







Article

Accurate Evaluation of Feature Contributions for Sentinel Lymph Node Status Classification in Breast Cancer

Angela Lombardi ^{1,2} , Nicola Amoroso ^{2,3} , Loredana Bellantuono ^{2,4}, Samantha Bove ⁵, Maria Colomba Comes ⁵, Annarita Fanizzi ^{5,*} , Daniele La Forgia ⁵ , Vito Lorusso ⁵, Alfonso Monaco ^{1,2} , Sabina Tangaro ^{2,6} , Francesco Alfredo Zito ⁵, Roberto Bellotti ^{1,2,†} and Raffaella Massafra ^{5,†}

- ¹ Dipartimento Interateneo di Fisica, Università degli Studi di Bari, Via E. Orabona 4, 70125 Bari, Italy; angela.lombardi@uniba.it (A.L.); alfonso.monaco@ba.infn.it (A.M.); roberto.bellotti@uniba.it (R.B.)
- ² Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Via E. Orabona 4, 70125 Bari, Italy; nicola.amoroso@uniba.it (N.A.); loredana.bellantuono@uniba.it (L.B.); sabina.tangaro@uniba.it (S.T.)
- ³ Dipartimento di Farmacia-Scienze del Farmaco, Università degli Studi di Bari Aldo Moro, Via A. Orabona 4, 70125 Bari, Italy
- ⁴ Dipartimento di Scienze Mediche di Base, Neuroscienze e Organi di Senso, Università degli Studi di Bari Aldo Moro, Piazza G. Cesare 11, 70124 Bari, Italy
- ⁵ I.R.C.C.S. Istituto Tumori “Giovanni Paolo II”, Viale Orazio Flacco 65, 70124 Bari, Italy; s.bove@oncologico.bari.it (S.B.); m.c.comes@oncologico.bari.it (M.C.C.); d.laforgia@oncologico.bari.it (D.L.F.); vitolorusso@me.com (V.L.); a.zito@oncologico.bari.it (F.A.Z.); r.massafra@oncologico.bari.it (R.M.)
- ⁶ Dipartimento di Scienze del Suolo, della Pianta e degli Alimenti, Università degli Studi di Bari Aldo Moro, Via A. Orabona 4, 70125 Bari, Italy
- * Correspondence: a.fanizzi@oncologico.bari.it
- † These authors contributed equally to this work.



Citation: Lombardi, A.; Amoroso, N.; Bellantuono, L.; Bove, S.; Comes, M.C.; Fanizzi, A.; La Forgia, D.; Lorusso, V.; Monaco, A.; Tangaro, S.; et al. Accurate Evaluation of Feature Contributions for Sentinel Lymph Node Status Classification in Breast Cancer. *Appl. Sci.* **2022**, *12*, 7227. <https://doi.org/10.3390/app12147227>

Academic Editor: Maria Rizzi

Received: 25 May 2022

Accepted: 14 July 2022

Published: 18 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: The current guidelines recommend the sentinel lymph node biopsy to evaluate the lymph node involvement for breast cancer patients with clinically negative lymph nodes on clinical or radiological examination. Machine learning (ML) models have significantly improved the prediction of lymph nodes status based on clinical features, thus avoiding expensive, time-consuming and invasive procedures. However, the classification of sentinel lymph node status represents a typical example of an unbalanced classification problem. In this work, we developed a ML framework to explore the effects of unbalanced populations on the performance and stability of feature ranking for sentinel lymph node status classification in breast cancer. Our results indicate state-of-the-art AUC (Area under the Receiver Operating Characteristic curve) values on a hold-out set (67%) while providing particularly stable features related to tumor size, histological subtype and estrogen receptor expression, which should therefore be considered as potential biomarkers.

Keywords: sentinel lymph node; imbalanced dataset; data augmentation; breast cancer; machine learning; interpretability

1. Introduction

Current recommendations advise sentinel lymph node biopsy (SLNB) for breast cancer patients experiencing clinically negative lymph nodes on clinical or radiological evaluation [1]. While this procedure is currently the most accurate assessment, it is also the most expensive and time-consuming, as well as the most invasive, with a variety of potential adverse effects [2,3]. In addition, in patients with early-stage breast cancer, the incidence of axillary metastases is around 10–25% [4,5], therefore SLNB is often an unnecessary procedure. In this context, the definition of less expensive and invasive alternative procedures for the prediction of lymph node involvement represents a current task of interest for the clinical and scientific community.

In literature, several models have been proposed to predict non-sentinel lymph nodes status by using different features [6,7]. On the other side, a small number of studies with the goal of predicting the sentinel lymph nodes status have yielded promising findings [8–10].

Nevertheless, the classification of sentinel lymph nodes provides a typical example of an unbalanced classification problem. In healthcare data analysis applications including cancer diagnosis and disease risk prediction, classifying unbalanced datasets has been identified as a major challenge. When the quantity of samples in a dataset is substantially uneven, an unbalanced classification problem emerges. In binary classification tasks, the unbalanced classification problem occurs when one class (i.e., the minority class) has much fewer observations than the other class (i.e., the majority class) [11]. For unbalanced datasets, most of the machine learning (ML) algorithms usually perform poorly since they aim to optimize overall classification accuracy and thus assume an equal incidence of all classes. As a result, these methods exhibit classification bias towards the majority class, failing to identify examples of the minority class [12].

Several strategies have been proposed to overcome these limitations. Resampling methods have been introduced to rebalance the datasets. Resampling schemes include random oversampling of the minority class, undersampling (or subsampling) of the majority class and some advanced synthetic sampling methods that aim at rebalancing the class distribution at the data level. These rebalancing solutions, however, present several drawbacks. For example, loss of information is an inherent consequence of undersampling [13], but oversampling by randomly repeating the minority class sample could result in overfitting [14].

Although several works have investigated the effect of different data balancing strategies on the overall classification performance [15–18], there is a lack of work exploring the effect of these methods on feature ranking. This issue is particularly crucial in clinical contexts, where the interpretability of decisions and therefore the accurate identification of the feature contribution to the decisions of the algorithms is a key requirement [19,20].

