Adaptive Machine Learning Approach for Importance Evaluation of Multimodal Breast Cancer Radiomic Features

Giulio Del Corso[1*†], Danila Germanese[1†], Claudia Caudai[1†], Giada Anastasi[2,3], Paolo Belli[4,5], Alessia Formica[2], Alberto Nicolucci[6], Simone Palma[4], Maria Antonietta Pascali[1], Stefania Pieroni[2], Charlotte Trombadori[4], Sara Colantonio[1,o], Michela Franchini[2,o*], Sabrina Molinaro[2,o]

[1]Institute of Information Science and Technologies *"A. Faedo"* (ISTI), National Research Council of Italy (CNR), Pisa, Italy.
[2]Institute of Clinical Physiology (IFC), National Research Council of Italy (CNR), Pisa.
[3]Department of Computer Science, University of Pisa, Pisa.
[4]Policlinico Gemelli IRCCS, Rome, Italy.
[5]Università Cattolica del Sacro Cuore, Rome, Italy.
[6]Studi Michelangelo srl, Firenze, Italy.
[o]These authors share last authorship.

*Corresponding author(s). E-mail(s): giulio.delcorso@isti.cnr.it; michela.franchini@cnr.it;
[†]These authors share the first authorship in ascending order of age.

Abstract

Breast cancer holds the highest diagnosis rate among female tumors and is the leading cause of death among women. Quantitative analysis of radiological images shows the potential to address several medical challenges, including the early detection and classification of breast tumors. In the P.I.N.K study, 66 women were enrolled. Their paired Automated Breast Volume Scanner (ABVS) and Digital Breast Tomosynthesis (DBT) images, annotated with cancerous lesions, populated the first ABVS+DBT dataset. This enabled not only a radiomic analysis for the malignant vs. benign breast cancer classification, but also the comparison of the two modalities.

For this purpose, the models were trained using a leave-one-out nested cross-validation strategy combined with a proper threshold selection approach. This approach provides statistically significant results even with medium-sized data sets. Additionally it provides distributional variables of importance, thus identifying the most informative radiomic features.
The analysis proved the predictive capacity of radiomic models even using a reduced number of features. Indeed, from tomography we achieved AUC-ROC **89.9%** using 19 features and **92.1%** using 7 of them; while from ABVS we attained an AUC-ROC of **72.3%** using 22 features and **85.8%** using only 3 features. Although the predictive power of DBT outperforms ABVS, when comparing the predictions at the patient level, only 8.7% of lesions are misclassified by both methods, suggesting a partial complementarity. Notably, promising results (AUC-ROC ABVS-DBT **71.8%-74.1%**) were achieved using non-geometric features, thus opening the way to the integration of virtual biopsy in medical routine.

Keywords: Breast Cancer, Radiomic, Adaptive Feature Selection, Model Reduction.

1 Introduction

Breast Cancer (BC) is the leading cause of death among women, and according to the Global Burden of Disease 2019, one in every eight new cancer cases was diagnosed as BC, making it the world's top most prevalent type of cancer [1]. There are 5 stages of BC, ranging from the non-invasive ductal carcinoma in situ (stage 0) to the more invasive ones (stages I-IV). To date, stages 0 and I exhibit an almost 100% 5-year survival rate, in contrast to stages II and III which show survival rates of 93% and 72% respectively [1]. Hence, the detection and classification of early-stage BC has a crucial impact on patients' prognosis, as it may allow for less invasive surgical procedures.

The screening procedure relies on the assessment of radiological images, essentially based on mammography (MX) [2, 3], breast ultrasound (US) [4, 5], or contrast-enhanced magnetic resonance imaging (DCE-MRI) [6]. Nowadays, the cornerstone technique for BC screening is MX, which can reduce the mortality rate by 20-22% [7]. However, integrating MX with additional techniques (e.g., Digital Breast Tomosynthesis (DBT), US, or Automated Breast Volume Scanner (ABVS)) can significantly improve detection capability [8, 9]. In particular, DBT by eliminating the problem of tissue overlap and allowing enhancing the identification of parenchymal distortions, increases the Cancer Detection Rate (CDR) of breast lesions by 2.7/1000 compared to MX alone (CDR 5.3/1000) [10]. Also, US-based imaging techniques can improve CDR (4.9 per 100 in a population of women with MX-dense breast) at the cost of a higher false positive rate than DBT [11]. Among the US-based radiological methods, ABVS is a screening technique (for patients with intermediate risk and with MX dense breasts) characterized by a greater reproducibility compared to traditional US [12–14].

While imaging methods are primarily used for screening, biopsy is the only existing tool to classify a breast lesion as benign or malignant and to characterize the malignant ones by receptor expression/phenotype (ER, PR, and HER2 receptor). However, the

biopsy is an invasive, time-consuming, and expensive procedure that can cause anxiety and discomfort to the patient and it's also frequently done needlessly, even for lesions that could be benign [15]. To overcome the cost and limitations of biopsy, the ultimate goals of modern breast imaging encompass the early detection of BC, followed by the accurate classification of the lesion and the prediction of its clinical course and biological aggressiveness.

Among modern image-based mathematical approaches, radiomics is a quantitative approach which uses automated methods to extract valuable information from radiological images. By selecting the most relevant features and embedding them in a Machine Learning (ML) pipeline, radiomics enables the development of predictive models that support standard radiological techniques [16, 17] (e.g. to assess the aggressiveness of cancer lesions). Several published studies have highlighted the potential of radiomics in addressing medical challenges in BC care, such as early detection, classification, cancer sub-type determination and molecular profiling, prediction of lymph node metastases, and prognostication of treatment response [2–6, 18–21].

Despite the large number of published studies, most of the proposed strategies extract quantitative parameters from databases of unimodal medical images (such as DCE-MRI [6, 22], MX [2, 19], and ABVS [14, 23]). Only a minority obtains acquisition from multiple techniques, including DWI + DCE-MRI [18], ABVS + Elastography [24], and BM-US + Elastography [25]. However, there is a paucity of radiomic studies in the scientific literature using ABVS images. Although ABVS and DBT diagnostic performance have been previously compared [26], to the best of our knowledge there is no multimodal ABVS + DBT comparative radiomic analysis.

