A Study on the Role of Radiomics Feature Stability in Predicting Breast Cancer Subtypes

Isabella Cama*[a,b], Alejandro Guzman*[a], Sara Garbarino[c], Cristina Campi[b,c], Karim Lekadir[a, d], and Oliver Díaz[a, e]

[a]Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain
[b]Università degli Studi di Genova, Genova, Italy
[c]IRCCS Ospedale Policlinico San Martino, Genova, Italy
[d]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
[e]Computer Vision Center, Bellaterra, Spain

ABSTRACT

Imaging features (radiomics) have potential for predicting Triple Negative Breast Cancer and other subtypes using magnetic resonance images (MRI). This work uses 244 images from the Duke-Breast-Cancer-MRI dataset to investigate the complex interplay between radiomics feature stability, with respect to segmentation variability, and prediction results of machine learning models. Our analysis reveals that features demonstrating high stability across different segmentations tend to enhance model performance, whereas unstable features sensitive to small segmentation changes degrade predictive accuracy. This exploration underscores the importance of feature stability in the development of reliable models for breast cancer subtype classification.

Keywords: Breast cancer, Triple negative breast cancer, Radiomics, Machine Learning, Artificial Intelligence, Stability, Features, Segmentation Variability, Magnetic Resonance Imaging, Predictive Modeling

1. INTRODUCTION

Breast cancer, a complex and heterogeneous disease with diverse molecular subtypes, presents a formidable challenge for accurate prediction and targeted treatment strategies.[1] Traditionally, identifying breast cancer subtypes relies on techniques such as immunohistochemistry (IHC) or staining and fluorescence in situ hybridization (FISH) analyses.[1] However, recent advances in the literature indicate the potential for predicting molecular subtypes using image-based features through machine learning (ML) techniques.[2] These features, known as radiomics features, numbering in the thousands, offer a wealth of mineable data, encompassing information on morphology, intensity, texture and more within a region of interest (ROI), such as a tumor. The conventional radiomics pipeline requires an initial segmentation,[3,4] relying on the annotator's expertise and the visibility of ROI boundaries.[5] Subsequently, all feasible radiomics features are extracted from the segmented images, feature selection procedures are performed to discard irrelevant or redundant features,[6] and ultimately prediction models are trained.

In a recent work in radiomics for breast cancer, Son et al.[7] conducted a study to predict the molecular subtype using clinical and radiomics data. Synthetic mammography reconstructed from digital breast tomosynthesis and clinical data such as patient age, lesion size, and mammographic features were used. An elastic-net logistic regression model was used to create the radiomics signature of each lesion. They found that predictions using a combination of radiomics and clinical data outperformed those using clinical data alone, suggesting that radiomics signatures could serve as biomarkers for Triple Negative breast cancer (TNBC), the most aggressive type of tumour with a faster growth rate, aiding in treatment direction for these patients. As for radiomics on MR imaging, Leithner et al.[8] showed that combining radiomics data extracted from both dynamic contrast-enhanced (DCE) MRI and apparent diffusion coefficient (ADC) mapping could aid in differentiating TNBC from other subtypes. These works show that the radiomics approach could potentially improve patient stratification and aid treatment planning by offering a comprehensive and non-invasive means of analyzing tumor biology.

*These authors contributed equally to this work.

Further author information: (Send correspondence to O.D.)
O.D.: E-mail: oliver.diaz@ub.edu

In this context, the present study explores, for the first time, the impact of radiomics feature stability with respect to segmentation variability on the predictive performance of ML models designed to discriminate between TNBC and other molecular subtypes. Our hypothesis posits that early identification of consistent, robust and unstable features will enhance the feature selection stage, ultimately improving the predictive performance. This research represents a crucial step toward refining the radiomics approach for breast cancer subtype discrimination and underscores the potential for advancing personalised treatment strategies.

2. METHODS AND MATERIALS

This work aims at investigating the influence of radiomics feature stability, with respect to segmentation variability, in breast cancer subtype prediction of TNBC versus non-TNBC. Figure 1 shows the traditional radiomics pipeline, including image segmentation, feature extraction, and feature selection, with the addition of the proposed *feature stability assessment* module in yellow.

