Atlas-based Segmentation Consists of a Powerful Baseline for Brain Tissue Segmentation when Compared to a ML-based Approach

Antoine Biebuyck[1], Nathan Hoffman[1], Sandro Scherrer[1]
[1]ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland
Emails: antoine.biebuyck@students.unibe.ch, nathan.hoffman@students.unibe.ch, sandro.scherrer@students.unibe.ch

Abstract—**Alzheimer's disease is on the rise and presents a threat to our aging populations. Reliable biomarkers are required to permit early detection and comprehensive monitoring. One such class of biomarkers arise from volumetric and structural information obtained from magnetic resonance imaging sequences. However, such characterizations rely on precise segmentation which, when performed manually, is too resource-heavy to become adopted in routine clinical workflows. Therefore, automatic segmentation processes are essential to investigate. In this study, we are interested in comparing the performance of an atlas-based method to a machine learning Random Forest model, to evaluate whether the former can match or outperform the latter. Our findings suggest that it depends, in that, the atlas-based method performed better on the smaller brain structures whereas the Random Forest model performed better on the larger brain structures.**

Index Terms—**Alzheimer's disease, Magnetic resonance imaging, Image segmentation, Machine learning, Atlas**

I. Introduction

> What is real? How do you define 'real'? If you're talking about what you can feel, what you can smell, what you can taste and see, then "real" is simply electrical signals interpreted by your brain.
>
> – Morpheus, The Matrix

The human brain's remarkable capacity to navigate and master the intricate complexities of our lives is nothing short of *mind-blowing*—pun intended. Equally thought-provoking, nevertheless, is the question of why it can malfunction.

Dementia is a progressive neurodegenerative disease which results in the loss of cognitive functions such as memory, thought and reasoning skills and negatively interferes with a person's activities of daily life [1]. The World Health Organization estimates that more than 55 million people world-wide are suffering from dementia, with Alzheimer's disease being the most dominant form [2].

While the trigger and driving force behind the progression of Alzheimer's disease still remain uncertain, the neuropathological hallmarks of the disease are the selective loss of cortical neurons within the hippocampus, the temporal lobe and frontal lobe, the gradual loss of synapses, the deposition of amyloid-β forming senile plaques and the presence of intraneuronal neurofibrillary tangles which contain highly phosphorylated microtubule-associated tau proteins. [3,4]

Current assessment of Alzheimer's disease is based on a qualitative psychometric clinical diagnosis relying on criteria established by the Diagnostic and Statistical Manual of Mental Disorders and the National Institute of Neurological and Communicative Disorders and Stroke/Alzheimer's Disease and Related Disorders Association with a definitive diagnosis only confirmed post-mortem through histological staining of brain tissue [4,5]. Considering the wealth of modern scientific advancements, these methods have become rather antiquated and frequently overlook an early diagnosis, crucial for any potential intervention to effectively delay the disease process. Hence the ongoing search for a robust biomarker which can identify, assess and monitor the unfolding of Alzheimer's disease [4].

While molecular biomarkers, requiring a blood or cerebrospinal fluid sample, have been proposed, morphological biomarkers, such as hippocampus volume and cortical thinning obtained from magnetic resonance imaging (MRI) or brain hypometabolism information obtained from positron emission tomography, are also promising [4,5]. To leverage and evaluate these biomarkers, the regions of interests must be extracted from the scan, a process called image segmentation. Manual segmentation is the simplest technique but suffers from drawbacks which include the need for trained professionals, being time- and labor-intensive and operator-dependency rendering it ill-suited for routine clinical practice [6]. Consequently, current research is investigating the use of automated methods.

One such method is atlas-based segmentation. A brain atlas is a map that shows the typical structure of the brain anatomy as it is built from population-based imaging data after registration, warping and overlay to a common reference frame and is based on the assumption that all brains, to a degree of deformation, resemble a prototypical template. Applying this concept of an atlas to segmentation tasks leads to label-based approaches, or probabilistic anatomy maps, in which manual anatomical segmentation labels from a library are mapped to the atlas image [7]. The registration from atlas space to the coordinate system of a newly-obtained brain image then automatically generates with it the tailored segmentation labels.