Difficulties in obtaining paired data, especially with modern techniques like ABVS, lead to undersized study databases. This difficulty exacerbates the already known problems of radiomic studies, in particular, the risk of overfitting given the high ratio of radiomic features to sample size [27–30], and requires ad hoc techniques to maximize information extracted from the database without data leakage.

This work presents a methodological approach for studying radiomic databases of moderate sample size. The approach involves pre-selecting features through stability analysis, designing a validation scheme to maximize extracted information (using nested leave-one-out, LOO, cross validation), and generating distributed importance scores to define an adaptive augmentation procedure. We aimed to differentiate malignant from benign mass lesions using the radiomic features extracted from a medium-sized multi-modal dataset, including DBT and ABVS breast images [31]. To the best of our knowledge, the P.I.N.K database is the first to include both ABVS and DBT acquisitions for each patient. The data, collected in an ongoing longitudinal multicentre study, allow us for a rigorous and significant comparative study of the predictive capabilities of ML models trained on the different modalities.

The paper is structured as follows. The materials and methods section discusses the characteristics of the population and the collection protocol. The analysis of trait stability and the consequent reduction of independent features is included. We then present the nested LOO method, adapted to this database, for generating the distributional feature importance, exploited to select a minimal model using an adaptive procedure. The results show the features most stable to perturbations and the scores

of the trained models: the one with all (non-collinear) features, the model obtained by an adaptive procedure, and the one trained on texture features only. The proposed approach and the related results enabled to draw medical conclusions. Finally, future work may involve the idea of a virtual biopsy, which integrates the texture features-based information with the patient's medical history.

2 Materials and Methods

2.1 Study Population and Acquisition Protocol

The database used in this work, as defined in the P.I.N.K study protocol, includes 66 women over 40 years of age who have both DBT and ABVS acquisition in concurrent periods. The women presented spontaneously for routine breast examination at 2 diagnostic centers in Italy, both equipped with DBT (vendor-independent tomosynthesis - Siemens, GE, Hologic) and ABVS (ABVS ACUSON S2000TM - Siemens Medical Solutions, Inc, Mountain View, CA) devices. The DBT data collect cranio-caudal scans, while the ABVS images are mainly anterior-posterior views. Exclusion criteria include the presence of breast implants, pregnancy, or breastfeeding.

Table 1 Women's distribution by age and breast density. Scores reported as malignant/(malignant + benign).

Age classes	**Breast density (Bi-RADS levels)**				Total
	A	*B*	*C*	*D*	
<50 ys	-	3/4	8/15	0/6	11/25
50-59 ys	1/1	2/4	4/13	2/3	9/21
60-69 ys	2/2	3/5	2/3	-	7/10
>70 ys	1/2	4/4	3/4	-	8/10
Total	4/5	12/17	17/35	2/9	35/66

The ground truth used to train the model is binary tumor classification (malignant/benign, see Table 1), which is obtained from post-intervention histological data (62.3%), if available, or pre-intervention histological data (37.7%), otherwise. The (66) patients included in our study all have mass-forming lesions ($n = 69$), visible both in the DBT images and in the ABVS images.

2.2 Lesion Segmentation

Lesion segmentation in ABVS and DBT images was performed by three radiologists (with over 5 years of experience) in consensus using 3D Slicer [32], a free open source software platform for image analysis and visualization. To ensure a rigorous segmentation process, breast radiologists followed a strict pipeline, consisting of: (i) identifying the lesion in each ABVS and DBT image, (ii) manually delineating the contour of the lesion using annotation tools such as brush and intensity threshold, (iii) assessing the quality of each 2D mask, considering also the lesion volume and shape, (iv) refining and cleaning each segmentation (e.g. holes with an area < 9 pixels have been filled,

and regions of less than 30 pixels have been removed; when necessary, the borders have been smoothed by applying to the segmentation mask a binary closing as implemented in the Multidimensional image processing [1]), (v) validating the lesion segmentation.

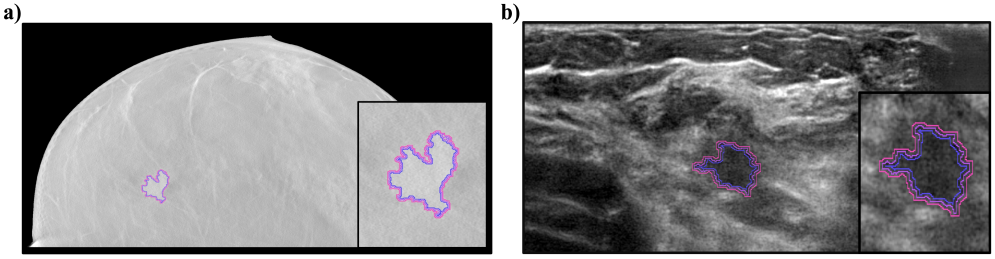

Fig. 1 The same lesion segmented for both modalities: *a)* DBT and *b)* ABVS. The manual segmentation (the standard one) has been slightly modified by applying standard morphological operators to produce two different annotation masks (i.e., reduced and increased) and assess the feature robustness against small variations, as reported in Section 2.3.

2.3 Features Extraction, Selection, and Stability Analysis

Twenty-five radiomic features, both geometric (e.g., Volume, Sphericity) and textural (e.g., Total energy, Coarseness) were extracted using Pyradiomics v.3.0.1 [33]. These features were computed exclusively from the *original images* and are not analytically correlated (Figure 3). We decided to work with a subset of radiomics features computed on original images in order to follow the guidelines of the Image Biomarker Standardization Initiative (IBSI, https://theibsi.github.io/ibsi2/), which have not been released yet for filtered images [34].

