Feature Extraction and Selection for Brain Image Segmentation

Abdelrahman Faqieh
Medical Image Analysis Laboratory
University of Bern - AIM

Alexander Aeschbacher
Medical Image Analysis Laboratory
University of Bern - AIM

Zejun Yan
Medical Image Analysis Laboratory
University of Bern - AIM

I. Introduction

Accurate medical image segmentation is a cornerstone of modern clinical workflows, directly impacting critical tasks such as diagnosis, treatment planning, and outcome assessment. For instance, in brain tumor analysis, precise delineation of tumor sub-regions is essential for surgical navigation and radiation therapy [1]. However, reliance on manual segmentation by clinicians creates significant bottlenecks: the process is notoriously time-consuming, subjective, and suffers from high inter-observer variability.

Automated segmentation methods have emerged to address these limitations. Among these, feature-driven approaches, ranging from traditional machine learning to hybrid deep learning systems, remain prevalent, especially in scenarios with limited annotated data. The performance of these methods hinges on a critical, yet often under-examined, step: feature engineering. This process involves the creation, extraction, and selection of quantitative descriptors from images that are most relevant for distinguishing anatomical or pathological structures.

The state of the art in feature engineering can be categorized as follows:

A. Handcrafted Feature Design

Early and many current non-deep learning methods rely on expert-designed features, such as Haralick textures, Gabor filters, and shape descriptors [2]–[4]. While interpretable and grounded in image processing principles, their design requires substantial domain expertise and may not generalize across diverse imaging protocols or pathologies.

B. Automated Feature Selection

To manage the high dimensionality of large feature pools, automated selection techniques like mutual information (MI), recursive feature elimination (RFE), and minimum redundancy maximum relevance (mRMR) are employed [5], [6]. These methods statistically identify informative subsets. However, their fundamental limitation is that they can only select from the features provided to them, and cannot assess the inherent suitability or completeness of the initial feature pool for the spatial segmentation task.

C. Deep Learning-Based Feature Learning

Convolutional Neural Networks (CNNs) [7] represent a paradigm shift by learning hierarchical features directly from data. While powerful, they demand large-scale annotated datasets, pose challenges in interpretability, and their learned features can be difficult to integrate with established clinical knowledge. Hybrid models that fuse handcrafted and learned features attempt to bridge this gap but can inherit the complexity and limitations of both approaches.

Open Question and Contribution:

A persistent, practical hypothesis in feature-driven segmentation is that a larger, more comprehensive feature pool inherently leads to better performance. Furthermore, in practice, Engineers must often choose between extracting features using standard image processing libraries (e.g., scikit-image) or specialised radiomics frameworks (e.g., PyRadiomics), yet clear guidance on the performance implications of this choice remains limited.

In this work, we conduct a controlled, practical investigation to address these questions. We systematically evaluate the impact of feature pool composition and size on segmentation accuracy. Specifically, we test the hypothesis that ”more features are always better” and provide a comparative analysis of features sourced from general-purpose libraries versus dedicated radiomics toolkits [8]. Our goal is to derive evidence-based insights for constructing effective and efficient feature sets, thereby offering practical guidance for developing robust feature-based segmentation pipelines.

II. Materials and Methods

A. Review on MIA pipeline

A structured Feature-Oriented Medical Image Analysis pipeline was applied, incorporating image registration, preprocessing, and feature extraction. Two experimental pathways were evaluated: using all radiomics features and applying MI–based feature selection to assess the benefit of informative feature subsets [10]. Features were classified using a Random Forest configured with 100 decision trees, a maximum depth of 30, and balanced class weights to address class imbalance [9]. Post-processing and segmentation reconstruction followed classification. Segmentation performance was quantitatively