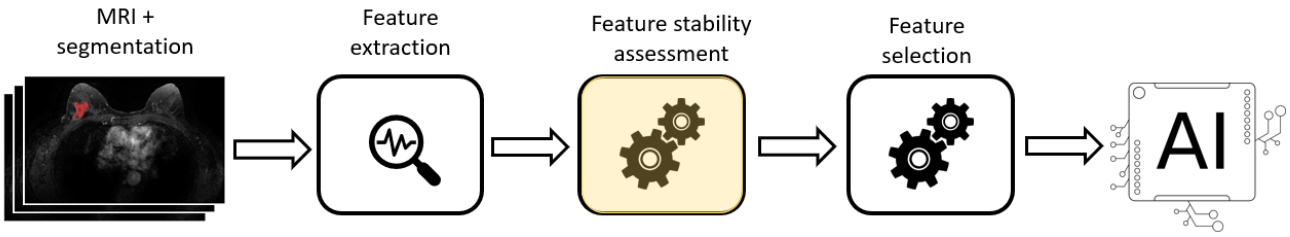

Figure 1. Radiomics pipeline used in this work. In yellow, the proposed *feature stability assessment* module to improve radiomics results.

The Duke-Breast-Cancer-MRI dataset[9] was used. For each patient, the dataset contains images from multiple time points, capturing both pre-contrast and post-contrast phases. The dataset also contains demographic, clinical, pathology and treatment information, and outcomes of patients (e.g., response to treatment, recurrence, follow-up). Pre-operative DCE-MR first post-contrast images of breast cancer patients with tumour locations were employed for this study. Specifically, data from 244 patients were analyzed, including 71 with TNBC and 173 non-TNBC (30% vs 70% of the dataset, respectively). Such images were manually segmented as described by Caballo et al.[10] To assess the stability of the features with respect to segmentation accuracy, we introduced variability in the segmentation masks by simulating annotations from two other experts. This was achieved through morphological operations, specifically opening and closing with a kernel size of 3, applied on the original manual segmentations. In the rest of the document the original, opening and closing segmentations are refereed as mask A, B, and C, respectively (see Figure 2). The morphological operations employed in this work introduce light variations to the original segmentation. As a result, over 70% of masks B and C exhibit a Dice similarity coefficient (DSC) greater than 0.8 (over 1.0), when compared to mask A.

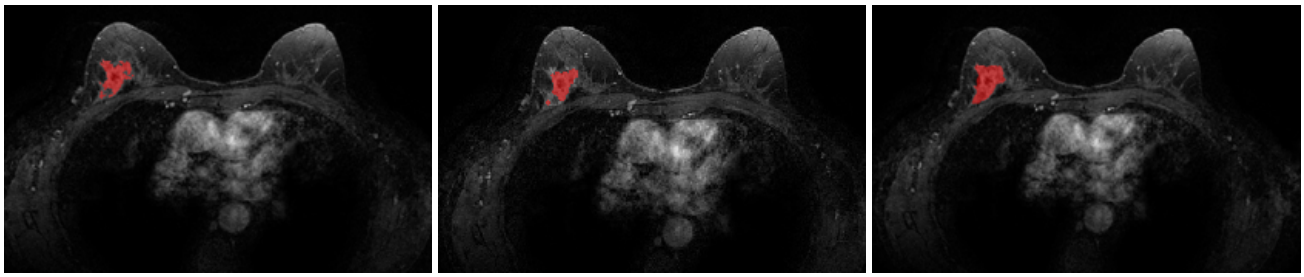

Figure 2. Sample of MR slice images with segmentations of breast cancer in red. From left to right: segmentation mask A (original), B (opening), and C (closing).

Radiomics features were extracted from the masks (A, B and C) after image z-score normalization and fixed-bin count discretization with 50 bins using PyRadiomics 3.1.0 library.[11] The extracted features (n=1030) included shape, texture, matrix-based, wavelet, and Laplacian of Gaussian (LoG) features (σ = 1, 2, 3).

A preliminary feature selection based on correlation (threshold fixed at 0.7) was performed before the grid search to facilitate the identification of informative features to use during training and 5-fold cross-validation. The optimal model identified through grid search was assessed using balanced accuracy, which accounts for dataset imbalance.

Then, in order to benchmark our feature selection procedure (based on the stability with respect to segmentation accuracy) against state-of-the-art methods, we conducted an alternative feature selection procedure by means of SHAP (SHapley Additive exPlanations[12]). SHAP is a feature importance tool, based on a game theoretic approach, used in ML for explaining the output of a model by quantifying the importance of each feature. SHAP identifies the most relevant features that contribute to the model's predictions by calculating SHAP values for each feature in the dataset: features with higher SHAP values are considered more influential in the model's predictions, while features with lower SHAP values have less impact. Specifically, SHAP was performed across distinct data splits, recording the top 10 selected features each time. Subsequently, for each of these features the frequency of occurrence was counted and their mean SHAP weight was computed. Amongst this set of SHAP-selected features, we further selected those with the highest frequency (> 5) and used them as an alternative set of features, the *best SHAP features*, for comparison with our model.