In this work, we developed a ML framework to explore the effects of unbalanced populations on the performance and stability of feature ranking for sentinel lymph node status classification in breast cancer. In particular, we assessed the predictive power of different clinical and immunohistochemical features by exploiting two classifiers. We embedded a module specifically devoted to analysis the importance of features in relation to the variation of the training set obtained by different sampling techniques. Our work aims to address some questions: (i) How do the different models perform? (ii) What are the most important predictors? (iii) Are they stable across populations?

2. Materials

2.1. Data

From 2015 to 2017, we collected 635 patients enrolled at Istituto Tumori “Giovanni Paolo II” in Bari (Italy) according to the following eligibility criteria: (i) no evidence of metastatic lymph nodes on palpation or radiological examination, and (ii) patient undergoing SBLN. In particular, the one-step nucleic acid amplification (OSNA) procedure is performed in our Institute. OSNA is the intra-operative exam with a sensitivity and specificity of 87.5–100% and 90.5–100%, respectively [21], but it is an expensive and time-consuming process. Overall, 214 patients had clinically positive lymph nodes, whereas 421 patients resulted negative. For each patient, we collected age at breast cancer diagnosis and several prognostic factors related to the tumor evaluated on pre-operative stage. The retrospective observational study was approved by the Scientific Board of the Istituto Tumori “Giovanni Paolo II” and carried out according the Helsinki Statement. All patients who agreed to have their data used for research were enrolled.

2.2. Histological Features

We gathered information from our pathological anatomy department’s immunohistochemistry analyses, such as: tumor size stage (T: staging system classify), histological grade (G, Elston–Ellis scale: 1, 2, 3), estrogen receptor expression (ER, Pos/Neg), histological subtype (i.e., ductal, lobular, other special types), progesterone receptor expression (PgR, Pos/Neg), cellular marker for proliferation (Ki67, Pos/Neg with cut-off 20%), tumor

multiplicity (Pos/Neg), human epidermal growth factor receptor-2 (HER2/neu: 0, 1+, 2+, 3+), presence of carcinoma in situ associated with invasive component (Pos/Neg), and the sentinel lymph nodes status (N, Pos/Neg) required in the classification approach. A lower grade denotes a better prognosis, and the Elston–Ellis adaptation of the Scarff–Bloom–Richardson grading system uses a three-grade scale to describe the tumor grade G: G1 (low grade), G2 (intermediate grade), or G3 (high grade) [22]. The histological examination was carried out using multiple 14–16 G core biopsy sampling while being guided by ultrasound. The characteristics of the samples are summarized in Table 1.

Table 1. Characteristics of the samples collected in this study.

	N. Patients	N. Positive/ N. Negative
Overall	635	214/421
Histologic Type		
<i>Ductal</i>	512	185/327
<i>Lobular</i>	67	20/47
<i>Special type</i>	56	9/47
Diameter		
<i>T1a</i>	31	3/28
<i>T1b</i>	125	18/107
<i>T1c</i>	281	88/193
<i>T2</i>	198	105/93
ER		
<i>Positive</i>	571	193/378
<i>Negative</i>	64	21/43
Grading		
<i>G1</i>	175	35/140
<i>G2</i>	287	111/176
<i>G3</i>	173	68/105
HER2		
<i>0</i>	471	161/310
<i>1</i>	78	23/55
<i>2</i>	46	21/25
<i>3</i>	39	9/30
Multifocality		
<i>Positive</i>	143	61/82
<i>Negative</i>	492	153/339
In situ component		
<i>Positive</i>	369	114/255
<i>Negative</i>	266	100/166

3. Methods

In this work, we investigated the effectiveness of two classification methods (Random Forest and Logit Lasso) with three training strategies, i.e., unbalanced strategy, oversampling of the minority class and subsampling of the majority class. As depicted in Figure 1, we adopted a hold-out approach, randomly selecting $M = 100$ samples for the independent test (i.e., the hold-out test). A repeated k-fold validation ($N = 10$ iterations, $k = 10$ folds) was selected as cross-validation approach on the training set. In detail, each of the three strategies was combined with each of the two classification algorithms resulting in six training schemes. For each scheme, we evaluated both the importance and variability of the features. In addition, the consensus degree on the probability scores of the different strategies was assessed as the correlation between each couple of schemes. Each step is detailed in the following sections. We used a PC with the following hardware configuration: Intel Core i7-8550U CPU @ 1.99GHz × 4, and 16GB RAM to run the experiments. All the analyses were performed by using R Statistical Software (v4.1.1., R Core Team 2021) with a run-time of 9.8 min for the execution of the entire framework.

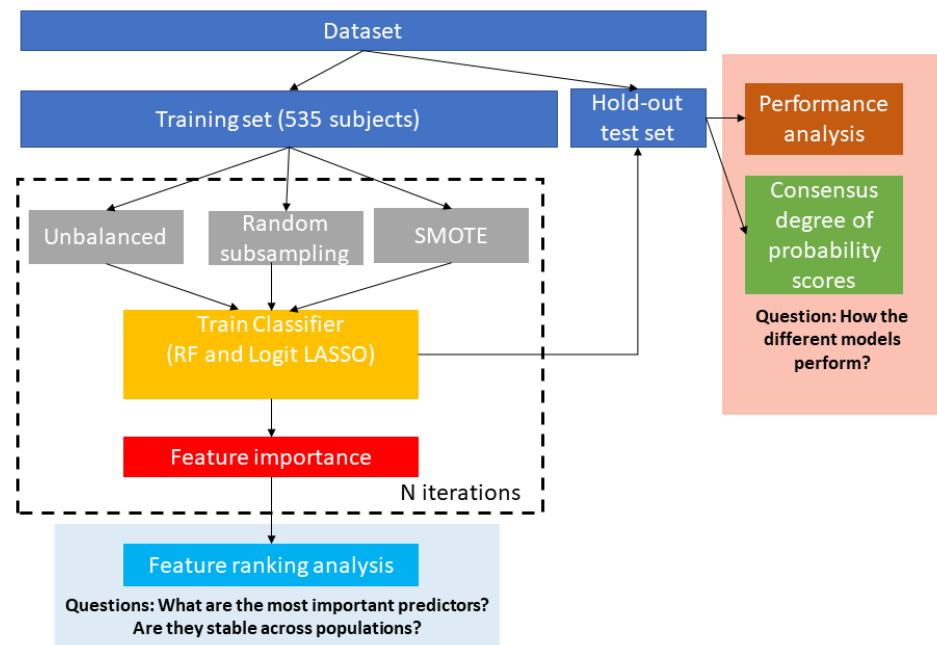


Figure 1. Overview of the proposed framework.

