Feature Extraction

Report for Medical Image Analysis Lab

Ackermann Denise Nicole
Universität Bern - BME
Bern, Switzerland
denise.ackermann@students.unibe.ch

Krzeminska-Sciga Alicja Agata
Universität Bern - AIM
Bern, Switzerland
alicja.krzeminska-sciga@students.unibe.ch

Palmgrove Noel Roy
Universität Bern - BME
Bern, Switzerland
noel.palmgrove@students.unibe.ch

Abstract—During the project, the hypothesis that the performance of brain MRI segmentation improves with an increasing number of features was investigated. A provided Medical Image Analysis (MIA) pipeline was used and adapted, focusing on improving the feature extraction. The pipeline features Random Forest as the default classifier. Preprocessing steps were added for image registration, skull stripping, and intensity normalization. Various combinations of features, intensities, gradients, edges, coordinates, and texture were evaluated using the mean Dice score as the performance metric. The results largely confirmed the hypothesis, demonstrating that more features generally enhance segmentation performance. In addition to the Mean Dice score, the Hausdorff distance showed a consistent reduction with introducing more features. Texture and gradient intensity showed the most significant improvement for segmenting smaller complex regions such as the amygdala and hippocampus. However, a detailed analysis shows that the effectiveness of adding features depends on the specific combinations utilized. Improving segmentation performance has significant implications for clinical applications. It increases the accuracy and precision of detecting abnormalities more reliably. In treatment planning for radiotherapy, this could aid in delineating tumours or lesions and monitor their disease progression.

This project's code can be found on GitHub.

I. INTRODUCTION

Medical images generated from a magnetic resonance imaging (MRI) scan display anatomical landmarks, including detailed brain and spinal cord views. By adjusting the repetition time (TR), which is the time between two pulses, and the echo time (TE), the time between a pulse and its returning echo, MRI can enhance different structures while suppressing others through two primary sequences. T1-weighted (T1-w) images often provide detailed anatomical contrast, which facilitates identifying grey and white matter in the brain. Here, structures with a high-fat content will have a hyperintense or bright signal and a hypointense or dark signal for fluid-filled areas in the image. Conversely, T2-weighted (T2-w) images assist in visualizing fluids such as water and blood, which can be effective in detecting oedema, lesions, and fluid-filled areas within tissues [1].

The precise and accurate segmentation of anatomical landmarks and pathological regions of interest is crucial for diagnosis, treatment planning, and monitoring of disease progression. Current segmentation methods in brain MRI rely on manually assigning and defining labels to different brain regions. However, this process is susceptible to inconsistencies among radiologists while increasing time complexity and manual labour of the task [2][3].

Automated segmentation methods based on deep learning architectures have been developed to address these inconsistencies. Among all the features that can be extracted from the image, tissue characteristics can be well-represented using radiomic features [4]. Despite advanced algorithms powered by carefully selected features, the complexity and heterogeneity of medical images still limit the performance of these models [5].

To address these limitations, preprocessing methods such as skull stripping and intensity normalization were implemented to enhance the structure visibility and improve the performance of the segmentation based on features like voxel intensity, edges and texture, so the model can highlight organ abnormalities more effectively [6].

This project investigated whether increasing the number of extracted features could improve segmentation performance by providing more detailed information about image characteristics. This approach revealed additional challenges, including feature correlation, increased computational and time complexity, and the need to balance the quantity and relevance of features.

II. MATERIALS & METHODS

A. Image Dataset

The Human Connectome Project (HCP) dataset was used to train the model [7]. It consists of MR images of the brain from 30 unrelated healthy patients, measured at 3 Tesla. The T1-weighted and T2-weighted images are not skull-stripped but have undergone bias field correction and defacing for anonymity [8]. Ground truth label maps, brain masks and affine transformations are provided in the native subject space. The HCP dataset was previously divided into 20 training patients and 10 test patients to evaluate the model's accuracy.

B. MIA Pipeline

Figure 1 shows the medical image analysis (MIA) pipeline used in this project. Initially, the bias-corrected T1-weighted

Figure 1. MIA pipeline with focus on the feature extraction

and T2-weighted images were registered to the atlas using the given affine transformations. This aligns the images to each other and moves them to the same voxel space. After registration, the exact position in the object space represents the same voxel coordinates. Following registration was the preprocessing step, which involved skull stripping from the relevant brain tissues and voxel-wise intensity normalization to unify the intensities distributions between the images of different patients.

Both registration and preprocessing are crucial to ensure that extracted features are meaningful and precise, enabling reliable feature combination testing - the main focus of this project. During the feature extraction step, features such as intensity, gradients, edges, and texture were computed.

To accelerate the experiments, results from the preprocessing and feature extraction steps were saved as .mha files. This approach eliminated the need to recompute features for each combination, enabling faster experimentation.