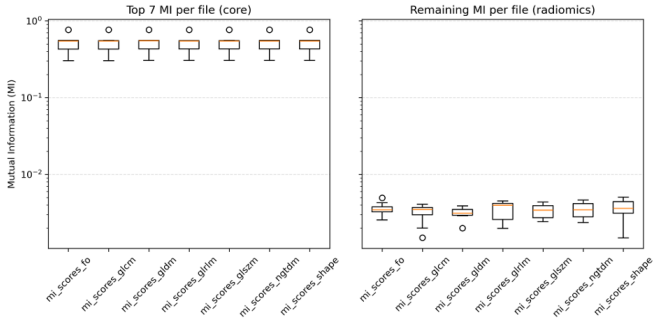


Fig. 1: Distribution of Mutual Information (MI) scores for baseline (core) features and radiomics features.

evaluated using the Dice Similarity Coefficient (DSC) to assess volumetric overlap between predicted and ground-truth segmentations, and the Hausdorff Distance (HD) to measure boundary agreement, with higher DSC and lower HD values indicating better segmentation accuracy.

B. Dataset

Experiments were conducted using 3 Tesla head MR images from 30 unrelated healthy subjects obtained from the Human Connectome Project. For each subject, T1-weighted and T2-weighted image volumes in native T1 space were available, along with corresponding ground-truth label maps, brain masks, and affine atlas transformations. All images underwent bias field correction and anonymisation prior to registration and pre-processing, after which quantitative radiomics features were extracted.

C. Experimental design

The experimental design aimed to assess whether segmentation performance improves through the use of all radiomics features or through a reduced set of informative features. A baseline configuration using voxel-wise features (spatial coordinates, intensity, and gradient intensity) was first established. Two feature representations were then evaluated: the full radiomics feature set and MI–selected feature subsets [11]. Controlled experiments investigated the effect of augmenting the baseline with single or multiple radiomics feature groups, with and without MI-based feature selection. Feature selection and model training were performed exclusively on the training data to avoid information leakage. The dataset was split into 20 training and 10 independent test subjects. All configurations were trained using a Random Forest classifier with optimized hyperparameters, and segmentation performance was evaluated on the independent test set. This design enabled a direct comparison between exhaustive feature inclusion and selective feature augmentation, allowing the stated hypothesis to be rigorously tested.

III. Results

A. Mutual Information Analysis

Figure 1 illustrates the distribution of Mutual Information (MI) scores for the baseline voxel-wise features and the radiomics feature groups. A clear separation is observed between the two feature categories. The baseline (core) features, consisting of spatial coordinates, intensity values, and gradient information, exhibit consistently high MI values across subjects. In contrast, radiomics features from all groups show MI values that are one to two orders of magnitude lower, with distributions concentrated close to zero. This result indicates that voxel-wise features encode substantially more discriminative information with respect to tissue labels than radiomics features. Importantly, this gap persists across all radiomics families, including texture-based and shape descriptors, suggesting that the limitation is not restricted to a specific radiomics group.

These findings provide a quantitative explanation for the degradation in segmentation performance observed when radiomics features are used in isolation or without aggressive feature selection. Low MI values imply weak statistical dependency between radiomics features and voxel-level labels, making them poorly suited for direct voxel-wise classification.

B. Baseline Performance

TABLE I: Dice score (mean $\pm$ standard deviation) for different feature configurations.

Structure	Baseline	Neighborhood	Top-3 (GLCM)
Amygdala	0.38 ± 0.04	0.35 ± 0.04	$\mathbf{0.39 \pm 0.05}$
Hippocampus	0.32 ± 0.02	0.27 ± 0.02	$\mathbf{0.30 \pm 0.02}$
Thalamus	0.60 ± 0.03	0.48 ± 0.03	0.59 ± 0.03
Grey Matter	0.70 ± 0.02	0.70 ± 0.01	$\mathbf{0.71 \pm 0.01}$
White Matter	0.80 ± 0.04	$\mathbf{0.81 \pm 0.03}$	0.80 ± 0.03

TABLE II: Hausdorff distance in mm (mean $\pm$ standard deviation) for different feature configurations.

