Revitalize the Potential of Radiomics: Interpretation and Feature Stability in Medical Imaging Analyses through Groupwise Feature Importance

Anna Theresa Stüber[1,2,3], Stefan Coors[2] and Michael Ingrisch[1,3]

[1]Department of Radiology, University Hospital, LMU Munich
[2]Department of Statistics, LMU Munich
[3]Munich Center for Machine Learning (MCML), LMU Munich

Abstract

Radiomics, involving analysis of calculated, quantitative features from medical images with machine learning tools, shares the instability challenge with other high-dimensional data analyses due to variations in the training set. This instability affects model interpretation and feature importance assessment. To enhance stability and interpretability, we introduce grouped feature importance, shedding light on tool limitations and advocating for more reliable radiomics-based analysis methods.

Keywords

radiology, radiomics, feature (importance) instability, grouped feature importance

1. Introduction

Radiomics [1] is a field of study that aims to extract quantitative features from medical images using machine learning (ML) and statistical analysis. These features can be used to identify patterns and associations that may not be apparent from visual inspection alone. Radiomics have become increasingly popular in medical imaging as they provide a non-invasive and efficient way to extract biomarkers from medical images. [2]

Radiomics analyses typically involve three main steps: image acquisition and segmentation, feature extraction, and statistical analysis (see Fig. 1). In the first step, medical images are acquired and segmented to isolate the region of interest. In the second step, quantitative features are extracted from the segmented region using mathematical algorithms and statistical methods. These features can include shape, texture, and intensity-based metrics, among others. In the final step, statistical / machine learning (ML) based analyses are performed to identify patterns and associations between the extracted features and clinical outcomes, such as disease diagnosis, prognosis, and treatment response. [3]

Radiomics relies on measuring feature importance to understand their impact on predictions. [4] [5] ML models like Random Forests [6] generate scores indicating a feature's contribu-

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal
Theresa.Stueber@med.uni-muenchen.de (A. T. Stüber)
0000-0001-5236-4373 (A. T. Stüber); 0000-0002-7465-2146 (S. Coors); 0000-0003-0268-9078 (M. Ingrisch)

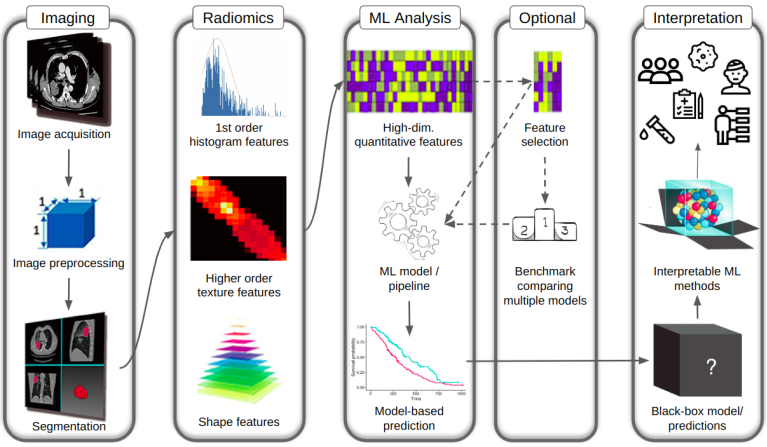

Figure 1: Workflow of radiomics analyses including imaging segmentation, radiomics feature calculation, the analysis via an ML model (pipeline) and their interpretation.

tion to prediction accuracy. Ensuring model robustness across datasets requires understanding the stability of these measures. However, like in other high-dimensional data analyses, the sensitivity of feature selection to training set variations [7] [8] [9] is restricted. Hence, methods to restore feature stability (FS) in radiomics-based analyses are indispensable. This entails assessing the coherence in feature importance scores across datasets or utilizing stability selection to identify consistent features across numerous model building iterations.

Hypothesizing low feature stability (FS) in radiomics-based prediction models, similar to other high-dimensional data analyses, we propose grouped feature importance [11] [12] [13] for assessing radiomics-based ML models, aiming to enhance stability and simplify interpretation.