Our feature selection tool, on the other hand, was implemented by assessing feature stability with respect to segmentation accuracy, using the method proposed by Cama et al.[13] This tool allows to compute, for each feature, at varying segmentation mask, a set of indexes pertaining features' quality, consistency, robustness, and instability. In our specific example, the segmentation variability arises from the differences between mask A and B, as well as between mask A and C. We defined as highly consistent/robust/unstable features the ones which exhibited a higher consistency/robustness/instability score [13] than their median value across the set of uncorrelated radiomics features. Then, we performed three different alternative feature selection procedures, selecting each time highly consistent/robust/unstable features. In the following we focus on the stability indexes computed for mask A and B (data for A and C were similar and are not shown).

In summary, following the radiomics pipeline presented in Figure 1, TNBC prediction was investigated for the original segmentation mask (mask A):

- prediction with all the uncorrelated radiomics features;
- prediction with the best SHAP radiomics features;
- prediction with high-consistency radiomics features (consistency score greater than its median), meaning features' error varies linearly with segmentation variation;
- prediction with high-robustness radiomics features (robustness score greater than its median), meaning features' values remain stable with any segmentation variation;
- prediction with high-instability radiomics features (instability score greater than its median), i.e., features that vary largely within small changes in the segmentation mask.

Two of the most common ML models where trained from MRI-derived features to predict TNBC, Random Forest (RF) and Support Vector Machine (SVM) algorithms. The models were fine-tuned through GridSearchCV for hyperparameter optimization. The dataset was divided into a 70-30 stratified proportion for training and testing, and a class weighting strategy was implemented during training using the *inverse class frequency* technique in order to weight the classes (TNBC vs non-TNBC) based on their frequency.

The resulting baseline prediction models, leveraging all the uncorrelated radiomics features, best SHAP radiomics features, as well as high-consistency, high-robustness, and high-instability features, underwent rigorous evaluation for various performance metrics: accuracy, balanced accuracy, f1, precision, average precision, recall and area under the Receiver Operating Characteristic Curve (ROC-AUC).

The performances of the best models were boosted by training them over 85 random splits of the dataset, and the distributions of the scores (balanced accuracy) were compared via the Student's t-test,[14] a statistical test suitable for comparing two independent normal distributions (normality assessed via Shapiro-Wilk test[15]).

3. EXPERIMENTS AND RESULTS

The learning curves depicted in Figure 3 illustrate the trend of balanced accuracy for the classifiers, RF and SVM, both on training and test set, across various training set sizes, showing how much we benefit from adding more training data. For both RF and SVM, the test score and the training score do not converge to a value with increasing size of the training set, across various data splits, and regardless of the employed feature selection technique. Plots in Figure 3 show the trend of the balanced accuracy when training with all the uncorrelated radiomics features. The curves are plotted with the mean scores, while variability during cross-validation is shown with the shaded areas (standard deviation for all cross-validations). From this result, we conjecture that the model is underfitting because this particular set of radiomics features lacks predictive capability for distinguishing between TNBC and non-TNBC cases. This leads us to examine the impact of different feature selection methods to ascertain their effectiveness in improving model performance and increasing generalization. The experiments we conducted seek to investigate how radiomics' stability impact the training of the classifier, in relation to predictions made using the best SHAP features or all the uncorrelated features. For the description of these experiments, greater emphasis was placed on the training phase, in alignment with SHAP selection process. This decision was also influenced by the limited generalizability shown by the classifier during testing phase, as discussed above.

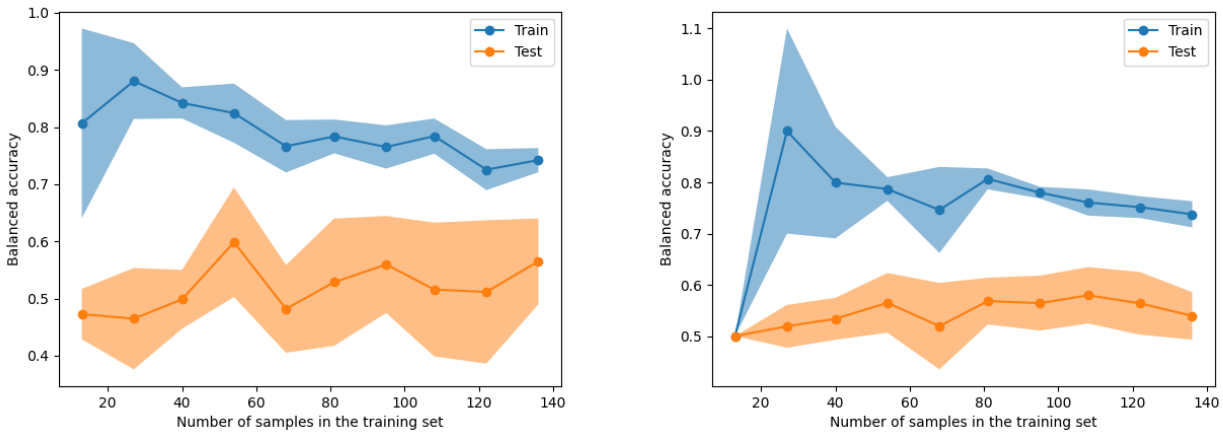

Figure 3. Learning curves for RF (left) and SVM (right) when training with all the uncorrelated radiomics features. The curve is drawn with the mean scores (balanced accuracy). The shaded areas display variability during cross-validation (standard deviation).