The 25 features were subjected to a Principal Component Analysis (PCA) to assess the cardinality of the principal components and thus to identify the presence of linearly correlated features. To determine a subset of features that are non-redundant and stable to small variations, we used the following procedure: (i) the segmentation annotated by radiologists (i.e., *standard*) has been used as the reference mask, (ii) each radiomic feature has been tested varying both the bin width[2] It is defined as: $(max(\text{gray level}) - min(\text{gray level}))/\#bins$. and the mask type (reduced, standard, increased), as shown in Figure 1, (iii) for each combination of bin width ($bw \in \{15,20,25,30,35\}$) and mask ($m \in \{\text{reduced, standard, increased}\}$), the instability of the i^{th} feature Δ_i is estimated as follow:

$$\frac{1}{15 \cdot \#p} \sum_{m,bw,p} \frac{|\mathrm{f}_i(m,bw,p) - \mathrm{f}_i(st,25,p)|}{\mathbf{max}_{p^*}(\mathrm{f}_i(st,25,p^*)) - \mathbf{min}_{p^*}(\mathrm{f}_i(st,25,p^*))} \tag{1}$$

where $st := standard$, $\#p$ the number of patients, and $f_i(m,bw,p)$ is the i^{th} feature calculated using the mask size m, the bin width bw on the patient p, (iv) between two highly correlated features, the most unstable is defined as the one with the higher

[1] https://docs.scipy.org/doc/scipy/reference/ndimage.html
[2] *Bin width* is the width of each bin in the histogram that represents the gray level intensities of the image.

value of Δ and therefore is dropped. After variable reduction, PCA was repeated to ensure the redundancy of the dropped features.

2.4 Nested LOO Cross Validation

Tumor classification was performed using three ML approaches: Random Decision Forest (RDF , an ensemble of n_t random trees, with $n_t \in [50, 250]$), polynomial Support Vector Machine (SVM, with a polynomial kernel of degree 3 with a cost of $c \in [1, 10]$), and Logistic Model (Logit, binary classifier based on a binomial general linear model). These methods were trained on both DBT- and ABVS-derived radiomic features. However, the high ratio ($\sim 1/3$) between the number of independent features and the sample size, as well as the high variability of the population under study (regarding lesion shape, breast density, extension, branching, etc.), require the use of a robust *ad-hoc* technique for the optimization of the classifier's hyperparameters (i.e., the number of trees for the RDF and the cost for the SVM) and the subsequent evaluation of the performance of the models.

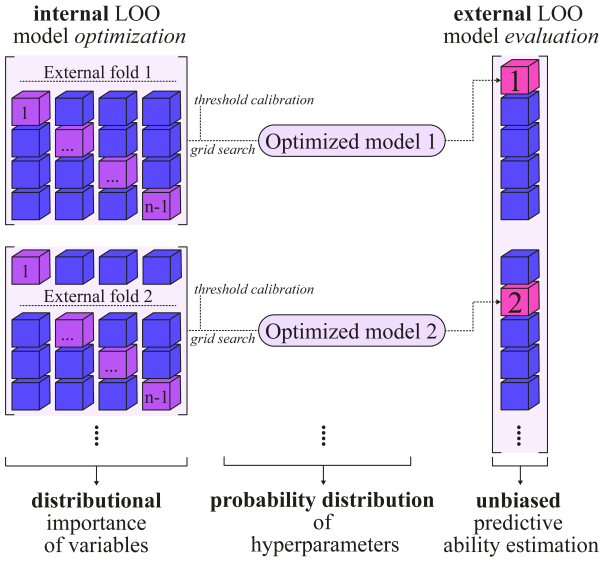

Fig. 2 Optimization and validation scheme adapted to the dimension ($n = 69$) of the dataset. The external LOO model evaluation uses a leave-one-out approach to provide an estimate of performance for each patient, at the cost of increased computational complexity. Each model evaluation is performed after an optimization (i.e., internal LOO) using an additional LOO strategy combined with a grid search for the optimal hyperparameter. To reduce positively biased estimates, every optimized model calibrates its internal threshold.

Referring to Figure 2, a LOO nested cross-validation approach was defined as an extension of classical nested cross-validation [35]. In order to adapt the implementation to the available data, the described procedure was implemented in-house from scratch using R (v4.2.2) and Python (v3.9.13). The dataset (of cardinality n) is partitioned into n variants (called external LOOs) by a leave-one-out scheme. For each

external LOO, $n-1$ data points are used as training. One data point is treated as an independent evaluation to estimate the generalization ability of the model. The model used in each step of the external LOO (i.e., optimized model i) is obtained by calibrating the hyperparameters using a grid search. The latter is defined as the one that maximizes the performance (i.e., the AUC-ROC, Area Under the Receiver Operating Characteristic Curve) in the internal LOO cross-validation. The same internal LOO cross-validation is used to estimate the i^{th} optimal classification threshold for the corresponding model (i.e., the probability value that discriminates categories) as the one maximizing the sum of specificity and sensitivity across the $n-1$ LOO evaluations. The i^{th} external LOO model evaluation then performs a prediction estimate based on the optimized i^{th} model (grid search + internal LOO cross validation) with the appropriately calibrated i^{th} threshold.

This approach provides a low-bias estimate of the model's generalization capability from the external LOO procedure, where the data employed is never used in the optimization and training phases and is therefore not affected by data leakage. Similarly, to further reduce positively-biased results, the optimal threshold is calibrated for each model (internal LOO model optimization). In addition, unlike the canonical partitioning of the available data into training, validation, and test-set required for both optimization and external validation, the use of nested leave-one-out cross-validation makes the most of the information content of this dataset.

A major drawback is that the computational cost of nested LOO cross-validation is quadratic in dataset cardinality. Indeed, for each choice (#GridSearch) of hyperparameters to be adjusted, each optimization of the internal LOO model requires $n-1$ model training sessions. For each external LOO model evaluation, this procedure must be performed n times. The total number of model training sessions is: $\#\text{GridSearch} \cdot n \cdot (n-1)$.

2.5 Distributional Feature Importance and Adaptive Selection

The optimization/validation scheme used does not produce a single optimal model, but rather a family of models (i.e., n=69 different models generated by the external LOO, one for each element in the dataset). Each of these models is obtained by nested LOO cross-validation and is the optimized model on the remaining $n-1$ (68) training data. The relative importance of the features can be calculated in an analytical way for each trained model. Features importance are: (i) the Mean Decrease Gini/Accuracy for RDF, (ii) SVM coefficient multiplied by its support vector for SVM, (iii) change in deviance (i.e., reduction of prediction error) for Logit. The external LOO scheme can be used to compute the probability distribution of the importance for each input feature (obtained from the $n = 69$ LOO-models) which is more informative than a canonical point estimate obtained from a single retrained model. Furthermore, the resulting distributions can be used to analyze the stability of the model. Indeed, low distribution variance corresponds to a feature whose relative importance is constant across trained models.

