauer, R. Wiest, L.-P. Nolte, and M. Reyes, ”A survey of MRI-based medical image analysis for brain tumor studies,” *Phys Med Biol*, vol. 58, no. 13, pp. R97–R129, 2013. https://doi.org/10.1088/0031-9155/58/13/R97

[5] A. Fawzi, A. Achuthan, and B. Belaton, ”Brain image segmentation in recent years: A narrative review,” *Brain Sciences*, vol. 11, no. 8, p. 1055, 2021. https://doi.org/10.3390/brainsci11081055

[6] M. K. Singh and K. K. Singh, ”A review of publicly available automatic brain segmentation methodologies, machine learning models, recent advancements, and their comparison,” *Annals of Neurosciences*, vol. 28, no. 1-2, pp. 82–93, 2021.

[7] G. Chen, Q. Li, F. Shi, I. Rekik, and Z. Pan, ”RFDCR: Automated brain lesion segmentation using cascaded random forests with dense conditional random fields,” *NeuroImage*, vol. 211, p. 116620, 2020. https://doi.org/10.1016/j.neuroimage.2020.116620

[8] Q. Dong, X. Zhu, and S. Gong, ”Single-Label Multi-Class Image Classification by Deep Logistic Regression,” Queen Mary University of London and Vision Semantics Ltd., 2021.

[9] D. C. Van Essen et al., ”The WU-Minn Human Connectome Project: An overview,” *NeuroImage*, vol. 80, pp. 62–79, 2013.

[10] A. Jungo, O. Scheidegger, M. Reyes, and F. Balsiger, ”pymia: A Python package for data handling and evaluation in deep learning-based medical image analysis,” *Computer Methods and Programs in Biomedicine*, vol. 198, p. 105796, Jan 2021. http://dx.doi.org/10.1016/j.cmpb.2020.105796

[11] MIALab Contributors, ”MIALab: Medical Image Analysis Laboratory,” https://github.com/ubern-mia/MIALab, accessed: December 6, 2023.

[12] scikit-learn contributors, ”RandomForestClassifier documentation,” https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html, accessed: December 6, 2023.

[13] scikit-learn.org, ”sklearn.model_selection.GridSearchCV,” https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html, accessed: 01.12.2023.

[14] Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. Ecology, 26(3), 297–302. https://doi.org/10.2307/1932409

[15] A. Reinke et al., ”Common limitations of performance metrics in biomedical image analysis,” 2021. https://arxiv.org/abs/2104.05642

[16] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, *Comparing images using the Hausdorff distance, IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850-863, Sept. 1993, https://doi.org/10.1109/34.232073.