As described in Section 2, SHAP was used to identify features with significant impact on the model's predictions. Table 2 lists the highly influential features selected within the top 10 of each split, sorted by frequency of occurrence. Figure 4 shows the beeswarm plot of SHAP most influential features for one random data split. The best SHAP features (selected by frequency >5, see Table 2) are 31. We point out that the uncorrelated features largely overlap with the ones listed in Table 2.

The thresholds for the stability measures (i.e., consistency, robustness, and instability score) were set to the median of the score distribution within the uncorrelated features (median consistency 0.39, robustness 0.21, instability 0.08). This ensured that half of the feature set was included in each model training process. Table

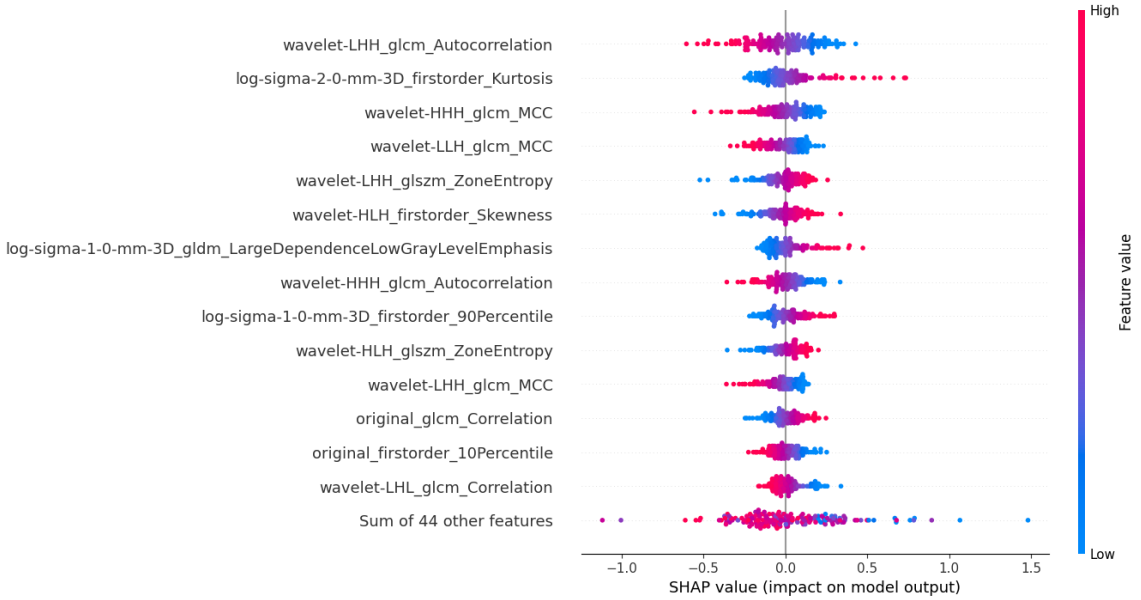

Figure 4. Beeswarm plot of SHAP values for the 15 most influential features selected by SHAP method for one random data split. Features are ordered (top to bottom) according to their importance for that split.

2 highlights features with specific combinations of characteristics. Features with high consistency and high instability are marked in green, those with high instability and high robustness in yellow, features exhibiting high consistency and high robustness are in light blue, and those demonstrating high consistency, high robustness, and high instability are highlighted in orange. In Table 2 the reliability scores have been calculated based on the segmentation variability introduced by the difference between mask A and mask B. A comparable selection of features is observed in the corresponding analysis involving RF classifier and the reliability scores computed between mask A and mask C.

Figures 5 and 6 show the distribution (i.e., histogram) of the balanced accuracy in training for 85 random data splits with all the uncorrelated features (green), best SHAP features (yellow), and high-consistency, robustness, and instability features (red) for both SVM and RF classifiers. Table 1 provides the means and standard deviations of these distributions, along with the p-values resulting from the t-tests conducted to compare the means within the scores distributions (normality assessed via Shapiro-Wilk test). P-value < 0.05 indicates evidence against the null hypothesis of equal score means. In all the cases, training the model with the best SHAP features led to better balanced accuracy scores than training with all the uncorrelated radiomics features. The observed behavior is more pronounced with the SVM classifier compared to RF. However, it is important to recall that the training procedure exhibited underfitting on this dataset, as depicted in Figure 3.