The resulting relative importance is used to define an adaptive procedure for the selection of a subset of the features.

Algorithm 1 Adaptive feature selection

Require: $\mathbf{v} = \{v_j\}_{j \in [0,m]}$ ▷ Features ordered by importance
Ensure: $\mathbf{v}^*$ ▷ Selected features
  $\mathbf{v}^* \leftarrow$ 2-Cluster($\{v_j\}$) ▷ Starting relevant features
  $k \leftarrow \#(\mathbf{v}^*)$ ▷ Cardinality of $\mathbf{v}^*$
  best_Aucroc $\leftarrow$ **AUCROC_NestedLOO**$(M(\mathbf{v}^*))$
  for $s \in [k, m]$ **do**
    temp_Aucroc $\leftarrow$ **AUCROC_NestedLOO**$(M(\mathbf{v}^*, v_s))$
    if temp_Aucroc $>$ best_Aucroc **then**
      $\mathbf{v}^* \leftarrow \{\mathbf{v}^*, v_s\}$ ▷ Add the new feature
    end if
  end for

A 2-clustering algorithm divides the features into 2 groups (most relevant/less relevant), as reported in Algorithm 1. The most relevant features are added to the model, whose performance is evaluated using the nested LOO cross validation procedure previously introduced. The performance score used is the LOO-AUCROC obtained from the external LOO cycle (summarized by the function **AUCROC_NestedLOO** in the algorithm). The procedure adds the features to the model in the order of their relative importance (mean decrease accuracy). After adding each feature, the performance score is evaluated: if the model performs better, the feature is retained; if not, it is dropped. This procedure returns the minimal model with the best performance by performing a total of $86940 * 19 \sim 2 \cdot 10^6$ simulations ($\sim$ 13 hours on 1.6 GHz Dual Core Intel i5 CPU). In contrast, a brute-force exploration of all the combinations of features would have required $86940 \cdot 2^{19} \sim 10^{10}$ simulations.

3 Results

3.1 Feature Selection, Redundancy Correction, and Stability Analysis

Among the starting subset of original features reported in Figure 3, we performed a selection based on the elimination of redundant ones. In particular, within pairs of correlated variables (Pearson Correlation above 0.95), we decided to keep the feature more stable with respect to the stability measure (Section 2.3).
The PCA analysis showed that the number of principal components (99% of the cumulative variance) does not change before and after the redundancy correction (16 components for ABVS and 14 for DBT).
For ABVS, redundant features are: the Intensity Histogram Entropy (highly correlated with the Intensity Histogram Uniformity), the Median (highly correlated with the Root Mean Squared (RMS)) and GLSZM LAE - Large Area Emphasis (highly correlated with GLSZM LALGLE - Large Area Low Gray Level Emphasis). For DBT, redundant features are: the Total Energy (highly correlated with the Volume), the Intensity Histogram Entropy (highly correlated with the Intensity Histogram Uniformity), the Median (highly correlated with the Root Mean Squared (RMS)), GLSZM

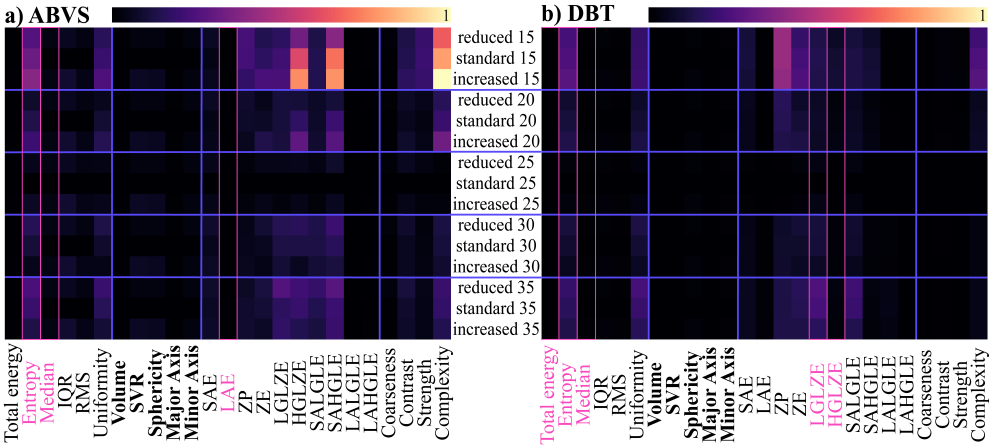

Fig. 3 The score defined in Eq. 1 has been computed to assess the radiomic feature stability for ABVS (Panel *a*) and DBT (Panel *b*). Each row corresponds to a different extraction: reduced/standard/increased represents which mask was used in the computation, while the numbers [15-35] are bin width used to extract the features. Geometric features are shown in bold. Features dropped after the redundancy correction are marked with a pink box.

LGLZE - Low Gray Level Zone Emphasis (highly correlated with GLSZM SALGLE - Small Area Low Gray Level Emphasis), GLSZM HGLZE - High Gray Level Zone Emphasis (highly correlated with GLSZM SAHGLE (Small Area High Gray Level Emphasis).

Figure 3 represents the heat maps of the stability of the features with respect to the bin width and the perturbation (increase/decrease) of the segmentation mask. They show that the shape features (in bold) are generally more stable compared to the texture ones. Notably, the least variability induced by geometrical perturbation of the mask is obtained with the default value of Pyradiomics for the bin width (25).

3.2 Full models: ABVS-DBT Comparison

The performance of the models trained on the same set of patients (called respectively **RDF-ABVS/SVM-ABVS/Logit-ABVS Model** and **RDF-DBT/SVM-DBT/Logit-DBT Model**) are reported in Table 3.2.