2. Material and methods

2.1. Assessing feature stability of un-grouped radiomics features

To investigate the instability of radiomics-based analyses, we constructed a working example utilizing 136 pre-calculated radiomics features [14] from the Lung Image Database Consortium image collection (LIDC-IDRI) [15]. This database contains thoracic computed tomography (CT) scans with annotated lesions, classifing 616 benign and 281 malignant nodules. We established a standard machine learning (ML) classification pipeline using the R package mlr3 [16]. The pipeline encompassed a random forest (rf) with feature preprocessing (imputation, factor encoding, and correlation-based feature selection). To refine the pipeline, we utilized nested resampling with 5-fold cross-validation for both outer and inner loops. Hyperparameters, including the fraction filtered in preprocessing, the count of features randomly sampled per decision tree split, and the number of trees within the rf model, were fine-tuned. The correlation-based filter's optimization involved a correlation cutoff range from 0.1 to 0.9. AUC served as the basis for optimization and performance assessment.

Using a methodology akin to test-retest, we trained our ML pipeline over 1000 bootstrap iterations, varying solely the underlying training set (seed) [17]. For model-agnostic feature

assessment, we employed minimal depth (MD) and permutation feature importance (PFI). The variable's 'importance' threshold was determined based on the mean PFI or MD within one iteration, revealing how often the variable was deemed important across 1000 bootstraps.

2.2. Feature stability of grouped radiomics features

To test our hypothesis on improved model stability with grouped radiomics features, we'll form these groups and assess their stability using Group Feature Importance (GFI) methods.

2.2.1. Grouping radiomics features

We will consider various approaches to organizing radiomics features into groups:

1. Grouping based on Semantic Meaning / Clinical Relevance: Categories according to the anatomical or physiological aspects (shape, intensity, texture).
2. Feature Type Grouping: Groups based on calculation nature, e.g., original vs. processed (wavelet, log-filter) image features.
3. Statistical Grouping: Use statistical techniques like clustering or intercorrelation analysis to group features based on their statistical properties.
4. Task-Specific Grouping: Adapt feature grouping to the research question; e.g. for predicting treatment response, cluster treatment-related features.
5. Expert Knowledge-based Grouping: Categories guided by physicians or domain experts, based on clinical significance or feature relevance.

2.2.2. Grouped feature importance

To assess grouped feature importance [18] in radiomics analyses, we will use permutation-based [19], refitting[20], and Shapley-based [21] methods.

1. Permutation-based method: Randomly permuting grouped features measures their impact on the model's predictive accuracy.
2. Refitting method: Fit the model multiple times, excluding specific feature groups, to assess the change in performance.
3. Shapley-based method: Assign values to grouped features based on their contribution to predictions using cooperative game theory.

Furthermore, we'll employ the combined features effect plot (CFEP) [11] to visualize grouped feature impact. CFEP presents a sparse and interpretable linear combination, offering insights into the collective effect of grouped features and their combined influence on predictions.

3. Results

In our classification demonstration, employing 136 pre-calculated radiomics features from the LIDC-IDRI dataset, we trained a ML pipeline over 1000 bootstrap rounds. The models achieved

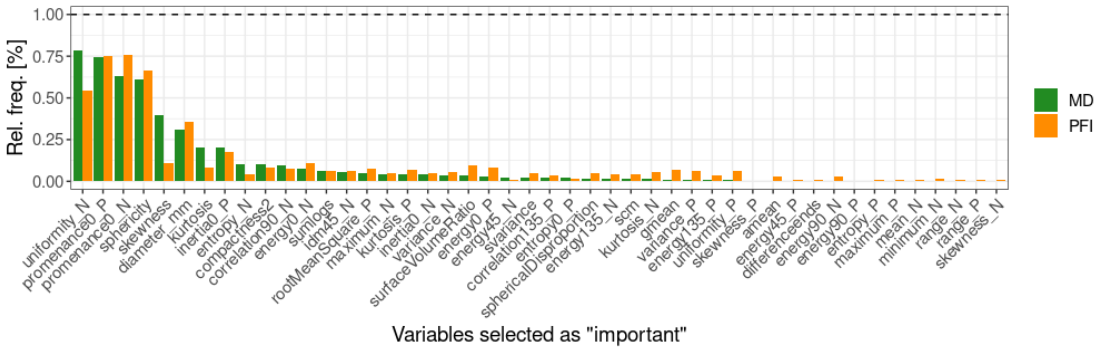

Figure 2: Relative frequency of features selected as "important" (threshold: > mean MD/PFI across all iterations) based on Minimal Depth (MD) and Permutation Feature Importance (PFI) measures. Derived from a test-retest analysis using a single radiomics-based ML model.

an average AUC of 0.880 (inter-quartile range: [0.871; 0.891]). In each bootstrap iteration, we calculated MD and PFI values, marking variables exceeding the mean MD/PFI as 'important'.