Figures 5 and 6 in the left panel display that, in the training phase, consistent features perform similar to or better than all the uncorrelated features (RF, $p = 0.91$; SVM, $p < 10^{-5}$), but also similar to the best SHAP features (RF, $p < 10^{-3}$; SVM, $p = 0.04$), on average. We recall that the values of consistent features strongly

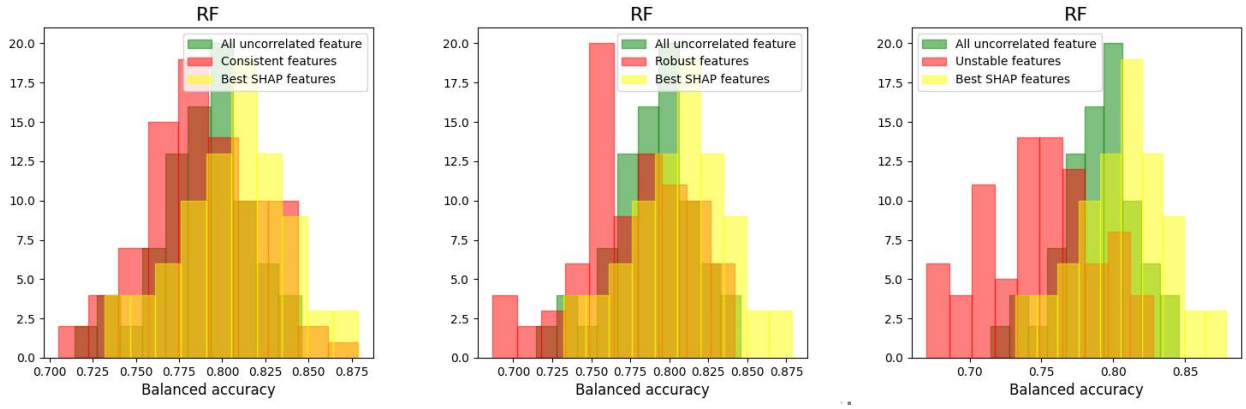

Figure 5. Distribution of the balanced accuracy computed on 85 random data splits on the training set for RF classifier. Green: all uncorrelated features. Yellow: best SHAP features. Red, from left to right: high-consistency, high-robustness, and high-instability features.

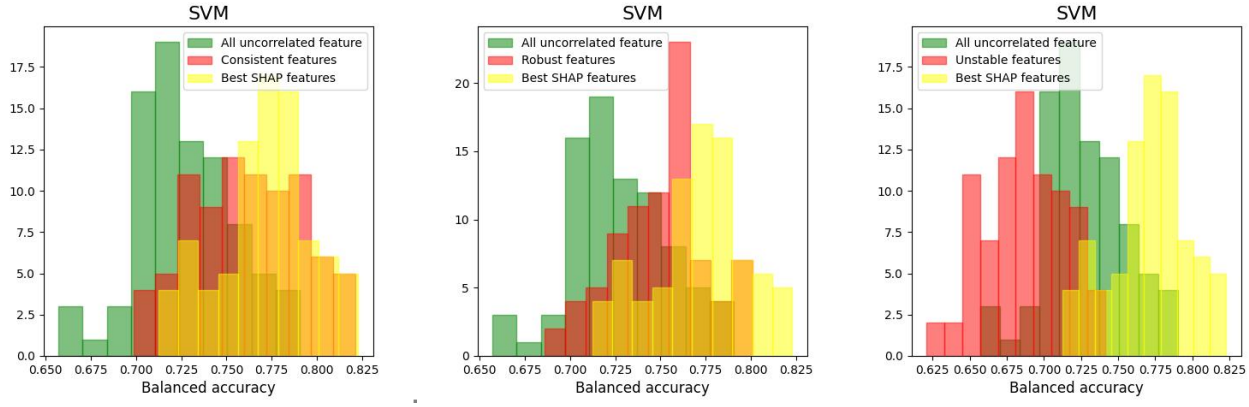

Figure 6. Distribution of the balanced accuracy computed on 85 random data splits on the training set for SVM classifier. Green: all uncorrelated features. Yellow: best SHAP features. Red, from left to right: high-consistency, high-robustness, and high-instability features.

correlate with the segmentation result, with variations in the segmentation mask typically resulting in a linear increase of the feature error.

In the middle panel of Figures 5 and 6, the distribution of the scores obtained using robust features, which are the ones that remain unaffected by variations in the segmentation mask. During training, the distributions of the scores obtained using the robust or all the uncorrelated features almost overlap for RF ($p < 10^{-2}$) while, for SVM, robust features improved. In both cases, the balanced accuracy is degraded with respect to the best SHAP features.