For the three models, the DBT Models always outperform the ABVS ones. Indeed, DBT-based models have a higher AUC-ROC compared to ABVS for each of the trained ML methods (69.9/73.0/89.9% vs 67.8/66.7/72.3% for SVM/Logit/RDF respectively). DBT is also better than ABVS at identifying pathological cases for all three models (94.7/81.6/84.2% vs 71.0/55.3/68.4% for SVM/Logit/RDF respectively), while the number of false positives is comparable between DBT and ABVS. The comparison between the three ML models highlights the strength of ensemble methods, with RDF proving to be the most accurate in almost all performance metrics, particularly for DBT data (AUC-ROC 89.9%, Accuracy/Specificity/Precision and Recall $> 80\%$). In addition, the simplest model (Logit) has very low accuracy, precision and recall.

As reported in Section 2.5, we calculated the Distributional Feature Importance for DBT and ABVS (see Figure 4). In terms of the most effective model, it emerges

Table 2 Full RDF-, SVM-, and Logit-ABVS and DBT Models. The best full model (RDF) is also retrained on the texture feature only (RDF-ABVS tx. and RDF-DBT tx.) and using the adaptive feature selection strategy. The performance metrics are computed by selecting the threshold using the inner nested LOO evaluation.

	AUC-ROC	**Accuracy**	**Specificity**	**Precision**	**Recall**
SVM-ABVS	67.8%	68.1%	64.5%	71.0%	71.0%
SVM-DBT	69.9%	72.5%	58.0%	67.9%	**94.7**%
Logit-ABVS	66.7%	58.0%	61.3%	63.6%	55.3%
Logit-DBT	73.0%	73.9%	64.5%	73.8%	81.6%
RDF-ABVS	72.3%	68.1%	67.7%	72.2%	68.4%
RDF-DBT	**89.9**%	**80.7**%	**80.7**%	**84.2**%	84.2%
RDF-ABVS tx.	71.8%	71.1%	77.1%	70.3%	62.1%
RDF-DBT tx.	74.1%	74.4%	82.9%	75.0%	65.5%
RDF-ABVS Reduced	85.8%	76.8%	71.0%	77.5%	81.6%
RDF-DBT Reduced	**92.1**%	**85.5**%	**87.1**%	**88.9**%	84.2%

that Sphericity is the most important feature and, particularly for DBT acquisitions, geometric features play an important role in model classification. Among the other features, glszm SAE (Small Area Emphasis) and Strength, are relevant for Full ABVS Model. The high concordance of Mean Decrease Accuracy and Mean Decrease Gini is a good indicator of models stability. Similar results are obtained for the SVM, with sphericity being the most important feature for ABVS and DBT. Conversely, Logit mainly uses non-geometric NGTDM features to make its prediction, but this leads to unreliable predictions and affects both model accuracy and specificity.

The uniqueness of the P.I.N.K dataset (consisting of a dual DBT+ABVS acquisitions) allows for a comparison of the most effective models (RDF) at the patient level. It can be verified that 26.1% of the lesions (4 malignant cases and 8 benign cases correctly predicted *only by* DBT and 3 malignant cases and 3 benign cases correctly predicted *only by* ABVS) are correctly identified by only one of the two models. Conversely, only 8.7% (3 malignant and 3 benign cases) are misclassified by both models. Consequently, even considering the better performance of the Full RDF-DBT Model compared to the RDF-ABVS one, this suggests that the two modalities are partially complementary.

3.3 Reduced models: Adaptive Features Selection

An optimal set of the non-collinear features (Figure 4) was identified by applying the Adaptive Selection Algorithm 1 to the most effective full model (i.e., ABVS/DBT RDF). The corresponding models trained on these subsets are the **RDF-ABVS Reduced Model** and the **RDF-DBT Reduced Model**. Starting from *Sphericity* for DBT and *Sphericity, glszm SAE, and Strength* for ABVS, the procedure iteratively adds only features with positive AUC-ROC contribution. The reduced ABVS model has an AUC-ROC of 85.8% using only the following 3 starting features: Sphericity, SAE (Small Area Emphasis), and Strength. On the other hand, the reduced DBT model uses 7 features to obtain an even higher AUC-ROC of 92.1%: Sphericity, LALGLE (Large Area Low Gray Level Emphasis), ZP (Zone Percentage), ZE (Zone Entropy), Coarseness, SAE (Small Area Emphasis), and RMS (Root Mean Squared). Notably,