In Figure 2, the relative frequency of each variable's importance is depicted. Among the 136 radiomics features, 46 were chosen at least once by PFI, and 38 by MD. The top four vital features were selected between 54.7% and 75.6% of the time for PFI, and 61.0% to 78.3% for MD. These four features—promenance0_N, promenance0_P, sphericity, and uniformity_N—appeared as the most important in different orders. The fifth feature (PFI: diameter_mm, MD: skewness) was chosen less than 40% of the time in both cases.

4. Conclusion and outlook

In conclusion, our study highlights challenges in achieving feature stability and interpretability in radiomics-based analyses, notably during feature importance interpretation. Using a test-retest approach training one ML model (pipeline) 1000 times on varied training set combinations, we found that the most crucial feature was selected in only about 75% of cases, clearly revealing variability and uncertainty. This inconsistency persisted despite incorporating a correlation filter into our ML pipeline, signifying that the issue extends beyond feature correlations [10].

To tackle this challenge and augment the interpretability of radiomics-based ML models, we advocate for the adoption of grouped feature importance techniques. Among these, the combined features effect plot (CFEP) [11] shows promise, visually representing feature group influence through a sparse and understandable linear combination.

Instability in radiomics feature calculations arises from various factors like acquisition modes, reconstruction parameters, and segmentation thresholds. [22] [23] This instability extends to radiomics-based ML model performance due to high-dimensional test set variations. Hence, future research should focus on refining grouped feature importance methods, including CFEP, to enhance feature stability and interpretability.

Ultimately, by fortifying the stability and interpretability of radiomics-based ML models, we can *revitalize the potential of radiomics* in medical imaging, enabling more precise diagnoses, prognoses, and informed treatment choices for various medical conditions.

References

[1] C. McCague, S. Ramlee, M. Reinius, I. Selby, D. Hulse, P. Piyatissa, V. Bura, M. Crispin-Ortuzar, E. Sala, R. Woitek. Introduction to radiomics for a clinical audience. Clin Radiol. 2023 Feb;78(2):83-98. doi: 10.1016/j.crad.2022.08.149. PMID: 36639175.

[2] Joon Young Choi. "Radiomics and Deep Learning in Clinical Imaging: What Should We Do?." Nuclear medicine and molecular imaging vol. 52,2 (2018): 89-90. doi:10.1007/s13139-018-0514-0

[3] Janita E. van Timmeren, Davide Cester, Stephanie Tanadini-Lang, Hatem Alkadhi, Bettina Baessler. Radiomics in medical imaging—"how-to" guide and critical reflection. Insights Imaging 11, 91 (2020). https://doi.org/10.1186/s13244-020-00887-2

[4] Andrei Mouraviev, Jay Detsky, Arjun Sahgal, Mark Ruschin, Young K Lee, Irene Karam, Chris Heyn, Greg J Stanisz, Anne L Martel. Use of radiomics for the prediction of local control of brain metastases after stereotactic radiosurgery, Neuro-Oncology, Volume 22, Issue 6, June 2020, Pages 797–805, https://doi.org/10.1093/neuonc/noaa007

[5] Sohi Bae, Chansik An, Sung Soo Ahn, Hwiyoung Kim, Kyunghwa Han, Sang Wook Kim, Ji Eun Park, Ho Sung Kim, Seung-Koo Lee. Robust performance of deep learning for distinguishing glioblastoma from single brain metastasis using radiomic features: model development and validation. Sci Rep 10, 12110 (2020). https://doi.org/10.1038/s41598-020-68980-6

[6] Johanna S. Enke, Jan H. Moltz, Melvin D'Anastasi, Wolfgang G. Kunz, Christian Schmidt, Stefan Maurus, Alexander Mühlberg, Alexander Katzmann, Michael Sühling, Horst Hahn, Dominik Nörenberg, Thomas Huber. Radiomics features of the spleen as surrogates for CT-based lymphoma diagnosis and subtype differentiation. Cancers, 14(3), 713 (2022).