Finally, the right panel of Figures 5 and 6 show that unstable features tend to deteriorate the prediction outcomes, pushing them closer to random guesses in the training phase ($p < 10^{-10}$ for both RF and SVM). This degradation could derive from the fact that the value of unstable features is highly dependent on minor segmentation variations, making them non-robust in the context of tumor segmentation variability.

Table 1. Mean and standard deviation of the distributions of balanced accuracy across experiments performed with different feature selection methods, using SVM and RF classifier. In the 4th column, p-value of the t-test performed on uncorrelated features vs best SHAP, consistent, robust, and unstable features; in the last column p-value of the t-test performed on best SHAP features vs consistent, robust, and unstable features. P values < 0.05 provides evidence against equal score means.

	Features	**Mean**	**Standard Deviation**	**P value (vs uncorrelated)**	**P value (vs best SHAP)**
RF	Uncorrelated	0.79	0.03	-	-
	Best SHAP	0.81	0.03	$< 10^{-2}$	-
	Consistency	0.79	0.03	0.91	$< 10^{-2}$
	Robustness	0.78	0.04	0.01	$< 10^{-2}$
	Instability	0.75	0.04	$< 10^{-2}$	$< 10^{-2}$
SVM	Uncorrelated	0.73	0.03	-	-
	Best SHAP	0.77	0.03	$< 10^{-2}$	-
	Consistency	0.76	0.03	$< 10^{-2}$	0.04
	Robustness	0.75	0.03	$< 10^{-2}$	$< 10^{-2}$
	Instability	0.69	0.03	$< 10^{-2}$	$< 10^{-2}$

Table 2: Features selected by SHAP in the top 10 across 85 data splits, with frequency of occurrence. The first 31 (frequency > 5) are the best SHAP features, having the highest impact in the prediction of TNBC with SVM classifier. The symbols *, +, and ° indicate high consistency, robustness, and instability, respectively. Green: high consistency and high instability. Yellow: high instability and high robustness. Light blue: high consistency and high robustness. Orange: high consistency, high robustness, and high instability.

Feature	**Frequency**	**Mean weight**
wavelet-HHH_glcm_MCC *	75	0.12
wavelet-HLH_firstorder_Skewness °	75	0.12
wavelet-LLH_glcm_MCC * +	49	0.09
wavelet-HHL_glcm_MCC * °	48	0.10
wavelet-HLH_glrlm_LongRunLowGrayLevelEmphasis * °	42	0.09
wavelet-LHH_glcm_Autocorrelation * ° +	40	0.09
wavelet-LHH_glszm_ZoneEntropy +	34	0.09
original_firstorder_Skewness °	30	0.10
wavelet-HLH_glszm_ZoneEntropy +	28	0.09
original_firstorder_Minimum +	27	0.09
original_firstorder_10Percentile *	24	0.09
wavelet-HLH_glcm_Autocorrelation * ° +	24	0.09
log-sigma-1-0-mm-3D_firstorder_Skewness °	23	0.11
wavelet-HLL_firstorder_Skewness ° +	21	0.09
log-sigma-2-0-mm-3D_firstorder_Skewness ° +	20	0.09
wavelet-HHH_glcm_Autocorrelation *	19	0.10
wavelet-LHL_glcm_Correlation * +	19	0.08
wavelet-LLH_glcm_Autocorrelation * °	19	0.09
log-sigma-2-0-mm-3D_firstorder_Kurtosis * +	18	0.09
wavelet-LLH_glszm_ZoneEntropy +	16	0.08
log-sigma-1-0-mm-3D_firstorder_90Percentile * °	15	0.09