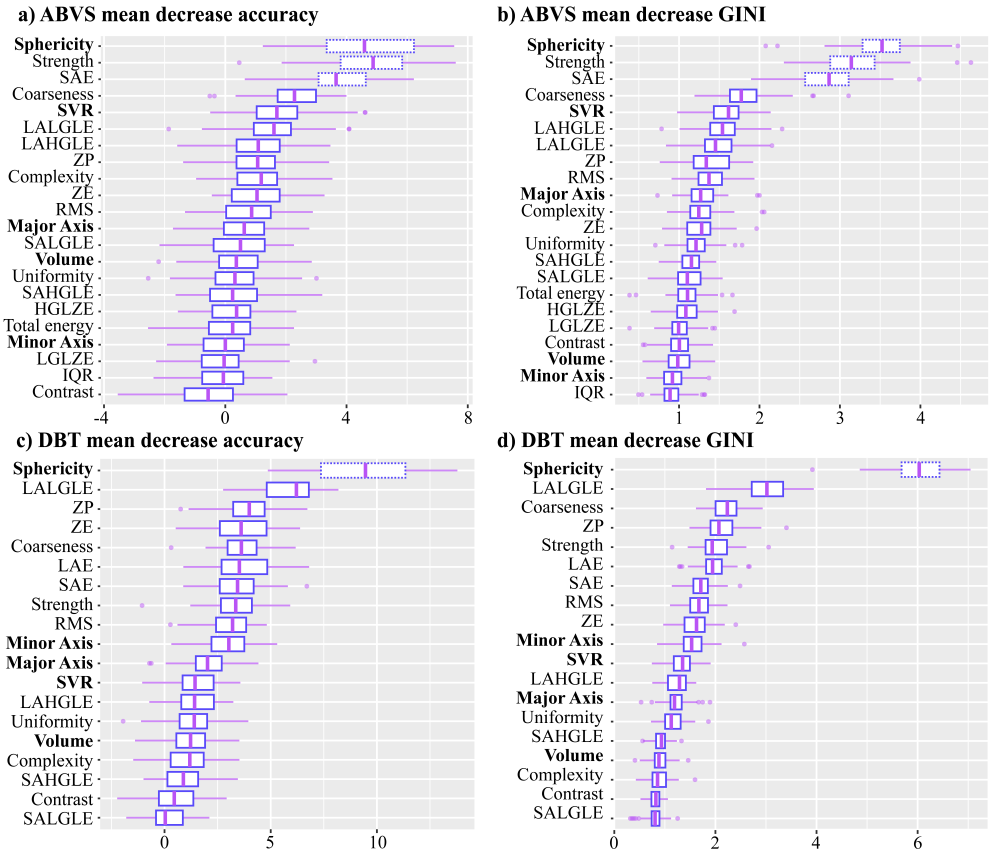

Fig. 4 Distributional Feature Importance of RDF-ABVS analysis (panels a-b) and RDF-DBT (panels c-d). The scores are reported as mean value and IQR (3° and 4° quantiles), calculated from the nested LOO external procedure. Dashed boxes indicate features that are significantly more relevant features (Sphericity, SAE, and Strength for ABVS, while Sphericity for DBT).

the selected features include a mixture among geometric (Sphericity), neighbouring gray tone difference (Coarseness), and texture features (both on small and large areas) comprehensively covering the different radiomic characteristics of the lesion. Both reduced models perform better than the full models because the excluded features degrade the classification (Figure 5).

3.4 Texture models: Towards a virtual biopsy

Currently, biopsy is the standard technique for lesion classification and characterization, focusing on local and visual features of the sampled tissue. Hence, to simulate the biopsy, it would be sufficient to consider only the texture features of the segmented lesion. To investigate and compare the informative content of the texture and geometric features (e.g., Volume, Sphericity), the **RDF-ABVS Texture Model** and **RDF-DBT Texture Model** were introduced. These are the most effective models

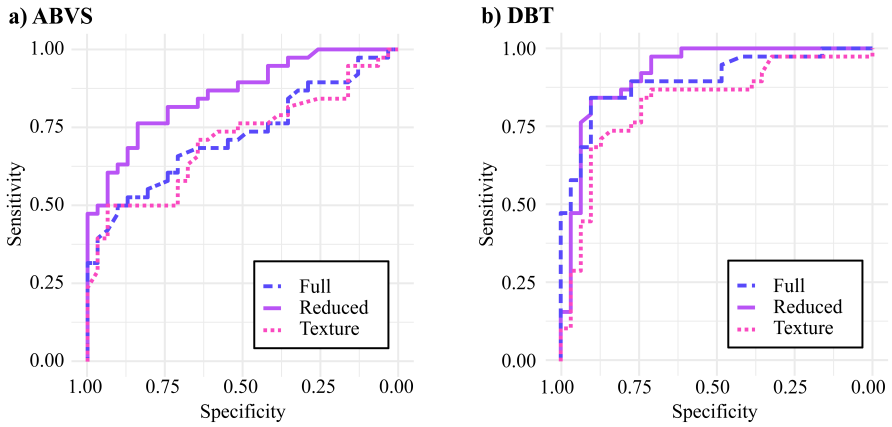

Fig. 5 Receiver Operating Characteristic (ROC) curves for RDF-ABVS (Panel a) and RDF-DBT (panel b). These curves represent the performance on the LOO external validation for the Full RDF model (trained on the whole set of radiomic features), the Reduced RDF model (features obtained from the adaptive procedure), and the RDF Texture one (without geometrical/shape features).

(i.e., RDF) that have been retrained to use only texture features to make their predictions. The Texture Models, even neglecting the significant contribution of geometric features, prove to train adequate classifiers (Figure 5). Indeed, the AUC-ROC of the RDF-DBT Texture Model is 74.1% (compared to 89.9% of the RDF-DBT Model). Similarly, the AUC-ROC of the RDF-ABVS Texture Model is 71.8% (compared to 72.3% of the RDF-ABVS Model). Note that the ABVS/DBT performance gap is reduced when the geometric features are neglected.

4 Discussion

In this work, we compared ABVS/DBT capability to classify benign/malignant breast tumors in a population of 66 women (69 lesions) using radiomic features. Three Machine Learning method were employed: Random Decision Forests, Support Vector Machines and Logistic Regression. They were trained and validated on a novel dataset of paired ABVS/DBT acquisitions using an *ad hoc* nested LOO cross-validation procedure. This approach allows us to avoid data-leakage among training and validation sets and, consequently, to obtain a low-biased estimate of generalization capability of the model even with a small sample size. Furthermore, an adaptive selection strategy was successfully applied to obtain a minimal highly informative subset of features, so that derived models were computationally lighter and less affected by overfitting. The first major finding of this study is to highlight the greater effectiveness of ensemble methodology (RDF) to provide efficient prediction of tumor classification using radiomic features compared to single-prediction methods (Logit, SVM). RDF radiomic-based models for both ABVS and DTB acquisition prove to efficiently discriminate malignant/benign lesions (AUC-ROC: RDF-ABVS 72.3%, RDF-DBT 89.9%, using respectively 22/19 features). Nevertheless, even this reduced set of features is likely to contain redundant information. In fact, the adaptive selection strategy leads to a minimal

subset of features with even greater classification power compared to the full set (AUC-ROC: ABVS 85.8% with 3 features, DBT 92.1% with 7 features). The latter suggests the importance of complementing classical radiomic analyses (based on hundreds or even thousands of features) with appropriate selection strategies to reduce the presence of confounding variables, especially in small/medium size datasets. As detailed in Section 3, independently of the set (or subset) of features used to train the classification model, using DBT data resulted in higher classification performances with respect to ABVS data, almost surely due to the image resolution. However, some kind of complementarity cannot be excluded: when comparing the predictions at a patient level, only 8.7% of lesions are misclassified by both the Full RDF Models. Finally, the removal of the (highly influential) geometric information from the model results in less accurate but still valid predictions (AUC-ROC: RDF-ABVS tx. 71.8%, RDF-DBT tx. 74.1%). This confirms radiomics as a tool capable of extracting information beyond the human eye. A limitation of this work is the size of the dataset (66 subjects, 69 lesions) deeply influenced by the difficulty of collecting reliable annotated images from DBT and ABVS covering the same lesions and by the time-consuming and demanding clinician-guided image segmentation process. Consequently, we focused on a binary classification task instead of a more complex tumor stage stratification. Such a limitation has been mitigated by LOO cross-validation and the use of the adaptive selection algorithm; of course, a larger study population will increase the statistical power of the results. In this respect, the ongoing activities of the P.I.N.K project will help, by collecting multimodal data from additional centers.