Structure	Baseline	Neighborhood	Top-3 (GLCM)
Amygdala	12.1 ± 1.3	16.5 ± 4.8	$\mathbf{11.9 \pm 1.2}$
Hippocampus	13.8 ± 0.9	24.1 ± 6.1	$\mathbf{15.4 \pm 1.1}$
Thalamus	10.4 ± 2.0	35.7 ± 6.6	13.2 ± 3.1
Grey Matter	2.38 ± 0.63	2.43 ± 0.36	$\mathbf{2.36 \pm 0.46}$
White Matter	2.89 ± 0.75	3.51 ± 1.03	3.61 ± 1.13

The baseline configuration used voxel-wise machine learning features, including spatial coordinates, intensity values, and gradient intensity features. This setup achieved strong segmentation performance for large anatomical structures such as White Matter and Grey Matter, with Dice scores of 0.80 ± 0.04 and 0.70 ± 0.02, respectively (Table I). Boundary accuracy for these structures was also high, reflected by low Hausdorff distances (Table II).

In contrast, segmentation performance was substantially lower for smaller subcortical structures. The Hippocampus and Amygdala achieved Dice scores of only 0.32 ± 0.02 and 0.38 ± 0.04, respectively, accompanied by comparatively large Hausdorff distances. This behavior highlights the impact of class imbalance and the inherent difficulty of accurately segmenting small, spatially compact brain regions using voxel-level classifiers.

C. Neighborhood-Based Features

In this experiment, the baseline voxel-wise feature set was augmented with a comprehensive set of neighborhood-based radiomics features. For each voxel, statistical descriptors were extracted from a local $3 \times 3 \times 3$ neighborhood, including first-order statistics such as mean, variance, skewness, entropy, energy, and many more totaling to over 35 features.

The resulting feature representation substantially increased the dimensionality of the input space while retaining the original baseline features. Despite this extensive feature augmentation, segmentation performance performed the worse. As shown in Table I, Dice scores for small subcortical structures such as the Hippocampus and Amygdala decreased to 0.27 ± 0.02 and 0.35 ± 0.04, respectively. In addition, Hausdorff distances increased markedly (Table II), indicating degraded boundary localization and reduced spatial consistency.

Mutual Information analysis further revealed that the majority of the added neighborhood-based radiomics features exhibited significantly lower relevance scores compared to the baseline voxel-wise features. This suggests that, when densely extracted at the voxel level without selection, neighborhood radiomics features contribute limited discriminative information and may introduce noise and redundancy. Consequently, the increased feature dimensionality leads to overfitting and reduced generalization, particularly for small anatomical structures.

D. Baseline Combined with MI-selected Radiomics Features

To reduce feature redundancy, Mutual Information-based feature selection was applied to the radiomics features, and only the top-K features were retained and combined with the baseline feature set. Three values of K were evaluated: $K = 3$, $K = 5$, and $K = 10$. Across all radiomics groups, the best performance was consistently achieved with $K = 3$. In particular, the combination of baseline features with the top three GLCM features yielded the highest Dice scores for small structures, with the Amygdala and Hippocampus reaching 0.39 ± 0.05 and 0.30 ± 0.02, respectively (Table I). Importantly, this improvement was accompanied by reduced Hausdorff distances compared to both the baseline and the full radiomics configuration, indicating improved boundary consistency (Table II). Increasing K beyond three did not lead to further improvements and often resulted in slightly reduced Dice scores and increased boundary errors.

Figure 2 illustrates the effect of MI–based feature selection when augmenting the baseline with GLCM radiomics features. Across all anatomical structures, the baseline configuration already exhibits stable Dice score distributions, particularly for large tissues such as White Matter and Grey Matter. For small structures, notably the Hippocampus and Amygdala, the use of all GLCM features does not improve performance and in several cases increases variance. In contrast, selecting a small number of informative GLCM features ($K = 3$) leads to modest but consistent improvements in Dice scores and reduced variability. This effect is most pronounced for the Amygdala, where both the median Dice score and interquartile