3.1. Training Strategies

The simple subsampling strategy involves randomly eliminating samples from the majority class in order to match a certain fixed percentage of samples from the minority class. In this work, we selected the parameter of 50% of the samples in the minority class to achieve a reasonable trade-off between sample representativeness and computational efficiency.

On the other hand, SMOTE (Synthetic Minority Over-sampling Technique) creates artificial samples from the minority class leveraging the information in the data [23]. For each sample from the minority class x_i , firstly $R = 5$ samples from the minority class with the smallest Euclidean distance from the original sample were identified (i.e., the top R nearest neighbors $x_j^{NN}, j = 1, \dots, R$), then, one of these samples is randomly chosen (x_s^{NN}) and a new synthetic SMOTE sample is defined as:

$$x^{SMOTE} = x + u \cdot (x_s^{NN} - x), \tag{1}$$

where u is randomly chosen from the uniform distribution $U(0, 1)$ and is the same for all variables, but differs for each SMOTE sample in order to ensure that the SMOTE sample is on the line joining the two original samples used to generate it [24]. Thus, SMOTE is an augmentation technique to increase the minority class resulting in an output balanced training set.

3.2. Classification Algorithms

3.2.1. Logistic LASSO

In binary classification studies, the dichotomous response variable y_i is usually coded as 1 for cases and 0 for controls. In order to model logistic regression, the probability $p_i = Pr(y_i = 1)$ of case i given the predictor vector x_i can be expressed as:

$$p_i = \frac{e^{\beta_0 + x_i^T \beta}}{1 + e^{\beta_0 + x_i^T \beta}}, \tag{2}$$

The parameter vector $\Theta = (\beta_0, \beta_1, \dots, \beta_p)^T$ is usually estimated by maximizing the log-likelihood function:

$$L(\Theta) = \sum_{i=1}^N [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (3)$$

Regularization methods such as shrinkage or penalized regression could reduce the likelihood of overfitting by modifying the loss function with a penalty term to shrink the coefficients in the regression towards zero and selecting the nonzero variables in the final model [25]. This approach has the additional advantage of selecting the most important variables for the predictions from a large and potentially multicollinear set of features, resulting in a more relevant and interpretable set of predictors [26–28]. The logistic LASSO model involves L1 penalty term, resulting in:

$$L(\Theta) + \lambda \sum_{j=1}^P |\beta_j| \quad (4)$$

In order to tune the penalty constant λ , without introducing bias and overfitting, we used an inner round of k-fold validation within each training round, with $k = 3$. The AUC metric was computed for the test set as the performance measure to tune λ . We used the “glmnet” package to fit the logistic LASSO regression.

3.2.2. Random Forest

Random Forest (RF) is also recognized as bagged decision trees. This algorithm works on using different weak learners to implement strong learners [29]. The target outcome for each sample y_i is individually forecast by each tree, while the final predictions are based on the majority of trees voting.

Two types of randomization are included: (I) a subset of observations is picked at random for each tree, and (II) a random set of $mtry$ candidate predictors is chosen to produce a split within each tree. As a result, a purity measure and a decision threshold are used to divide the node input samples into two groups. Each tree is built until the nodes have divided their inputs into subsets and assigned a single final label to each of them. The out-of-the-bag (OOB) set for that tree contains the samples that were not utilized for that tree [30]. The accuracy of RF is assessed by using the samples of the OOB as:

$$MSE_{OOB} = \frac{1}{N} \sum_{i=1}^N (y_i - \tilde{y}_{iO})^2, \quad (5)$$

where \tilde{y}_{iO} denotes the average prediction for the i th observation from all trees for which this observation has been OOB. We used the “RandomForest” R Package with the default parameter $mtry = P/3$ and number of trees = 100.

3.3. Performance Evaluation

The model decisions that arise may be categorized into four different groups: true positives (TP), which occur when the model correctly predicts the positive class, true negatives (TN), which occur when the model correctly predicts the negative class, and false positives (FP) and false negatives (FN), which occur when the model incorrectly predicts the positive and negative classes, respectively. Given these four cases, we considered the following metrics to evaluate the performance of the classification models:

- Accuracy

$$\frac{TP + TN}{TP + TN + FP + FN}$$

- Sensitivity

$$\frac{TP}{TP + FN}'$$

- Specificity

$$\frac{TN}{TN + FP}'$$

- Area Under the Receiver Operating Characteristics (ROC) Curve (AUC) that adjusts the decision threshold to plot sensitivity versus specificity.

3.4. Feature Ranking Analysis

Both RF and Lasso Logistic regression provide embedded feature selection techniques. As a result of the model, a subset of the features with non-zero weights could be retrieved, and the coefficients of the least relevant features are shrunk to zero [31]. Moreover, the absolute values of the Lasso coefficients could be used for feature ranking since the weight β_i quantifies the impact that each feature has in the regression model. We exploited the averaged absolute values of the weights across the validation rounds in order to obtain a final feature ranking, regardless of the specific training fold.

On the other side, RF feature importance can be assessed by applying the permutation-based MSE reduction criterion [32]. The relevance of each feature per each tree is determined by permuting the feature's OOB data for the tree and calculating the difference between the permuted and true OOB-MSE. By averaging these differences over all of the forest's trees, the final MSE decrease for each feature is produced. The basic logic behind this technique is that if a feature has no effect on performance, the difference in accuracy estimated using the true values of the feature and that computed using its permuted values is unlikely to be significant.

For each classification scheme (i.e., each sampling algorithm with each of the two machine learning methods), we averaged the importance scores of the features across the rounds to obtain a single feature importance vector. In addition, the quartile coefficient of dispersion [33] was used to assess the variability across rounds of each feature.