Future directions of research aim at the development of a mixed model of ABVS and/or DBT that includes also patient clinical history. In this perspective, a larger dataset is crucial also to tackle the more difficult task of a multi-class analysis, to enable both phenotype and tumor stage characterization. The results described and discussed above indicate that for BC the **virtual biopsy**, i.e., a radiomic-based ML, which uses only image data to characterize the lesion, is not so far.

Declarations

Ethical Approval

The P.I.N.K. study was submitted and approved by the ethical committees of each participating center and by the CNR Research Ethics and Integrity Committee (Prot. n. 0065051/2018 on the 4th of October 2018). The main approval was provided by the Regional Ethics Committee for Clinical Trials of the Tuscany Region CEAVNO (Prot. n. 9047 on the 19th of February 2018). The study is being conducted in line with the principles set out in the original Declaration of Helsinki and later amendments. Informed consent has been obtained by all participants and data are handled accordingly. Prior to any preprocessing step, each imaging study was pseudoanonymized.

Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Acknowledgments and Fundings

This work was sponsored and co-founded by the Umberto Veronesi Foundation.

Author Contributions

Study definition and research objectives: CC, DG, SC, MAP, GDC, SPi, MF. Data collection and curation: MAP, SPi, MF, GA, PB, AN, SPa, CT. Funding acquisition/Project administration: MF, SM, SPi. Methodology development: CC, DG, GDC, SC. Software implementation: CC, GDC. Manuscript preparation/Analysis: CC, DG, GDC, MAP.

References

[1] Alkabban, F., Ferguson, T.: Breast Cancer. Treasure Island (FL): StatPearls Publishing, PMID: 29493913 (2023)

[2] Ma, W., Zhao, Y., Ji, Y., Guo, X., Jian, X., Liu, P., Wu, S.: Breast cancer molecular subtype prediction by mammographic radiomic features. Academic radiology **26**(2), 196–201 (2019)

[3] Wang, G., Shi, D., Guo, Q., Zhang, H., Wang, S., Ren, K.: Radiomics based on digital mammography helps to identify mammographic masses suspicious for cancer. Frontiers in oncology **12**, 843436 (2022)

[4] Romeo, V., Cuocolo, R., Apolito, R., Stanzione, A., Ventimiglia, A., Vitale, A., Verde, F., Accurso, A., Amitrano, M., Insabato, L., *et al.*: Clinical value of radiomics and machine learning in breast ultrasound: a multicenter study for differential diagnosis of benign and malignant lesions. European Radiology **31**, 9511–9519 (2021)

[5] Lee, S.E., Han, K., Kwak, J.Y., Lee, E., Kim, E.-K.: Radiomics of us texture features in differential diagnosis between triple-negative breast cancer and fibroadenoma. Scientific reports **8**(1), 1–8 (2018)

[6] Braman, N., Etesami, M., Prasanna, P., Dubchuk, C., Gilmore, H., Tiwari, P., Pletcha, D., Madabhushi, A.: Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast dce-mri. Breast Cancer Research **19** (2017)

[7] Marmot, M.G., Altman, D., Cameron, D., Dewar, J., Thompson, S., Wilcox, M.: The benefits and harms of breast cancer screening: an independent review. British journal of cancer **108**(11), 2205–2240 (2013)

[8] Buchberger, W., Geiger-Gritsch, S., Knapp, R., Gautsch, K., Oberaigner, W.: Combined screening with mammography and ultrasound in a population-based screening program. European Journal of Radiology **101**, 24–29 (2018) https://doi.org/10.1016/j.ejrad.2018.01.022

[9] Tan, T., Rodriguez-Ruiz, A., Zhang, T., Xu, L., Beets-Tan, R.G., Shen, Y., Karssemeijer, N., Xu, J., Mann, R.M., Bao, L.: Multi-modal artificial intelligence for the combination of automated 3d breast ultrasound and mammograms in a population of women with predominantly dense breasts. Insights into Imaging **14**(1), 10 (2023)

[10] Ciatto, S., Houssami, N., Bernardi, D., Caumo, F., Pellegrini, M., Brunelli, S., Tuttobene, P., Bricolo, P., Fantò, C., Valentini, M., *et al.*: Integration of 3d digital mammography with tomosynthesis for population breast-cancer screening (storm): a prospective comparison study. The lancet oncology **14**(7), 583–589 (2013)

[11] Tagliafico, A.S., Mariscotti, G., Valdora, F., Durando, M., Nori, J., La Forgia, D., Rosenberg, I., Caumo, F., Gandolfo, N., Sormani, M.P., *et al.*: A prospective comparative trial of adjunct screening with tomosynthesis or ultrasound in women with mammography-negative dense breasts (astound-2). European Journal of Cancer **104**, 39–46 (2018)

[12] Zelst, J.C., Mann, R.M.: Automated three-dimensional breast us for screening: technique, artifacts, and lesion characterization. Radiographics **38**(3), 663–683 (2018)

[13] Brem, R.F., Tabár, L., Duffy, S.W., Inciardi, M.F., Guingrich, J.A., Hashimoto, B.E., Lander, M.R., Lapidus, R.L., Peterson, M.K., Rapelyea, J.A., *et al.*: Assessing improvement in detection of breast cancer with three-dimensional automated breast us in women with dense breast tissue: the somoinsight study. Radiology **274**(3), 663–673 (2015)