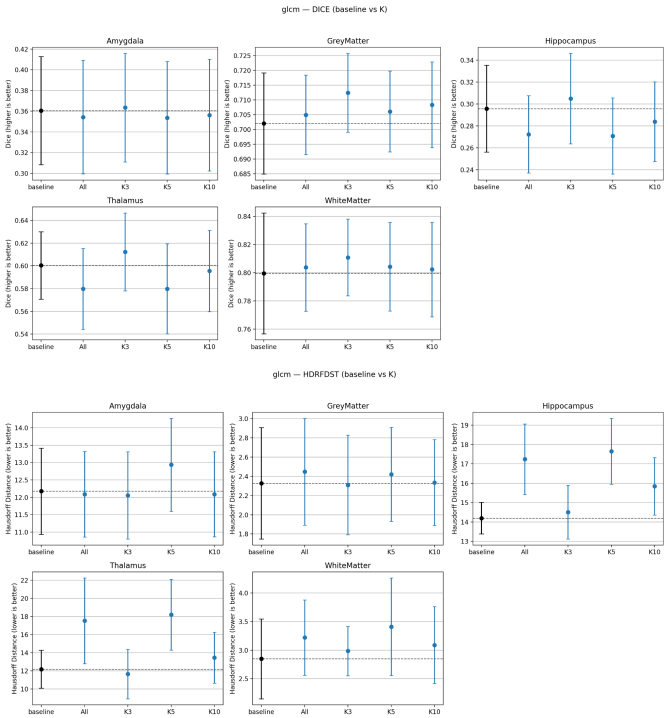


Fig. 2: Performance comparison of GLCM-based feature augmentation using MI selection. Top: Dice scores. Bottom: Corresponding Hausdorff distance distributions.

range improve relative to the baseline. The Hausdorff distance distributions further support this observation.

Using all GLCM features substantially increases boundary error and variance, indicating unstable and spatially inconsistent predictions. The $K = 3$ configuration achieves lower or comparable Hausdorff distances relative to the baseline, while larger values of K again degrade performance. Overall, these results demonstrate that while GLCM features can provide complementary information, their benefit is limited to a very small, carefully selected subset. Including additional radiomics features beyond this point introduces redundancy and noise, reducing both accuracy and robustness.

E. Other Experiments

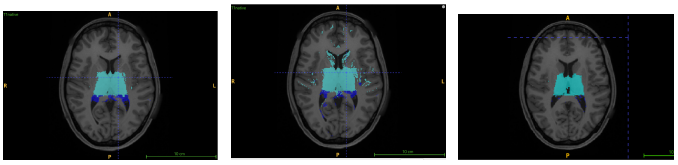
Combining multiple radiomics feature groups did not improve segmentation performance and instead reduced accuracy and increased variability, particularly for small structures, likely due to overfitting from increased feature dimensionality. Applying Mutual Information–based selection mitigated this effect; however, selecting more features (K = 5) did not outperform a smaller subset (K = 3). These results highlight the necessity of aggressive feature selection and demonstrate that adding more features does not inherently improve segmentation accuracy.

F. Feature Complementarity Analysis

To verify that the selected baseline feature families provide non-identical information, a group-level mean absolute

TABLE III: Group-level mean absolute Pearson correlation between baseline feature groups. Low off-diagonal values indicate that different feature families capture complementary information, while larger diagonal values indicate intra-group redundancy.

	Coordinates	**Gradient**	**Intensity**
Coordinates	0.346	0.065	0.086
Gradient	0.065	0.901	0.209
Intensity	0.086	0.209	0.669



(a) Baseline (b) Neighborhood (c) Top-3 GLCM

Fig. 3: Qualitative comparison of Thalamus segmentation. The baseline model produces spatially consistent predictions. Dense neighborhood-based feature augmentation introduces spurious peripheral misclassifications. The proposed Top-3 GLCM configuration reduces false positives and improves spatial coherence, yielding the most anatomically plausible segmentation.

Pearson correlation matrix across the baseline feature groups was calculated (Table III). Off-diagonal correlations were low (Coordinates–Gradient: 0.065, Coordinates–Intensity: 0.086, Gradient–Intensity: 0.209), indicating that spatial location, intensity, and gradient cues are largely complementary.In contrast, within-group correlations were substantially higher, especially for gradient features (0.901) and intensity features (0.669), suggesting redundancy inside each family. This supports the use of Mutual Information selection and limited feature augmentation, since adding features from a different group is more likely to contribute new information than adding many highly correlated features from the same group.