Extracted features were used to train a Random Forest (RF) classifier, a supervised machine learning method. The hyperparameters, including the number of estimators and tree depths, were optimized to evaluate performance. The final results were based on 100 estimators and a tree depth of 5.

The last part of the MIA pipeline was post-processing the results. In post-processing, artifacts should be removed to achieve more robust results. However, the post-processing step was outside the scope of this project.

C. Pipeline: Feature Extraction

This paragraph provides a detailed description of the Feature Extraction step in the pipeline, as it is the most important component of this project. The following features were considered:

1) Intensity Features: Intensity features were extracted to provide a direct representation of voxel intensity values. These features capture variations in brightness within the images, offering essential information about tissue properties. It was directly extracted from the images, ensuring the spatial integrity of the original data. Intensity features provided a foundational layer of information, complementing more complex features like texture and gradients.

2) Gradient Intensity Features: Gradient intensity features were extracted to highlight changes in image intensity and identify structural boundaries through the magnitude of the intensity gradient. It reflects the intensity change rate, providing valuable insights into structural edges and transitions. The gradient intensity magnitude was calculated for the T1-weighted and T2-weighted images, ensuring consistency across modalities. Gradient intensity features complemented texture and edge-based features by providing additional information about structural transitions and fine details. This approach provided a wider range of functions and allowed a detailed analysis of anatomical structures and their intensity patterns.

3) Coordinates Features: Coordinate features were extracted to incorporate spatial context into the analysis using atlas-based reference coordinates by aligning images to a standard brain atlas. This ensured spatial consistency across T1-weighted and T2-weighted images, complementing intensity and texture features. This methodology provides a robust framework for integrating spatial and structural information, enabling a detailed examination of anatomical patterns and relationships.

4) Edges Feature: Edge features were extracted using the Sobel filter, a well-established method for identifying intensity gradients, providing insights into image boundaries and structural details. The Sobel operator computed edge features by detecting gradients along each axis and combining them to produce a magnitude map of edges. This method allowed to identify prominent structural details and boundaries within anatomical images. The resulting edge maps retained the spatial integrity of the original images, enabling a comprehensive analysis of structural patterns [9].

5) Texture Features: Texture features were extracted using the Gray Level Co-occurrence Matrix (GLCM), which captures spatial relationships between pixel intensities for a quantitative texture assessment. GLCM captures specific angles and distances of the features and maps them back to the original image dimension. The extraction focused on localized texture through patch-wise computation across T1-weighted and T2-weighted images. Features such as contrast, dissimilarity, and correlation provide insights into texture homogeneity and variability [10]. To extract these radiomic features within the image, the scikit-image library

was used. A small moving window over the entire image was used to assess the region's characteristics.

After extracting all features, every possible configuration was tested as input for the RF classifier. This approach enabled the assessment of the hypothesis regarding the dependency between the number of features used and the performance.

D. Evaluation of feature extraction

The feature extraction pipeline was evaluated using segmentation performance metrics to assess the effectiveness and robustness of the extracted features. The metrics used were designed to capture both spatial conformity and structural accuracy between predicted segmentation and ground truth segmentations:

1) Mean Dice Coefficient: The Mean Dice coefficient was used to quantify the predicted and ground truth segmentation overlap. This metric provides a measure of spatial agreement, with values closer to 1 indicating a higher overlap and better segmentation performance. It is beneficial for assessing the similarity of segmentations for regions of interest, such as anatomical structures.

2) Hausdorff Distance: The Hausdorff Distance measured the largest spatial difference between the predicted and ground truth segments. This metric evaluates the maximum distance between the surfaces of the segmented regions, providing insights into the worst-case segmentation accuracy. It is particularly valuable for identifying outlier errors where segmentations deviate significantly.

Since the Hausdorff distance could not always be computed (e.g., when the structure was not detected), the Dice coefficient was used as the primary metric for experiments and analysis.

III. Results

All experiments were conducted using the MIA pipeline on the HCP dataset. Figure 2 presents the Mean Dice Score (window size of 5 voxels and 100 estimators in the RF classifier) for all tested combinations of features. The scores were evaluated over the individual regions and plotted against the number of used features. It was observed that with the addition of more features, the Mean Dice Score improves. Furthermore, it was noted that the best scores were linked to the test runs where the intensity feature was toggled along with the other features. The best scores represented the larger anatomical regions such as the White and Grey Matter. Looking at the individual regions, the Mean Dice for the smaller regions, such as the Amygdala and Hippocampus, remained relatively constant with no major increase as the number of features increased. However, the larger regions, such as the White and Grey Matter as well as the Thalamus, showed a steady and linear increase with more features added.
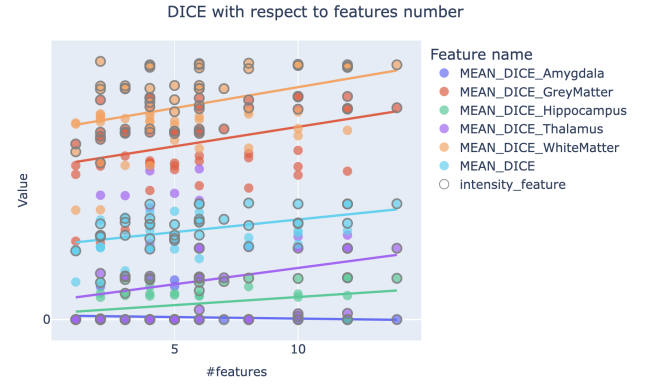