Of interest in this research study is the performance comparison between an altas-based and machine-learning based segmentation of several brain regions—the white matter, the

All authors declare that they have no conflicts of interest.

gray matter, the hippocampus, the amygdala and the thalamus. Machine learning approaches carry several associated costs such as high computational demands, potentially slow inference times, lack of interpretability and transparency that pose a challenge in clinical adoption, potential to overfit and consequent ineffectiveness on unseen data, and dependency on pre-processing, such as skull stripping, bias field correction and normalization which can introduce additional sources for error [8,9]. Accordingly, the motivation for the juxtaposition in this study is to evaluate whether a so-called more simplistic atlas-based approach can match or even outperform an esoteric and not without drawbacks machine learning Random Forest model, calling into question whether the more sophisticated approach is sufficiently warranted for the task of brain tissue segmentation in this setting.

II. Materials and Methods

In medical image analysis, a typical workflow involves executing several sequential algorithmic steps, a process collectively referred to as a pipeline [8]. The pipeline in this study consists of pre-processing, registration, feature extraction, classification, post-processing and evaluation.

Regarding the experimental data, we utilized 20 training sets and 10 test sets of T1-weighted (T1w) and T2-weighted (T2w) MRI sequences, each accompanied by ground truth segmentation labels, brain masks and an affine transformation matrix to a provided reference atlas.

A. Pre-processing

As an initial step, pre-processing functions to refine and standardize all images in a dataset. Some common actions include background removal, noise reduction, intensity normalization and resampling [9].

In our implementation, the pipeline began with intensity normalization, executed via Z-score normalization. This method computed the mean and standard deviation of pixel intensities, standardizing them to a zero-mean, unit-variance distribution. This step mitigated inconsistencies across the dataset. Next, skull stripping was performed, which applied binary masks to remove non-brain regions. This ensured only relevant anatomical structures were retained.

Additionally, we introduced artificial salt-and-pepper noise. This step was not part of standard pre-processing but was incorporated to test model robustness under lower image quality conditions.

B. Registration

Registration is used to align multiple images to a common reference frame. When only rotations and translations are involved, it is referred to as a rigid transformation while, when scale and skew factors are also included, it is referred to as an affine transformation. Nonlinear, deformable transformations also exist [10]. Since the corresponding affine transformations were provided in the dataset, the registration step was skipped during the Random Forest training.

C. Feature Extraction

Any extractable characteristic or property that describes the underlying medical image and is used in the analysis is coined as a feature. Some examples of image features are intensity, shape, and texture information [8]. In this study, we focused on extracting features that encapsulate spatial, textural, and statistical properties from the T1w and T2w MRI sequences.

Textural features, derived from local intensity distributions, were used to capture patterns and variations within tissue regions. These included metrics such as mean intensity, variance, skewness, and entropy, which are particularly valuable for distinguishing between homogeneous and heterogeneous structures. These features enable the identification of subtle differences in tissue composition. Statistical sampling of voxel intensities provided additional insights by focusing on representative subsets of the data. By generating masks to ensure balanced sampling across different labels, we accounted for variations in tissue representation and minimized the impact of class imbalance. This approach not only improved the quality of the feature set but also enhanced the training process for the Random Forest model.

D. Classification

Classification is the core of the automated segmentation process, where the chosen machine learning algorithm determines the label for each voxel in the image [8]. In this study, a Random Forest classifier was utilized due to its robustness in handling high-dimensional data and its resistance to overfitting.

To optimize the classifier's performance, a grid search was conducted to determine the best hyperparameters for both the original and the simulated noisy datasets. Interestingly, the same optimal hyperparameters were identified for both conditions, indicating the classifier's adaptability to different data qualities. The model achieved the best results with 700 decision trees, a maximum tree depth of 45, and a square root selection of features at each split.