[7] Alexandros Kalousis, Julien Prados, Melanie Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl Inf Syst 12, 95–116 (2007). https://doi.org/10.1007/s10115-006-0040-8

[8] Barbara Pes. Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. Neural Comput & Applic 32, 5951–5973 (2020). https://doi.org/10.1007/s00521-019-04082-3

[9] Salem Alelyani, Zheng Zhao, Huan Liu. A Dilemma in Assessing Stability of Feature Selection Algorithms. IEEE International Conference on High Performance Computing and Communications, Banff, AB, Canada, 2011, pp. 701-707, doi: 10.1109/HPCC.2011.99.

[10] Utkarsh Mahadeo Khaire and R. Dhanalakshmi. 2022. Stability of feature selection algorithm: A review. J. King Saud Univ. Comput. Inf. Sci. 34, 4 (2022), 1060–1073. https://doi.org/10.1016/j.jksuci.2019.06.012

[11] Quay Au, Julia Herbinger, Clemens Stachl, Bernd Bischl, Giuseppe Casalicchio: Grouped feature importance and combined features effect plot. Data Mining and Knowledge Discovery (2022). 36. 1-50. 10.1007/s10618-022-00840-5.

[12] Cheng Zhu, Huili Gong, Zhongren Li, Chunxia Yu. Application of High Dimensional Feature Grouping Method in Near-Infrared Spectra of Identification of Tobacco Growing Areas. 3rd International Conference on Information Science and Control Engineering (ICISCE), Beijing, China, 2016, pp. 230-234, doi: 10.1109/ICISCE.2016.58.

[13] Zhigang Shang, Mengmeng Li. Feature Selection Based on Grouped Sorting. 9th Interna-

tional Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 2016, pp. 451-454, doi: 10.1109/ISCID.2016.1111.

[14] Anthony Jatoba, Igor Theotonio, Eduardo Moraes, Vinicius Costa. deep_radiomics (2020). https://github.com/anthonyjatoba/deep_radiomics

[15] Samuel G. Armato III, Geoffrey McLennan, Michael F. McNitt-Gray, Charles R. Meyer, David Yankelevitz, Denise R. Aberle, Claudia I. Henschke, Eric A. Hoffman, Ella A. Kazerooni, Heber MacMahon, Anthony P. Reeves, Barbara Y. Croft, Laurence P. Clarke. Lung image database consortium: developing a resource for the medical imaging research community. Radiology 232.3 (2004): 739-748.

[16] Michael Lang, Martin Binder, Jakob Richter, Patrick Schratz, Florian Pfisterer, Stefan Coors, Quay Au, Giuseppe Casalicchio, Lars Kotthoff, Bernd Bischl (2019). "mlr3: A modern object-oriented machine learning framework in R." Journal of Open Source Software. doi:10.21105/joss.01903, https://joss.theoj.org/papers/10.21105/joss.01903

[17] Sarah Nogueira, Gavin Brown. Measuring the stability of feature selection. Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part II 16. Springer International Publishing, 2016.

[18] Christoph Molnar, Gunnar König, Bernd Bischl, Giuseppe Casalicchio. Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach. Data Min Knowl Disc (2023). https://doi.org/10.1007/s10618-022-00901-9

[19] Baptiste Gregorutti, Bertrand Michel, Philippe Saint-Pierre: Grouped variable importance with random forests and application to multiple functional data analysis. Computational Statistics & Data Analysis (2015). 90:15–35

[20] Brian D. Williamson, Peter B. Gilbert, Noah R. Simon, Marco Carone (2020): A unified approach for inference on algorithm-agnostic variable importance. arXiv:200403683

[21] Scott M. Lundberg, Su-In Lee: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA (2017), NIPS'17, p 4768–4777

[22] Stephen S. Yip, Hugo J. Aerts. Applications and limitations of radiomics. Phys Med Biol. 2016 Jul 7;61(13):R150-66. doi: 10.1088/0031-9155/61/13/R150. Epub 2016 Jun 8. PMID: 27269645; PMCID: PMC4927328.

[23] Kyriakos Flouris, Oscar Jimenez-del-Toro, Christoph Aberle, Michael Bach, Roger Schaer, Markus M. Obmann, Bram Stieltjes, Henning Müller, Adrien Depeursinge, Ender Konukoglu. Assessing radiomics feature stability with simulated CT acquisitions. Scientific reports 12.1 (2022): 4732.