4. Results and Discussion

4.1. Performance

Figures 2 and 3 show the performance metrics across the cross-validation rounds and for the independent hold-out test for each classification scheme, respectively. Tables 2 and 3 list the mean and standard deviation for each performance metric for cross-validation sets and hold-out test, respectively. The Logit Lasso algorithm is less affected by model overfitting as it can be noted from the smaller difference between training and test performance. Moreover, it is worth noting that among the various strategies, the SMOTE technique provides the best balance between sensitivity and specificity.

Overall, the achieved performance compares favorably with that obtained in our previous work [34]. In particular, in this analysis, we also found the logistic regression model as the method with the most stable performance between the training and test sets. However, in contrast to our previous work, here we used different training sampling strategies that showed that SMOTE represents a promising method for this clinical challenge and that simply undersampling does not guarantee equally effective results.

Other works addressed similar classification tasks, showing that the performance can be significantly improved by adding predictors extracted from imaging [35,36], genetics [9] and nomograms of clinical and pathologic variables [37] with more complex nonlinear models such as deep neural networks [38]. In our analysis, we used simpler and therefore more interpretable models in combination with known sampling strategies in order to highlight different effects on performance. In future developments, we will expand the presented framework to investigate the impact of other features and predictive models.

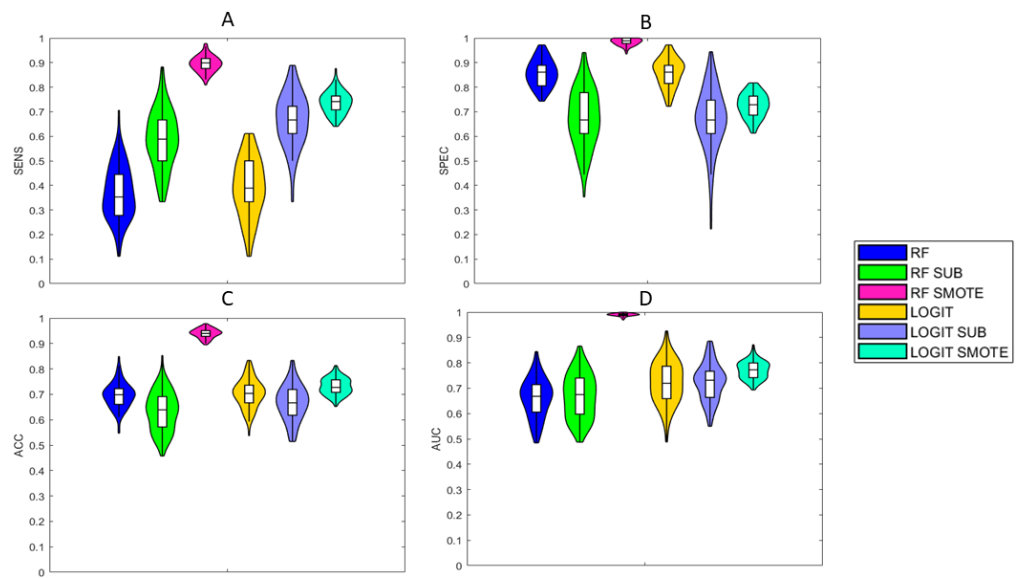


Figure 2. Performance metrics across the cross-validation rounds for each classification scheme: violin plots of the distributions of the sensitivity values (A), specificity values (B), accuracy values (C), and AUC values (D).

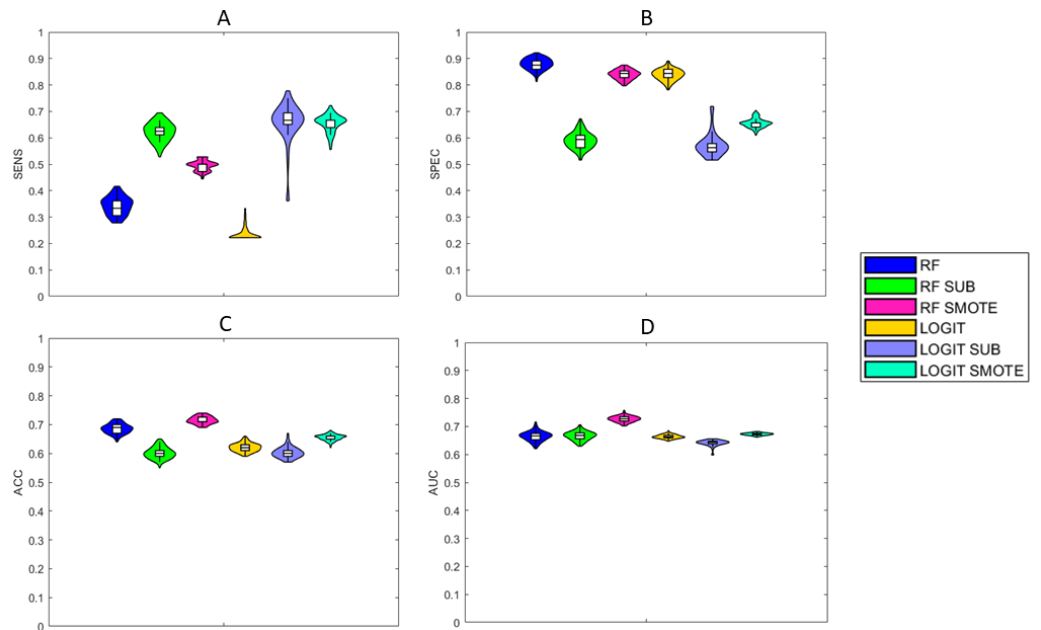


Figure 3. Performance metrics for the hold-out test for each classification scheme: violin plots of the distributions of the sensitivity values (A), specificity values (B), accuracy values (C), and AUC values (D).

Table 2. Performance of the ML models for for the cross-validation sets (mean ± std).