The Random Forest was trained on features extracted from the pre-processed images, and its predictions on test data provided voxel-level probabilities for each label. These outputs formed the basis for the subsequent evaluation step.

E. Post-processing

In this study, post-processing was not applied as it was not necessary to evaluate our hypothesis. The results were directly assessed to determine the segmentation performance, ensuring that the findings reflected the core methodologies without additional modifications.

F. Evaluation

The final step of the pipeline is to quantitatively assess its performance. Such a numeric score is a quick and easy way to compare and contrast different pipelines. Note that qualitative assessment of the resulting segmentation results is also very important to consider when deciding the clinically ”better” result.

Dice similarity coefficients quantify the segmentation quality for each tissue type (e.g., gray matter, white matter) [12]. Evaluator tools calculate subject-wise and aggregated statistics, including mean and standard deviation, across test samples. Results are logged and saved in timestamped directories for analysis.

G. Construction of Atlas Labels

A reference atlas serves as a common space for alignment. All training images were used for generating the atlas labels by registering the segmentation labels of each patient to the provided reference atlas using the given affine transformations. The atlas labels are constructed by averaging all the individual patient labels from the training set and then assigning each voxel the label that occurs most frequently. Additionally, morphological operations, such as median filtering and Opening and Closing, were applied to refine segmentation quality and eliminate small artifacts.

III. Results

TABLE I
Comparison of Dice similarity coefficients between the Random Forest and Atlas-based methods.

Region	Random Forest		Atlas	
Noise	No	Yes	No	Yes
Amygdala	0.47 ± 0.06	0.47 ± 0.06	$\mathbf{0.63 \pm 0.08}$	0.61 ± 0.07
Hippocampus	0.43 ± 0.05	0.43 ± 0.06	0.61 ± 0.10	$\mathbf{0.63 \pm 0.06}$
Thalamus	0.68 ± 0.10	0.68 ± 0.10	$\mathbf{0.79 \pm 0.04}$	0.78 ± 0.05
Grey Matter	0.73 ± 0.01	$\mathbf{0.74 \pm 0.01}$	0.53 ± 0.02	0.52 ± 0.03
White Matter	$\mathbf{0.83 \pm 0.02}$	$\mathbf{0.83 \pm 0.02}$	0.66 ± 0.03	0.66 ± 0.03
Time (s)	197.3 ± 6.3	204.4 ± 7.3	6.0 ± 1.8	$\mathbf{5.8 \pm 0.7}$

These are mean and standard deviation values across all 10 patients from the test data.

The mean Dice similarity coefficients along with their standard deviations for the various brain regions across all 10 test patients using the Random Forest and atlas-based methods can be seen in Table I. The mean computational times along with their standard deviations have also been included.

Generally speaking, for both methods, the obtained mean Dice similarity coefficients remain relatively the same for both the original and simulated noisy images. Therefore, it seems that this simulated noise had very little effect on the performance of the models.

For the amygdala, hippocampus and thalamus, the atlas-based method performed better than the Random Forest model as we obtained, respectively, mean Dice similarity coefficients of 0.63, 0.61 and 0.79 compared to 0.47, 0.43 and 0.68.

For the grey matter and white matter, the Random Forest model performed better than the atlas-based method as we obtained, respectively, mean Dice similarity coefficients of 0.73 and 0.83 compared to 0.53 and 0.66.

Regarding computational time, the atlas-based method was significantly faster, taking 6.0 ± 1.8 seconds without noise and 5.8 ± 0.7 seconds with noise while the Random Forest method required substantially more time, taking 197.3 ± 6.3 seconds without noise and 209.0 ± 7.5 seconds with noise.