Figure 2. Mean Dice Scores concerning the number of feature extraction methods applied.

To examine the impact of individual features on the Mean Dice Score, feature importance was analyzed. Figure 3 illustrates the importance of each feature in the RF classifier when all possible features were used for training. Figure 4 shows the features importance if the intensity features are excluded.
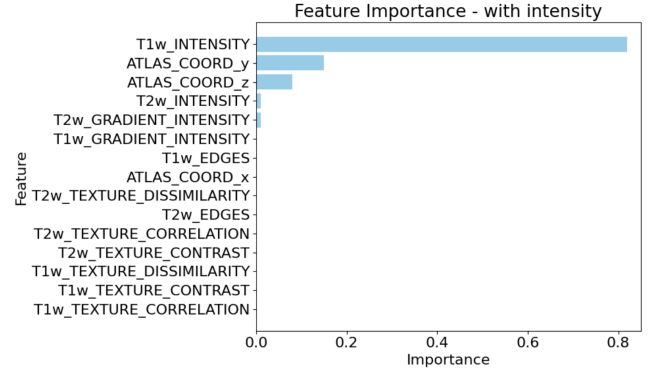

Figure 3. Features importance (all features used).

The comparison of Figure 3 and Figure 4 shows that intensity is the most important feature for the RF classifier when all features are used for training. However, in the experiment where intensities are excluded, gradients contribute most to the final decision. This highlights the features that are most relevant for the model.
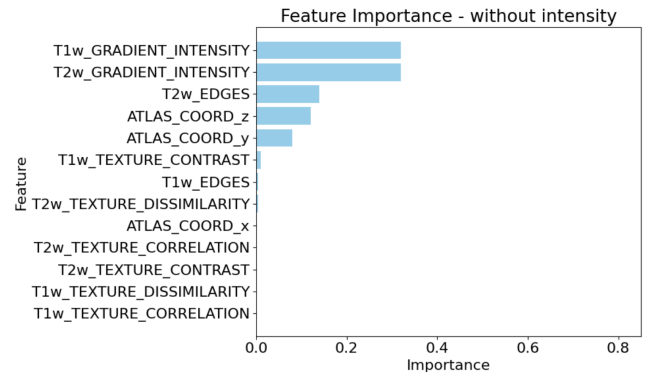

Figure 4. Features importance (intensity features excluded).

IV. Discussion

The results confirm the hypothesis that a larger number of extracted features improves segmentation performance, specifically in the larger structures such as the White and Grey Matter. Experiments to validate it included various feature combinations, such as intensity, texture, gradient, edge, and coordinates. In contrast to larger regions, performance on smaller ones, such as the Amygdala and Hippocampus, shows no significant improvement. This could be due to the complex and inconsistent boundaries, making the segmentation of these regions more challenging.

Another major challenge was noticed when incorporating more features for segmentation: the more methods used, the higher the complexity and computational effort required.

The Feature Extraction was evaluated using the Mean Dice Score and the Hausdorff Distance. These metrics gave information about the spatial coordinates of outliers and edge misalignments within small anatomical landmarks. A consistent increase was noted in Mean Dice Score from 0.21 to 0.39 while increasing the number of Feature Extraction methods. This supports our hypotheses that a larger number of features and their combinations improve Feature Extraction.

Incorporating the T2-weighted features into the pipeline further improved the performance by adding spatial and intensity details. The feature selection based on the feature correlation analysis could help address possible feature redundancy and reduce computational complexity.

Despite improving the Feature Extraction within the MIA pipeline, some limitations were still noticeable within the project. A major issue was the computational complexity when more features were used. This was overcome by utilizing the UBELIX server with high-performance computing. Additionally, using the platform Weights & Biases enabled us to save all the results to keep track on all metrics.

V. Conclusion

This project showed that using more features significantly improved the performance of segmentation of brain MRI. It presented an overall Mean Dice Score of 0.39 (0.77 for the White Matter, 0.65 for the Grey Matter and 0.29 for the Thalamus), while using all 14 feature extraction methods tested in this project. The lower Mean Dice Score for smaller structures indicates a voxel-wise discrepancy between predicted and ground truth segmentation. Our results highlight the importance of using diverse features to capture different anatomical landmarks and their characteristics.