Classification Scheme	SENS	SPEC	ACC	AUC
RF	0.37 ± 0.1	0.85 ± 0.05	0.7 ± 0.05	0.66 ± 0.07
RF SUB	0.59 ± 0.11	0.67 ± 0.11	0.63 ± 0.07	0.66 ± 0.08
RF SMOTE	0.89 ± 0.03	0.98 ± 0.01	0.93 ± 0.01	0.98 ± 0.005
Logit Lasso	0.39 ± 0.1	0.85 ± 0.05	0.70 ± 0.05	0.72 ± 0.08
Logit Lasso SUB	0.67 ± 0.1	0.67 ± 0.12	0.67 ± 0.07	0.72 ± 0.07
Logit Lasso SMOTE	0.73 ± 0.04	0.72 ± 0.04	0.73 ± 0.03	0.77 ± 0.03

Table 3. Performance of the ML models for for the hold-out set (mean ± std).

Classification Scheme	SENS	SPEC	ACC	AUC
RF	0.34 ± 0.03	0.87 ± 0.02	0.69 ± 0.01	0.66 ± 0.01
RF SUB	0.62 ± 0.003	0.58 ± 0.03	0.60 ± 0.01	0.66 ± 0.01
RF SMOTE	0.49 ± 0.03	0.83 ± 0.01	0.71 ± 0.01	0.73 ± 0.01
Logit Lasso	0.23 ± 0.02	0.84 ± 0.02	0.62 ± 0.01	0.66 ± 0.006
Logit Lasso SUB	0.65 ± 0.08	0.57 ± 0.04	0.60 ± 0.01	0.64 ± 0.01
Logit Lasso SMOTE	0.65 ± 0.03	0.65 ± 0.01	0.65 ± 0.01	0.67 ± 0.004

4.2. Feature Ranking Analysis

We computed the feature ranking list resulting from each classification scheme over the cross-validation rounds for the clinical interpretability of the results. Figure 4A shows the feature ranking for the six schemes. It can be noted that only the diameter results are relevant, regardless of the adopted scheme. The correlation between tumor size and lymph node status has been evaluated in several studies. In patients with breast cancer, increasing tumor diameter at diagnosis is associated with an increasing number of metastatic lymph nodes [39,40].

Moreover, it is worth noting that the three Lasso Logit models agree much more with each other on the final average feature ranking, outlining the features histology, diameter, ER and multifocality as the most important features. As a matter of fact, Logit Lasso exhibits greater stability of feature ranking compared to RF. Stability refers to the feature ranking’s sensitivity to perturbation of training samples and it is associated with the reproducibility power of the feature selection method [41]. When analyzing feature selection algorithms in different clinical scenarios, high stability could be just as crucial as high classification accuracy [42–44]. We used the quartile coefficient to objectively measure the stability of these features for each classification scheme. It can be seen in Figure 4B that the feature diameter is the most stable among all the features, regardless of the adopted scheme, while the HER2 feature exhibits the highest dispersion coefficient, showing high instability. Moreover, the weights of Age, PgR and Ki67 are shrunk to zero from the Lasso algorithm, highlighting their low impact on the classification task.

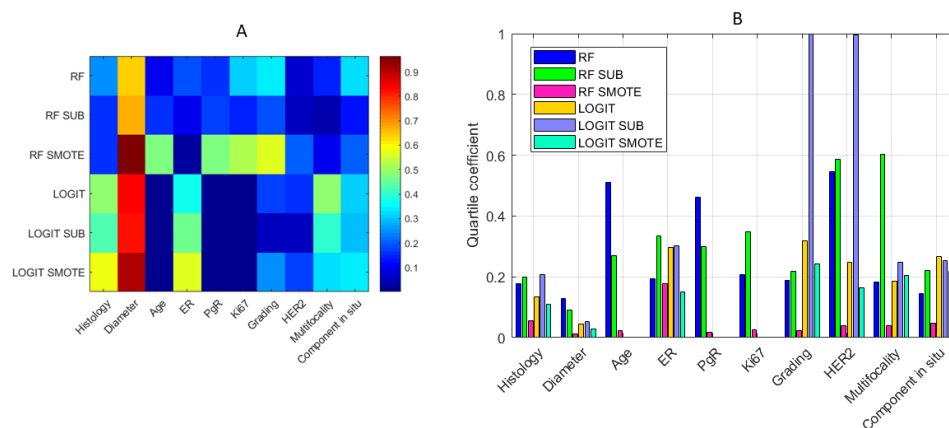


Figure 4. (A) Importance scores of the features averaged across the rounds to obtain a single feature importance vector for each classification scheme. (B) Variability across rounds of each feature for each classification scheme expressed as the quartile coefficient of dispersion.

4.3. Consensus Degree of the Probability Scores

We better analyzed the behavior of the classification schemes on the independent hold-out test set, by inspecting the correlation between the probability scores resulting from each couple of schemes as shown in Figure 5. It is important to underline that the RF algorithm exhibits the lowest correlations between the probability scores for the different training strategies, showing an inherent instability with respect to the sampling methods;

on the contrary, the Logit Lasso algorithm shows the highest correlations between the probability scores, underlining better stability of performance and less variability among the sampling strategies. This issue is of paramount importance in the clinical setting, where stability and generalization of algorithms supporting diagnosis are highly recommended. These two aspects are closely linked to each other since we need to evaluate how changes in the composition of the learning set (i.e., sampling randomness) influence the function produced by the algorithm (i.e., the probability scores) [45–47].