IV. Discussion

The Random Forest method demonstrated better performance in segmenting larger structures, such as white matter and grey matter. Features capturing intensity, texture, and spatial distribution play a key role in distinguishing tissue types [15]. Larger structures (i.e. white matter) tend to exhibit more uniform and distinctive intensity patterns, which are more easily captured by statistical features such as mean intensity and variance. Additionally, textural features, which quantify patterns and variations, are more consistent and pronounced in larger regions, making them easier to classify. In contrast, smaller structures (i.e. thalamus) present challenges due to their limited size and less distinct textural properties. The statistical sampling of features in Random Forest models is designed to handle diverse and high-dimensional data, but smaller structures contribute fewer representative samples during the training process. This imbalance can lead to a bias in favor of larger structures that dominate the feature set. Consequently, the Random Forest method's reliance on a multidimensional feature set enables it to perform better in segmenting larger regions where spatial and intensity variations are more readily learned, while smaller regions remain more challenging to classify accurately.

The atlas-based method showed better performance in segmenting smaller structures (i.e. thalamus) which exhibit a high degree of consistency in shape, size, and position across subjects [16]. This anatomical consistency makes these structures particularly well-suited for atlas-based approaches. The compact, ellipsoid shape of the thalamus further simplifies segmentation, reducing the likelihood of errors. Additionally, located centrally in the brain, the thalamus is thereby less affected by global registration misalignments as it experiences smaller shifts compared to more peripheral structures (e.g. assuming a rotational misalignment, due to angular divergence). When the atlas registration is performed accurately, these factors enable reliable identification of the thalamus and other small structures in proximity [17]. The inherent simplicity of the atlas-based method, which does not require a complex classifier, increases its effectiveness for smaller, well-defined regions that are consistently identifiable across subjects. However, the atlas-based method had difficulty segmenting grey matter and also performed more poorly on white matter compared to the Random Forest method. Grey matter presents a unique challenge because it is not as homogeneous as other tissue types, with complex anatomical boundaries and high variability between individuals in terms of size and shape [18]. Similarly, white matter segmentation suffers from low contrast with neighboring tissues and partial volume effects, which blur boundaries and reduce segmentation accuracy.

Additionally, the atlas-based method's reliance on the quality of the initial atlas can limit its performance for both grey and white matter. If the atlas was constructed from a population with different anatomical characteristics than the study dataset, segmentation accuracy may degrade significantly. The widespread and convoluted nature of grey matter, along with the complex organization of white matter tracts, increases sensitivity to registration errors. Misalignments during registration can lead to compounded inaccuracies, particularly for larger, less consistent structures like white matter. Furthermore, the static nature of the atlas-based method prevents it from adapting dynamically to individual anatomical variability, unlike the Random Forest method, which uses a large and diverse training dataset to flexibly address such variations.

In addition to segmentation performance, computational efficiency is a key consideration. The atlas-based method was significantly faster, completing segmentation in seconds compared to the minutes required by the Random Forest method. This efficiency stems from its reliance on predefined atlases and transformations, involving only resampling and registration steps, making it ideal for large datasets. This speed advantage is critical in clinical settings, where rapid processing is often necessary [19]. In contrast, the Random Forest method's extensive prediction steps make it more computationally demanding and less suitable for real-time applications without further optimization.

Both methods demonstrated robustness against salt and pepper noise, maintaining segmentation performance despite image disruptions. While it is expected that the atlas-based method, relying on spatial registration, would handle noise well if the process is robust, the Random Forest method's resilience was more surprising. Notably, the Random Forest performed well on salt and pepper test images even when trained solely on normal data, showcasing its ability to generalize effectively to noisy conditions. This highlights an unexpected strength, particularly valuable in real-world applications where medical images often contain noise from artifacts or patient movement [20].

This study has several limitations that must be considered. First, the dataset used may not fully capture the variability in brain anatomy across different populations, such as those with diverse age groups or pathologies. This limitation could affect the generalizability of the findings to broader clinical settings. Second, the atlas-based method's performance heavily depends on accurate registration, and misalignments can significantly degrade segmentation quality, particularly in regions with high inter-subject variability. To address this, future work could involve constructing multiple atlases tailored to specific populations, allowing better capture of anatomical variations and improving segmentation accuracy. Lastly, while both methods were robust to salt and pepper noise, their resilience to other noise types, such as Gaussian or motion artifacts, was not evaluated, leaving their performance under such conditions uncertain.