Continued on next page

Table 2 continued from previous page

Feature	Frequency	Mean weight
wavelet-LHH_firstorder_Mean °	15	0.09
wavelet-LHH_glcm_MCC * °	14	0.09
wavelet-LHL_glcm_Autocorrelation * +	11	0.09
original_firstorder_Entropy +	10	0.08
wavelet-HLL_gldm_DependenceEntropy +	9	0.09
original_glcm_Correlation +	6	0.09
wavelet-HLL_glcm_MCC * +	6	0.09
wavelet-HLL_glcm_Autocorrelation * °	5	0.08
wavelet-LHL_glcm_MCC * +	5	0.08
wavelet-LLH_firstorder_Skewness °	5	0.08
log-sigma-1-0-mm-3D_gldm_LargeDependenceLowGrayLevelEmphasis * °	4	0.10
log-sigma-3-0-mm-3D_firstorder_Skewness ° +	4	0.09
log-sigma-3-0-mm-3D_glrlm_LongRunLowGrayLevelEmphasis * °	4	0.09
wavelet-HHL_glszm_ZoneEntropy +	4	0.11
log-sigma-1-0-mm-3D_glcm_Correlation +	3	0.07
original_glcm_Autocorrelation *	3	0.10
original_glszm_ZoneEntropy +	3	0.08
wavelet-HHH_firstorder_Entropy +	3	0.11
wavelet-HHL_glcm_ClusterShade °	3	0.08
wavelet-HLH_glcm_MCC *	3	0.07
wavelet-HLH_gldm_DependenceEntropy +	3	0.13
wavelet-HLL_glszm_ZoneEntropy +	3	0.08
wavelet-LHH_glcm_Correlation * °	3	0.07
log-sigma-3-0-mm-3D_glcm_ClusterProminence * °	2	0.12
original_glcm_ClusterProminence * °	2	0.09
original_glcm_Imc1 +	2	0.09
original_shape_Elongation *	2	0.09
original_shape_SurfaceVolumeRatio * °	2	0.08
wavelet-HHH_firstorder_Kurtosis * +	2	0.08
wavelet-HHH_firstorder_Skewness °	2	0.07
wavelet-HHH_glcm_ClusterShade °	2	0.07
wavelet-HHL_glcm_Autocorrelation * ° +	2	0.08
wavelet-HLL_glcm_ClusterShade ° +	2	0.07
wavelet-LHH_glrlm_LongRunLowGrayLevelEmphasis * °	2	0.08
log-sigma-1-0-mm-3D_glcm_Imc1 +	1	0.07
log-sigma-2-0-mm-3D_firstorder_Entropy +	1	0.08
original_glrlm_LongRunLowGrayLevelEmphasis * °	1	0.07
original_ngtdm_Coarseness °	1	0.08
original_shape_Flatness *	1	0.10
original_shape_LeastAxisLength *	1	0.09
wavelet-HHH_firstorder_Mean °	1	0.07
wavelet-HHH_glrlm_LongRunLowGrayLevelEmphasis * °	1	0.07
wavelet-HLH_firstorder_Mean °	1	0.07
wavelet-LHL_firstorder_Entropy * +	1	0.07
wavelet-LHL_firstorder_Skewness ° +	1	0.08
wavelet-LHL_gldm_DependenceEntropy +	1	0.09
wavelet-LHL_glszm_ZoneEntropy +	1	0.07

4. DISCUSSION

This paper presents a radiomics feature selection model based on radiomics features' stability with respect to image segmentation. By comparing this method with a state-of-the-art feature selection procedure SHAP, based on explainability, we retrieved similar performances in terms of balanced accuracy of TNBC prediction on a training set derived from the Duke-Breast-Cancer-MRI dataset.

Specifically, we defined a set of best SHAP features (those occurring in the most influential positions more than 5 times across data splits), a set of highly consistent (whose consistency is $>$ the median), a set of highly robust (robustness $>$ the median), and a set of unstable (instability $>$ the median), and for all of them trained a ML model to predict TNCB (Table 2 for stability details). We note that the results obtained using the highly consistent features are particularly compatible with those obtained using the best SHAP features (Figures 5 and 6), suggesting that our feature selection tool could potentially serve as a reliable method in case of dataset with variability in the image segmentations (e.g., different experts). On the contrary, the experiments showed that unstable features tend to degrade the accuracy of the classifiers.

From a more descriptive side, we observe that, among the best SHAP features, 17 features are robust, 17 are consistent, and 15 are unstable (Table 2). It is worth noting that while consistent features are reasonably sensible to segmentation variations (the feature error linearly decreases with segmentation variations), consistency may introduce some instability due to the generous boundary allowed on feature error (up to 20% of relative error on the feature value for Dice coefficient greater than 0.9, up to 30% for Dice coefficient between 0.8 and 0.9, and so on[13]).

This study presents several limitations. First, the relatively small dataset size may limit the generalizability of our findings to broader patient populations. Moreover, addressing segmentation variability using only three segmentation masks and employing just two ML models may not adequately capture the full spectrum of complexities involved in tumor subtype classification. These limitations underscore the need for further research to validate and extend our our initial observations, exploring a wider datasets, addressing segmentation variability in a systematic way, and using different ML models to fully understand the relevance of radiomics features in classification tasks like the one addressed in this study.

5. CONCLUSIONS

This investigation delved into the impact of radiomics feature stability on the classification performance of a ML model designed to distinguish TNBC from other tumor subtypes. More specifically, this manuscript explored the possibility that variations in the segmentation mask, achieved through opening and closing operations, could elucidate feature behavior in relation to prediction outcomes. High-instability features were identified as the ones with the worst impact on the model training, while the selection of highly consistent features enhanced the classification performance, suggesting that feature selection based on stability could serve as a reliable method in the case of high variability in data segmentation.