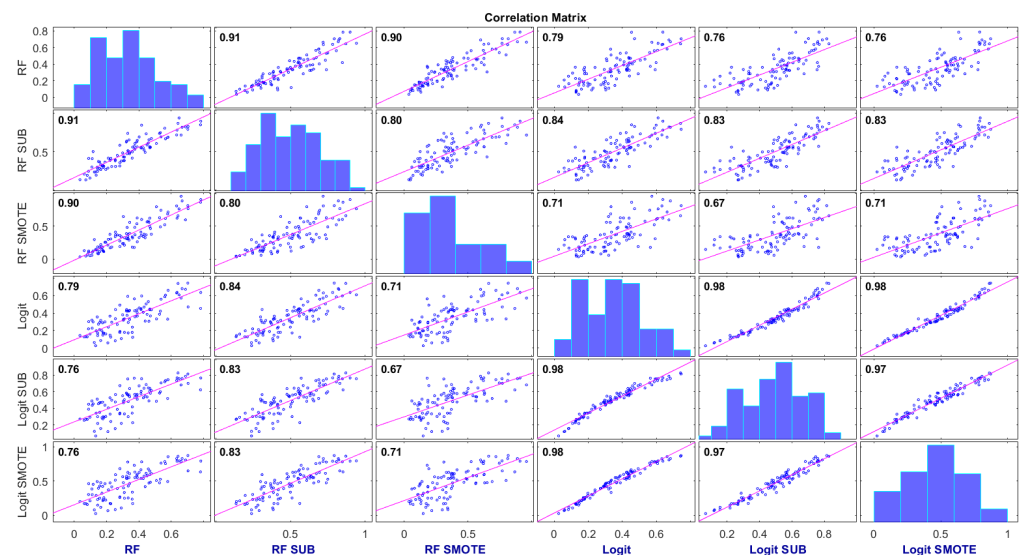


Figure 5. Correlation between the probability scores of the samples in the hold-out test set resulting from each classification scheme.

5. Landscape

Most of the studies proposed in the literature aimed at predicting the lymph node status in general terms (i.e., non-sentinel lymph nodes) using nomograms of clinical and pathological variables [6,7,48,49]. On the contrary, to the best of our knowledge, few works aimed at predicting the status of the sentinel lymph node [9,50]. The model proposed in [9] reached an average AUC value of 88.3%. by using genetic characteristics, tumor size and lymph vascular invasion. In [50], the authors proposed an analogous model for patients with ductal carcinoma in situ by stating an AUC value of 75%. In our preliminary works, comparable results were reached to the-state-of-the-art [10,34]. Recent studies, radiomic features extracted from magnetic resonance images [35,36,51] or ultrasounds [52] were exploited to identify a better performing customized model. However, none of the works reported in the literature specifically addressed the issue of class imbalance, as proposed in this work.

6. Limitations and Future Perspectives

In this work, we analyzed a dataset of limited size, therefore we avoided more sophisticated strategies such as those based on deep neural networks that require large sample sizes, preferring two simpler methods such as RF and Lasso. Such methods are rather well established in the literature for classification tasks in various clinical contexts. Moreover, these methods offer the advantage of providing the most important features for the algorithm's decisions proving a higher explainability of the impact of the clinical predictors on the overall performance. This issue is particularly relevant as it enables the integration of the perspectives of both clinical specialists and algorithm developers, allowing synergy between the different entities and improving the machine learning paradigms in each specific clinical context.

Although our methodological choices were conditioned by the dimensionality of the dataset, it should be emphasised that for high-dimensional datasets, other approaches should be evaluated in conjunction with Lasso. As an example, Ren et al. [53] proposed a network constrained regularization approach for classification exploiting networks to model interconnections among features, overcoming the existing shortcomings in addressing the correlations among features. In addition, when the number of features increases, robust feature selection techniques are critical for successful performance since these methods usually have better stability and reproducibility [54]. Accordingly, with higher dimensionality of the features and complex data structure, more feature selection algorithms should be evaluated in order to avoid the oversimplification of the predictive models. On the other hand, although the proposed framework is fully scalable and would allow a larger number of machine learning algorithms combined with different sampling strategies to be compared for future developments, it should be noted that the computational complexity and the execution time increase with the number of algorithms to be compared. Further analyses will be carried out to identify the best combinations of ML algorithms and sampling strategies that ensure the best balance among classification accuracy, computational complexity and interpretability of the models.

7. Conclusions

In this work, we presented a simple ML framework to explore the effect of different sampling strategies on both performance and feature stability for the prediction of the sentinel lymph node status in breast cancer. Due to the unbalanced configuration of the classification problem, we trained two different machine learning algorithms on tumor histopathology and clinical features in combination with the random undersampling method and a synthetic augmentation technique (SMOTE) to balance the training set. We found that the best combination of techniques consists in the use of the simpler Lasso classifier and the SMOTE strategy both in terms of the ability to generalize the predictive models on the hold-out test set and the stability of the clinical features. Although the proposed framework does not yet allow the implementation of a diagnosis support system due to moderate performance, it provides a set of clinical features worthy of attention, which should be considered as possible biomarkers alongside other imaging and genetic features.

Author Contributions: Conceptualization, A.L., S.T., R.M. and A.F.; methodology, A.L.; software, A.L.; methodological validation, R.M. and A.F.; clinical validation, D.L.F. and V.L.; formal analysis, A.L.; investigation, A.L.; resources, R.B. and F.A.Z.; data curation, A.F. and R.M.; writing—original draft preparation, A.L.; writing—review and editing, A.L., N.A., L.B., A.F., D.L.F., V.L., A.M., S.T., S.B., M.C.C., R.B.; visualization, A.L.; supervision, R.M., R.B. and S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by funding from Ricerca Finalizzata 2018.