V. Conclusion

This study compared the performance of atlas-based and Random Forest methods for brain tissue segmentation, focusing on regions such as the white matter, the grey matter, the thalamus, the amygdala and the hippocampus.

The Random Forest method excelled in segmenting larger structures, such as white and grey matter, leveraging its ability to learn from diverse features capturing intensity, texture, and spatial distribution. Its robustness extended to noisy test data, even when trained solely on clean datasets, showcasing strong generalization capabilities. However, the computational demands of the Random Forest approach may make it less suited for real-time applications.

Conversely, the atlas-based method performed better on smaller, anatomically consistent structures, such as the thalamus, where its reliance on registration-based alignment was highly effective. Despite its limitations in handling larger, variable structures, the atlas-based method demonstrated significant efficiency, completing segmentations in seconds, which can be advantageous in clinical settings.

Therefore, we cannot make the claim from our hypothesis that the atlas hands-down matches or outperforms the Random Forest model. As we've discovered, the answer isn't fixed—it depends. Future work should explore combining these methods to leverage the strengths of each—using the atlas-based approach for precise segmentation of small, consistent regions and Random Forest for larger, heterogeneous structures. Additionally, expanding datasets to capture greater anatomical variability and developing tailored atlases for diverse populations could improve the generalizability and accuracy of segmentation across clinical applications. This hybrid and adaptable approach could enhance automated brain segmentation's effectiveness, addressing the challenges identified in this study.

Acknowledgment

We would like to extend our heartfelt thanks to Shelley Shu and Amith Kamath for their guidance throughout this project and for fostering an engaging and enjoyable learning atmosphere during the lab. We would also like to express our sincere thanks to Professor Mauricio Reyes for imparting us with valuable knowledge throughout the lectures of the Medical Image Analysis course.

References

[1] "What Is Dementia? Symptoms, Types, and Diagnosis," National Institute on Aging. Accessed: Nov. 16, 2024. [Online]. Available: https://www.nia.nih.gov/health/alzheimers-and-dementia/what-dementia-symptoms-types-and-diagnosis

[2] "Dementia." Accessed: Nov. 16, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[3] A. A. Rostagno, "Pathogenesis of Alzheimer's Disease," IJMS, vol. 24, no. 1, p. 107, Dec. 2022, doi: 10.3390/ijms24010107.

[4] K. Gustaw-Rothenberg et al., "Biomarkers in Alzheimer's Disease: Past, Present and Future," Biomarkers Med., vol. 4, no. 1, pp. 15–26, Feb. 2010, doi: 10.2217/bmm.09.86.

[5] W. M. Van Oostveen and E. C. M. De Lange, "Imaging Techniques in Alzheimer's Disease: A Review of Applications in Early Diagnosis and Longitudinal Monitoring," IJMS, vol. 22, no. 4, p. 2110, Feb. 2021, doi: 10.3390/ijms22042110.

[6] B. Foster, U. Bagci, A. Mansoor, Z. Xu, and D. J. Mollura, "A review on segmentation of positron emission tomography images," Computers in Biology and Medicine, vol. 50, pp. 76–96, Jul. 2014, doi: 10.1016/j.compbiomed.2014.04.014.

[7] A. W. Toga and P. M. Thompson, "Chapter 43 - Image Registration and the Construction of Multidimensional Brain Atlases," in Handbook of Medical Image Processing and Analysis (Second Edition), I. N. Bankman, Ed., Academic Press, 2009, pp. 707–724. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123739049500532

[8] C. S. Perone and J. Cohen-Adad, “Promises and limitations of deep learning for medical image segmentation,” J Med Artif Intell, vol. 2, pp. 1–1, Jan. 2019, doi: 10.21037/jmai.2019.01.01.

[9] E. Goceri, “Challenges and Recent Solutions for Image Segmentation in the Era of Deep Learning,” in 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), Istanbul, Turkey: IEEE, Nov. 2019, pp. 1–6. doi: 10.1109/IPTA.2019.8936087.