Our future endeavors involve an exploration of whether varying threshold for stability-based feature selection can lead to improvements in prediction performance, leveraging insights gained from a larger dataset and different ML classifiers. This work aspires to contribute to a deeper understanding of the intricate interplay between feature stability in relation to segmentation and prediction accuracy, fostering the development of more reliable, generalizable, and clinically applicable models for breast cancer diagnosis and prognosis.

The work presented here has not been, nor is it being, submitted for publication or presentation elsewhere.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon Europe and Horizon 2020 research and innovation programme under grant agreement No 101057699 (RadioVal) and No 952103 (EuCanImage), respectively. Also, this work was partially supported by the project FUTURE-ES (PID2021-126724OB-I00) from the Ministry of Science and Innovation of Spain. IC acknowledges the financial support of the "Hub Life Science – Digital Health (LSH-DH) PNC-E3-2022-23683267 - Progetto DHEAL-COM – CUP: D33C22001980001". Also, this research was supported in part by the MIUR Excellence Department Project awarded to Dipartimento

di Matematica, Università di Genova, CUP D33C23001110001. IC is member of the Gruppo Nazionale per il Calcolo Scientifico - Istituto Nazionale di Alta Matematica (GNCS - INdAM).

REFERENCES

[1] Johnson, K. S., Conant, E. F., and Soo, M. S., "Molecular subtypes of breast cancer: a review for breast radiologists," *Journal of Breast Imaging* **3**(1), 12–24 (2021).

[2] Sha, Y. and Chen, J., "Mri-based radiomics for the diagnosis of triple-negative breast cancer: a meta-analysis," *Clinical Radiology* **77**(9), 655–663 (2022).

[3] Gillies, R. J., Kinahan, P. E., and Hricak, H., "Radiomics: images are more than pictures, they are data," *Radiology* **278**(2), 563–577 (2016).

[4] Zwanenburg, A., Vallières, M., Abdalah, M. A., Aerts, H. J., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R. J., Boellaard, R., et al., "The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping," *Radiology* **295**(2), 328–338 (2020).

[5] Rizzo, S., Botta, F., Raimondi, S., Origgi, D., Fanciullo, C., Morganti, A. G., and Bellomi, M., "Radiomics: the facts and the challenges of image analysis," *European radiology experimental* **2**(1), 1–8 (2018).

[6] Conti, A., Duggento, A., Indovina, I., Guerrisi, M., and Toschi, N., "Radiomics in breast cancer classification and prediction," in [*Seminars in cancer biology*], **72**, 238–250, Elsevier (2021).

[7] Son, J., Lee, S. E., Kim, E.-K., and Kim, S., "Prediction of breast cancer molecular subtypes using radiomics signatures of synthetic mammography from digital breast tomosynthesis," *Scientific reports* **10**(1), 21566 (2020).

[8] Leithner, D., Mayerhoefer, M. E., Martinez, D. F., Jochelson, M. S., Morris, E. A., Thakur, S. B., and Pinker, K., "Non-invasive assessment of breast cancer molecular subtypes with multiparametric magnetic resonance imaging radiomics," *Journal of clinical medicine* **9**(6), 1853 (2020).

[9] Saha, A., Harowicz, M. R., Grimm, L. J., Kim, C. E., Ghate, S. V., Walsh, R., and Mazurowski, M. A., "A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 dce-mri features," *British journal of cancer* **119**(4), 508–516 (2018).

[10] Caballo, M., Sanderink, W. B., Han, L., Gao, Y., Athanasiou, A., and Mann, R. M., "Four-dimensional machine learning radiomics for the pretreatment assessment of breast cancer pathologic complete response to neoadjuvant chemotherapy in dynamic contrast-enhanced mri," *Journal of Magnetic Resonance Imaging* **57**(1), 97–110 (2023).

[11] van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Bee ts Tan, R. G. H., Fillon-Robin, J. C., Pieper, S., and Aerts, H. J. W. L., "Computational radiomics system to decode the radiographic phenotype," *Cancer Research* **77**(21), e104–e107 (2017).

[12] Lundberg, S. M. and Lee, S.-I., "A unified approach to interpreting model predictions," *Advances in neural information processing systems* **30** (2017).

[13] Cama, I., Candiani, V., Roccatagliata, L., Fiaschi, P., Rebella, G., Resaz, M., Piana, M., and Campi, C., "Segmentation agreement and the reliability of radiomics features," *Advances in Computational Science and Engineering* **1**(2), 202–217 (2023).

[14] Mishra, P., Singh, U., Pandey, C. M., Mishra, P., and Pandey, G., "Application of student's t-test, analysis of variance, and covariance," *Annals of cardiac anaesthesia* **22**(4), 407–411 (2019).

[15] Shapiro, S. S. and Wilk, M. B., "An analysis of variance test for normality (complete samples)," *Biometrika* **52**(3-4), 591–611 (1965).