Institutional Review Board Statement: The study received approval from the Scientific Board of Istituto Tumori “Giovanni Paolo II”—Bari, Italy and was carried out in accordance with the Declaration of Helsinki’s standards. The authors affiliated to the Istituto Tumori “Giovanni Paolo II” RCCS, Bari are responsible for the views expressed in this article, which do not necessarily represent the ones of the Institute.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available upon reasonable request to the corresponding author. The data are not publicly available because they are the property of Istituto Tumori “Giovanni Paolo II”—Bari, Italy.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mazo, C.; Kearns, C.; Mooney, C.; Gallagher, W.M. Clinical decision support systems in breast cancer: A systematic review. *Cancers* **2020**, *12*, 369. [[CrossRef](#)] [[PubMed](#)]
2. Yan, M.; Abdi, M.A.; Falkson, C. Axillary management in breast cancer patients: A comprehensive review of the key trials. *Clin. Breast Cancer* **2018**, *18*, e1251–e1259. [[CrossRef](#)] [[PubMed](#)]
3. Cormier, J.N.; Askew, R.L.; Mungovan, K.S.; Xing, Y.; Ross, M.I.; Armer, J.M. Lymphedema beyond breast cancer: A systematic review and meta-analysis of cancer-related secondary lymphedema. *Cancer* **2010**, *116*, 5138–5149. [[CrossRef](#)] [[PubMed](#)]
4. Giuliano, A.E.; Ballman, K.V.; McCall, L.; Beitsch, P.D.; Brennan, M.B.; Kelemen, P.R.; Ollila, D.W.; Hansen, N.M.; Whitworth, P.W.; Blumencranz, P.W.; et al. Effect of axillary dissection vs no axillary dissection on 10-year overall survival among women with invasive breast cancer and sentinel node metastasis: The ACOSOG Z0011 (Alliance) randomized clinical trial. *JAMA* **2017**, *318*, 918–926. [[CrossRef](#)]
5. Galimberti, V.; Fontana, S.R.; Maisonneuve, P.; Steccanella, F.; Vento, A.; Intra, M.; Naninato, P.; Caldarella, P.; Iorfida, M.; Colleoni, M.; et al. Sentinel node biopsy after neoadjuvant treatment in breast cancer: Five-year follow-up of patients with clinically node-negative or node-positive disease before treatment. *Eur. J. Surg. Oncol. (EJSO)* **2016**, *42*, 361–368. [[CrossRef](#)]
6. Chen, K.; Liu, J.; Li, S.; Jacobs, L. Development of nomograms to predict axillary lymph node status in breast cancer patients. *BMC Cancer* **2017**, *17*, 1–10. [[CrossRef](#)]
7. Houvenaeghel, G.; Lambaudie, E.; Classe, J.M.; Mazouni, C.; Giard, S.; Cohen, M.; Faure, C.; Charitansky, H.; Rouzier, R.; Daraï, E.; et al. Lymph node positivity in different early breast carcinoma phenotypes: A predictive model. *BMC Cancer* **2019**, *19*, 1–10. [[CrossRef](#)]
8. Chen, J.Y.; Chen, J.J.; Yang, B.L.; Liu, Z.B.; Huang, X.Y.; Liu, G.Y.; Han, Q.X.; Yang, W.T.; Shen, Z.Z.; Shao, Z.M.; et al. Predicting sentinel lymph node metastasis in a Chinese breast cancer population: Assessment of an existing nomogram and a new predictive nomogram. *Breast Cancer Res. Treat.* **2012**, *135*, 839–848. [[CrossRef](#)]
9. Okuno, J.; Miyake, T.; Sota, Y.; Tanei, T.; Kagara, N.; Naoi, Y.; Shimoda, M.; Shimazu, K.; Kim, S.J.; Noguchi, S. Development of prediction model including microRNA expression for sentinel lymph node metastasis in ER-positive and HER2-negative breast cancer. *Ann. Surg. Oncol.* **2021**, *28*, 310–319. [[CrossRef](#)]
10. Fanizzi, A.; Pomarico, D.; Paradiso, A.; Bove, S.; Diotaiuti, S.; Didonna, V.; Giotta, F.; La Forgia, D.; Latorre, A.; Pastena, M.I.; et al. Predicting of sentinel lymph node status in breast cancer patients with clinically negative nodes: A Validation Study. *Cancers* **2021**, *13*, 352. [[CrossRef](#)]
11. Estabrooks, A.; Jo, T.; Japkowicz, N. A multiple resampling method for learning from imbalanced data sets. *Comput. Intell.* **2004**, *20*, 18–36. [[CrossRef](#)]
12. Weiss, G.M. Mining with rarity: A unifying framework. *ACM Sigkdd Explor. Newsl.* **2004**, *6*, 7–19. [[CrossRef](#)]
13. Tang, Y.; Zhang, Y.Q.; Chawla, N.V.; Krasser, S. SVMs modeling for highly imbalanced classification. *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)* **2008**, *39*, 281–288. [[CrossRef](#)] [[PubMed](#)]
14. Jo, T.; Japkowicz, N. Class imbalances versus small disjuncts. *ACM Sigkdd Explor. Newsl.* **2004**, *6*, 40–49. [[CrossRef](#)]
15. Zhao, Y.; Wong, Z.S.Y.; Tsui, K.L. A framework of rebalancing imbalanced healthcare data for rare events' classification: A case of look-alike sound-alike mix-up incident detection. *J. Healthc. Eng.* **2018**, *2018*, 6275435. [[CrossRef](#)] [[PubMed](#)]
16. Muhamed Ali, A.; Zhuang, H.; Ibrahim, A.; Rehman, O.; Huang, M.; Wu, A. A machine learning approach for the classification of kidney cancer subtypes using mirna genome data. *Appl. Sci.* **2018**, *8*, 2422. [[CrossRef](#)]
17. Jeong, B.; Cho, H.; Kim, J.; Kwon, S.K.; Hong, S.; Lee, C.; Kim, T.; Park, M.S.; Hong, S.; Heo, T.Y. Comparison between statistical models and machine learning methods on classification for highly imbalanced multiclass kidney data. *Diagnostics* **2020**, *10*, 415. [[CrossRef](#)]
18. Barbieri, D.; Chawla, N.; Zaccagni, L.; Grgurinović, T.; Šarac, J.; Čoklo, M.; Missoni, S. Predicting cardiovascular risk in Athletes: Resampling improves classification performance. *Int. J. Environ. Res. Public Health* **2020**, *17*, 7923. [[CrossRef](#)]
19. Tjoa, E.; Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4793–4813. [[CrossRef](#)]
20. Lombardi, A.; Diacono, D.; Amoroso, N.; Monaco, A.; Tavares, J.M.R.; Bellotti, R.; Tangaro, S. Explainable Deep Learning for Personalized Age Prediction With Brain Morphology. *Front. Neurosci.* **2021**, *15*, 578. [[CrossRef](#)]
21. Szychta, P.; Westfal, B.; Maciejczyk, R.; Smolarz, B.; Romanowicz, H.; Krawczyk, T.; Zadrożny, M. Intraoperative diagnosis of sentinel lymph node metastases in breast cancer treatment with one-step nucleic acid amplification assay (OSNA). *Arch. Med. Sci. AMS* **2016**, *12*, 1239. [[CrossRef](#)] [[PubMed](#)]
22. Egner, J.R. AJCC cancer staging manual. *JAMA* **2010**, *304*, 1726–1727. [[CrossRef](#)]
23. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [[CrossRef](#)]
24. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284.
25. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. [[CrossRef](#)]
26. Wu, T.T.; Chen, Y.F.; Hastie, T.; Sobel, E.; Lange, K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **2009**, *25*, 714–721. [[CrossRef](#)] [[PubMed](#)]
27. Kim, S.M.; Kim, Y.; Jeong, K.; Jeong, H.; Kim, J. Logistic LASSO regression for the diagnosis of breast cancer using clinical demographic data and the BI-RADS lexicon for ultrasonography. *Ultrasonography* **2018**, *37*, 36. [[CrossRef](#)]