Out of all features, the intensity feature showed the most importance in delineating larger brain regions, such as the White and Grey Matter. Further, adding T2-weighted features improved the analysis by including complementary structural and intensity information. This highlighted the importance of including multimodal feature extraction.

The results confirm the hypothesis that a larger number of features improves the segmentation performance. However, we found a need for feature selection to reduce computational complexity. In future work, optimization of the feature selection and extraction process through deep learning methods, the use of larger data sets, and the expansion to other imaging modalities could be further investigated. It could improve the generalizability of the segmentation and avoid bias.

To conclude, using a wider range of features generally improves segmentation performance, supporting the hypothesis that more features lead to better results. However, we also recognized the potential challenges posed by excessive features, but they can be addressed through feature selection based on importance and correlation analysis. This approach strengthens medical imaging analysis, contributing to more accurate and generalizable pipelines for radiotherapy treatment planning.

References

[1] D. Kawahara and Y. Nagata, "T1-weighted and T2-weighted MRI image synthesis with convolutional generative adversarial networks", en, *Reports of Practical Oncology and Radiotherapy*, vol. 26, no. 1, p. 35, Feb. 2021. DOI: 10.5603/RPOR.a2021.0005. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8086713/ (visited on 10/29/2024).

[2] D. Veiga-Canuto *et al.*, "Comparative Multicentric Evaluation of Inter-Observer Variability in Manual and Automatic Segmentation of Neuroblastic Tumors in Magnetic Resonance Images", en, *Cancers*, vol. 14, no. 15, p. 3648, Jan. 2022, Number: 15 Publisher: Multidisciplinary Digital Publishing Institute, ISSN: 2072-6694. DOI: 10.3390/cancers14153648. [Online]. Available: https://www.mdpi.com/2072-6694/14/15/3648 (visited on 11/25/2024).

[3] E. C. Covert *et al.*, "Intra- and inter-operator variability in MRI-based manual segmentation of HCC lesions and its impact on dosimetry", *EJNMMI Physics*, vol. 9, no. 1, p. 90, Dec. 2022, ISSN: 2197-7364. DOI: 10.1186/s40658-022-00515-6. [Online]. Available: https://doi.org/10.1186/s40658-022-00515-6 (visited on 11/25/2024).

[4] M. E. Mayerhoefer *et al.*, "Introduction to Radiomics", *Journal of Nuclear Medicine*, vol. 61, no. 4, pp. 488–495, Apr. 2020, ISSN: 0161-5505. DOI: 10.2967/jnumed.118.222893. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9374044/ (visited on 11/25/2024).

[5] M. H. Hesamian *et al.*, "Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges", en, *Journal of Digital Imaging*, vol. 32, no. 4, pp. 582–596, Aug. 2019, ISSN: 1618-727X. DOI: 10.1007/s10278-019-00227-x. [Online]. Available: https://doi.org/10.1007/s10278-019-00227-x (visited on 11/25/2024).

[6] J. Singh *et al.*, "A Concise Review Of MRI Feature Extraction And Classification With Kernel Functions", in *2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, Sep. 2023, pp. 1–5. DOI: 10.1109/CISCT57197.2023.10351471. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10351471 (visited on 10/29/2024).

[7] D. C. Van Essen *et al.*, "The WU-Minn Human Connectome Project: An overview", *NeuroImage*, Mapping the Connectome, vol. 80, pp. 62–79, Oct. 2013, ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2013.05.041. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811913005351 (visited on 11/25/2024).

[8] M. Milchenko and D. Marcus, "Obscuring Surface Anatomy in Volumetric Imaging Data", en, *Neuroinformatics*, vol. 11, no. 1, pp. 65–75, Jan. 2013, ISSN: 1559-0089. DOI: 10.1007/s12021-012-9160-3. [Online]. Available: https://doi.org/10.1007/s12021-012-9160-3 (visited on 11/25/2024).

[9] *Texture Analysis of Imaging: What Radiologists Need to Know*, en. DOI: 10.2214/AJR.18.20624. [Online]. Available: https://www.ajronline.org/doi/epdf/10.2214/AJR.18.20624?utm_source=chatgpt.com (visited on 11/25/2024).

[10] A. Kassner and R. E. Thornhill, "Texture Analysis: A Review of Neurologic MR Imaging Applications", en, *American Journal of Neuroradiology*, vol. 31, no. 5, pp. 809–816, May 2010, Publisher: American Journal of Neuroradiology Section: Research Perspectives, ISSN: 0195-6108, 1936-959X. DOI: 10.3174/ajnr.A2061. [Online]. Available: https://www.ajnr.org/content/31/5/809 (visited on 11/25/2024).