[10] S. Pereira, A. Pinto, J. Oliveira, A. M. Mendrik, J. H. Correia, and C. A. Silva, “Automatic brain tissue segmentation in MR images using Random Forests and Conditional Random Fields,” Journal of Neuroscience Methods, vol. 270, pp. 111–123, Sep. 2016, doi: 10.1016/j.jneumeth.2016.06.017.

[11] “Medical Image Preprocessing.” Accessed: Nov. 30, 2024. [Online]. Available: https://ch.mathworks.com/help/medical-imaging/ug/overview-medical-image-preprocessing.html

[12] “Medical Image Registration.” Accessed: Nov. 30, 2024. [Online]. Available: https://ch.mathworks.com/help/medical-imaging/ug/medical-image-registration.html

[13] M. Salvi, U. R. Acharya, F. Molinari, and K. M. Meiburger, “The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis,” Computers in Biology and Medicine, vol. 128, p. 104129, Jan. 2021, doi: 10.1016/j.compbiomed.2020.104129.

[14] “Understanding Evaluation Metrics in Medical Image Segmentation,” Medium. Accessed: Nov. 30, 2024. [Online]. Available: https://medium.com/@nghihuynh_37300/understanding-evaluation-metrics-in-medical-image-segmentation-d289a373a3f

[15] C. Liu, R. Zhao, W. Xie, and M. Pang, ”Pathological lung segmentation based on random forest combined with deep model and multi-scale superpixels,” *Neural Processing Letters*, vol. 52, no. 2, pp. 1631–1649, Oct. 2020, doi: 10.1007/s11063-020-10330-8. [Online]. Available: https://doi.org/10.1007/s11063-020-10330-8

[16] N. R. Damle, T. Ikuta, M. John, B. D. Peters, P. DeRosse, A. K. Malhotra, and P. R. Szeszko, ”Relationship among interthalamic adhesion size, thalamic anatomy and neuropsychological functions in healthy volunteers,” *Brain Structure and Function*, vol. 222, no. 5, pp. 2183–2192, Jul. 2017, doi: 10.1007/s00429-016-1334-6. [Online]. Available: https://doi.org/10.1007/s00429-016-1334-6

[17] A. Jakab, R. Blanc, E. L. Berényi, and G. Székely, ”Generation of Individualized Thalamus Target Maps by Using Statistical Shape Models and Thalamocortical Tractography,” *American Journal of Neuroradiology*, vol. 33, no. 11, pp. 2110–2116, Dec. 2012, doi: 10.3174/ajnr.A3140. [Online]. Available: https://www.ajnr.org/content/33/11/2110

[18] E. Radulescu, B. Ganeshan, L. Minati, F. D. C. C. Beacher, M. A. Gray, C. Chatwin, R. C. D. Young, N. A. Harrison, and H. D. Critchley, ”Gray matter textural heterogeneity as a potential in-vivo biomarker of fine structural abnormalities in Asperger syndrome,” *The Pharmacogenomics Journal*, vol. 13, no. 1, pp. 70–79, Feb. 2013, doi: 10.1038/tpj.2012.3. [Online]. Available: https://www.nature.com/articles/tpj20123

[19] C. A. S. J. Gulo, A. C. Sementille, and J. M. R. S. Tavares, ”Techniques of medical image processing and analysis accelerated by high-performance computing: a systematic literature review,” *Journal of Real-Time Image Processing*, vol. 16, no. 6, pp. 1891–1908, Dec. 2019, doi: 10.1007/s11554-017-0734-z. [Online]. Available: https://doi.org/10.1007/s11554-017-0734-z

[20] R. R. Kumar and R. Priyadarshi, ”Denoising and segmentation in medical image analysis: A comprehensive review on machine learning and deep learning approaches,” *Multimedia Tools and Applications*, May 2024, doi: 10.1007/s11042-024-19313-6. [Online]. Available: https://doi.org/10.1007/s11042-024-19313-6