[14] Chen, Y., Xie, Y., Li, B., Shao, H., Na, Z., Wang, Q., Jing, H.: Automated breast ultrasound (abus)-based radiomics nomogram: an individualized tool for predicting axillary lymph node tumor burden in patients with early breast cancer. BMC cancer **23**(1), 340 (2023)

[15] Hemmer, J.M., Kelder, J.C., Heesewijk, H.P.: Stereotactic large-core needle breast biopsy: analysis of pain and discomfort related to the biopsy procedure. European radiology **18**, 351–354 (2008)

[16] Ibrahim, A., Primakov, S., Beuque, M., Woodruff, H., Halilaj, I., Wu, G., Refaee, T., Granzier, R., Widaatalla, Y., Hustinx, R., *et al.*: Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. Methods **188**, 20–29 (2021)

[17] Pesapane, F., De Marco, P., Rapino, A., Lombardo, E., Nicosia, L., Tantrige, P., Rotili, A., Bozzini, A.C., Penco, S., Dominelli, V., *et al.*: How radiomics can improve breast cancer diagnosis and treatment. Journal of Clinical Medicine **12**(4), 1372 (2023)

[18] Fan, M., Yuan, W., Zhao, W., Xu, M., Wang, S., Gao, X., Li, L.: Joint prediction of breast cancer histological grade and ki-67 expression level based on dce-mri and dwi radiomics. IEEE journal of biomedical and health informatics **24**(6), 1632–1642 (2019)

[19] Mao, N., Yin, P., Wang, Q., Liu, M., Dong, J., Zhang, X., Xie, H., Hong, N.: Added value of radiomics on mammography for breast cancer diagnosis: a feasibility study. Journal of the American College of Radiology **16**(4), 485–491 (2019)

[20] Lee, J.Y., Lee, K.-s., Seo, B.K., Cho, K.R., Woo, O.H., Song, S.E., Kim, E.-K., Lee, H.Y., Kim, J.S., Cha, J.: Radiomic machine learning for predicting prognostic biomarkers and molecular subtypes of breast cancer using tumor heterogeneity and angiogenesis properties on mri. European radiology **32**, 650–660 (2022)

[21] Whitney, H.M., Taylor, N.S., Drukker, K., Edwards, A.V., Papaioannou, J., Schacht, D., Giger, M.L.: Additive benefit of radiomics over size alone in the distinction between benign lesions and luminal a cancers on a large clinical breast mri dataset. Academic radiology **26**(2), 202–209 (2019)

[22] Cain, E.H., Saha, A., Harowicz, M.R., Marks, J.R., Marcom, P.K., Mazurowski, M.A.: Multivariate machine learning models for prediction of pathologic response to neoadjuvant therapy in breast cancer using mri features: a study using an independent validation set. Breast cancer research and treatment **173**, 455–463 (2019)

[23] Wang, S.-j., Liu, H.-q., Yang, T., Huang, M.-q., Zheng, B.-w., Wu, T., Han, L.-q., Zhang, Y., Ren, J.: Machine learning based on automated breast volume scanner (abvs) radiomics for differential diagnosis of benign and malignant birads 4 lesions. International Journal of Imaging Systems and Technology **32**(5), 1577–1587 (2022)

[24] Ma, Q., Shen, C., Gao, Y., Duan, Y., Li, W., Lu, G., Qin, X., Zhang, C., Wang, J.: Radiomics analysis of breast lesions in combination with coronal plane of abvs and strain elastography. Breast Cancer: Targets and Therapy, 381–390 (2023)

[25] Jiang, M., Li, C.-L., Chen, R.-X., Tang, S.-C., Lv, W.-Z., Luo, X.-M., Chuan, Z.-R., Jin, C.-Y., Liao, J.-T., Cui, X.-W., *et al.*: Management of breast lesions seen on us images: dual-model radiomics including shear-wave elastography may match performance of expert radiologists. European Journal of Radiology **141**, 109781 (2021)

[26] Rahmat, K., Ab Mumin, N., Ng, W.L., Taib, N.A.M., Chan, W.Y., Hamid, M.T.R.: Automated breast ultrasound provides comparable diagnostic performance in opportunistic screening and diagnostic assessment. Ultrasound in Medicine & Biology **50**(1), 112–118 (2024)

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

[28] Chatterjee, A., Vallières, M., Dohan, A., Levesque, I.R., Ueno, Y., Bist, V., Saif, S., Reinhold, C., Seuntjens, J.: An empirical approach for avoiding false discoveries when applying high-dimensional radiomics to small datasets. IEEE Transactions on Radiation and Plasma Medical Sciences **3**(2), 201–209 (2018)

[29] Traverso, A., Wee, L., Dekker, A., Gillies, R.: Repeatability and reproducibility of radiomic features: a systematic review. International Journal of Radiation Oncology* Biology* Physics **102**(4), 1143–1158 (2018)

[30] Ubaldi, L., Valenti, V., Borgese, R., Collura, G., Fantacci, M., Ferrera, G., Iacoviello, G., Abbate, B., Laruina, F., Tripoli, A., *et al.*: Strategies to develop radiomics and machine learning models for lung cancer stage and histology prediction using small data samples. Physica Medica **90**, 13–22 (2021)

[31] Franchini, M., Pieroni, S., Montrucchio, E., Nori Cucchiari, J., Di Maggio, C., Cassano, E., Di Nubila, B., Giuseppetti, G.M., Nicolucci, A., Scaperrotta, G., *et al.*: The pink study approach for supporting personalized risk assessment and early diagnosis of breast cancer. International journal of environmental research and public health **18**(5), 2456 (2021)

[32] Kikinis, R., Pieper, S.D., Vosburgh, K.G.: In: Jolesz, F.A. (ed.) 3D Slicer: A Platform for Subject-Specific Image Analysis, Visualization, and Clinical Support, pp. 277–289. Springer, New York, NY (2014). https://doi.org/10.1007/978-1-4614-7657-3_19

[33] Van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.-C., Pieper, S., Aerts, H.J.: Computational radiomics system to decode the radiographic phenotype. Cancer research **77**(21), 104–107 (2017)

[34] 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)

[35] Stone, M.: Cross-validatory choice and assessment of statistical predictions. Journal of the royal statistical society: Series B (Methodological) **36**(2), 111–133 (1974)