G. Qualitative Analysis

Visual inspection of the segmentation outputs supported the quantitative findings. Baseline features produced stable and anatomically plausible segmentations for large tissue classes. The Top-3 (Baseline + GLCM) configuration yielded visibly improved delineation of small structures, particularly the Hippocampus and Amygdala. In contrast, models using neighborhood-based features exhibited increased boundary noise and occasional over-segmentation artifacts, consistent with the elevated Hausdorff distances observed quantitatively.

IV. Discussion

This study focused on voxel-wise classification using handcrafted features and a Random Forest classifier. Future work could explore region-level radiomics [11] or deep learning approaches [12] that implicitly learn multi-scale features. Additionally, evaluating feature relevance under different registration accuracies and across multi-centre datasets would further clarify the generalizability of these findings.

V. Conclusion

This study demonstrates that segmentation performance is primarily driven by voxel-wise features, which exhibit substantially higher MI with tissue labels than radiomics features, supporting the hypothesis that not all features are equally informative. Using all radiomics features degraded performance due to redundancy and noise, particularly for small structures, contradicting the assumption that increased feature dimensionality improves segmentation. In contrast, MI–based selection identified a small subset of complementary radiomics features that yielded modest but consistent improvements when combined with the baseline. These findings confirm that carefully selected feature subsets, rather than exhaustive feature inclusion, improve segmentation accuracy and robustness, especially under class imbalance.

The full experimental pipeline is publicly available on GitHub: https://github.com/zejunyan/mialab.

References

[1] Orringer, D.A., Golby, A. and Jolesz, F. (2012). *Neuronavigation in the surgical management of brain tumors: current and future trends*. Expert Review of Medical Devices, 9(5), pp. 491–500. doi: 10.1586/erd.12.42.

[2] Löfstedt, T., et al. (2019). *Gray-level invariant Haralick texture features*. PLOS ONE, 14(2), e0212110. doi: 10.1371/journal.pone.0212110.

[3] Fogel, I. and Sagi, D. (1989). *Gabor filters as texture discriminator*. Biological Cybernetics, 61(2), pp. 103–113.

[4] Ahonen, T., Hadid, A. and Pietikäinen, M. (2004). *Face recognition with local binary patterns*. In: Proceedings of the European Conference on Computer Vision (ECCV). Berlin, Heidelberg: Springer, pp. 469–481.

[5] Estévez, P.A., et al. (2009). *Normalized mutual information feature selection*. IEEE Transactions on Neural Networks, 20(2), pp. 189–201.

[6] Radovic, M., et al. (2017). *Minimum redundancy maximum relevance feature selection approach for temporal gene expression data*. BMC Bioinformatics, 18(1), p. 9. doi: 10.1186/s12859-016-1423-9.

[7] Malhotra, P., Gupta, S., Koundal, D., Zaguia, A. and Enbeyle, W. (2022). *Deep neural networks for medical image segmentation*. Journal of Healthcare Engineering, 2022, Article ID 9580991. doi: 10.1155/2022/9580991.

[8] Zhang, W., Guo, Y. and Jin, Q. (2021). *Radiomics and its feature selection: A review*. Symmetry, 13(3), p. 435. doi: 10.3390/sym13030435.

[9] Pereira, S., Pinto, A., Oliveira, J., Mendrik, A.M., Correia, J.H. and Silva, C.A. (2016). *Automatic brain tissue segmentation in MR images using Random Forests and Conditional Random Fields*. Journal of Neuroscience Methods, 270, pp. 111–123.

[10] Peng, H., Long, F. and Ding, C. (2005). *Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), pp. 1226–1238.

[11] van Griethuysen, J.J., Fedorov, A., Parmar, C., et al. (2017). *Computational radiomics system to decode the radiographic phenotype*. Cancer Research, 77(21), pp. e104–e107.

[12] Hutter, J., Price, A.N. and Hajnal, J.V. (2017). *Deep learning for brain MRI segmentation: State of the art and future directions*. Journal of Magnetic Resonance Imaging, 47(2), pp. 505–520.