28. McEligot, A.J.; Poynor, V.; Sharma, R.; Panangadan, A. Logistic LASSO regression for dietary intakes and breast cancer. *Nutrients* **2020**, *12*, 2652. [[CrossRef](#)]
29. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
30. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
31. Yamada, M.; Jitkrittum, W.; Sigal, L.; Xing, E.P.; Sugiyama, M. High-dimensional feature selection by feature-wise kernelized lasso. *Neural Comput.* **2014**, *26*, 185–207. [[CrossRef](#)] [[PubMed](#)]
32. Grömping, U. Variable importance assessment in regression: Linear regression versus random forest. *Am. Stat.* **2009**, *63*, 308–319. [[CrossRef](#)]
33. Bonett, D.G. Confidence interval for a coefficient of quartile variation. *Comput. Stat. Data Anal.* **2006**, *50*, 2953–2957. [[CrossRef](#)]
34. Fanizzi, A.; Lorusso, V.; Biafora, A.; Bove, S.; Comes, M.C.; Cristofaro, C.; Digennaro, M.; Didonna, V.; Forgia, D.L.; Nardone, A.; et al. Sentinel Lymph Node Metastasis on Clinically Negative Patients: Preliminary Results of a Machine Learning Model Based on Histopathological Features. *Appl. Sci.* **2021**, *11*, 10372. [[CrossRef](#)]
35. Dong, Y.; Feng, Q.; Yang, W.; Lu, Z.; Deng, C.; Zhang, L.; Lian, Z.; Liu, J.; Luo, X.; Pei, S.; et al. Preoperative prediction of sentinel lymph node metastasis in breast cancer based on radiomics of T2-weighted fat-suppression and diffusion-weighted MRI. *Eur. Radiol.* **2018**, *28*, 582–591. [[CrossRef](#)] [[PubMed](#)]
36. Liu, M.; Mao, N.; Ma, H.; Dong, J.; Zhang, K.; Che, K.; Duan, S.; Zhang, X.; Shi, Y.; Xie, H. Pharmacokinetic parameters and radiomics model based on dynamic contrast enhanced MRI for the preoperative prediction of sentinel lymph node metastasis in breast cancer. *Cancer Imaging* **2020**, *20*, 1–8. [[CrossRef](#)]
37. Klar, M.; Foeldi, M.; Markert, S.; Gitsch, G.; Stickeler, E.; Watermann, D. Good prediction of the likelihood for sentinel lymph node metastasis by using the MSKCC nomogram in a German breast cancer population. *Ann. Surg. Oncol.* **2009**, *16*, 1136–1142. [[CrossRef](#)]
38. Luo, J.; Ning, Z.; Zhang, S.; Feng, Q.; Zhang, Y. Bag of deep features for preoperative prediction of sentinel lymph node metastasis in breast cancer. *Phys. Med. Biol.* **2018**, *63*, 245014. [[CrossRef](#)]
39. Sopik, V.; Narod, S.A. The relationship between tumour size, nodal status and distant metastases: On the origins of breast cancer. *Breast Cancer Res. Treat.* **2018**, *170*, 647–656. [[CrossRef](#)]
40. Min, S.K.; Lee, S.K.; Woo, J.; Jung, S.M.; Ryu, J.M.; Yu, J.; Lee, J.E.; Kim, S.W.; Chae, B.J.; Nam, S.J. Relation between tumor size and lymph node metastasis according to subtypes of breast cancer. *J. Breast Cancer* **2021**, *24*, 75. [[CrossRef](#)]
41. Lombardi, A.; Amoroso, N.; Diacono, D.; Monaco, A.; Logroscino, G.; De Blasi, R.; Bellotti, R.; Tangaro, S. Association between structural connectivity and generalized cognitive spectrum in Alzheimer’s disease. *Brain Sci.* **2020**, *10*, 879. [[CrossRef](#)] [[PubMed](#)]
42. Awada, W.; Khoshgoftaar, T.M.; Dittman, D.; Wald, R.; Napolitano, A. A review of the stability of feature selection techniques for bioinformatics data. In Proceedings of the 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI), Las Vegas, NV, USA, 8–10 August 2012; pp. 356–363.
43. Nogueira, S.; Sechidis, K.; Brown, G. On the stability of feature selection algorithms. *J. Mach. Learn. Res.* **2017**, *18*, 6345–6398.
44. Khaire, U.M.; Dhanalakshmi, R. Stability of feature selection algorithm: A review. *J. King Saud-Univ.-Comput. Inf. Sci.* **2019**, *34*, 1060–1073. [[CrossRef](#)]
45. Bousquet, O.; Elisseeff, A. Stability and generalization. *J. Mach. Learn. Res.* **2002**, *2*, 499–526.
46. Kernbach, J.M.; Staartjes, V.E. Foundations of Machine Learning-Based Clinical Prediction Modeling: Part II—Generalization and Overfitting. In *Machine Learning in Clinical Neuroscience*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 15–21.
47. Futoma, J.; Simons, M.; Doshi-Velez, F.; Kamaleswaran, R. Generalization in clinical prediction models: The blessing and curse of measurement indicator variables. *Crit. Care Explor.* **2021**, *3*, e0453. [[CrossRef](#)] [[PubMed](#)]
48. Bonsang-Kitzis, H.; Mouttet-Boizat, D.; Guillot, E.; Feron, J.G.; Fourchette, V.; Alran, S.; Pierga, J.Y.; Cottu, P.; Lerebours, F.; Stevens, D.; et al. Medico-economic impact of MSKCC non-sentinel node prediction nomogram for ER-positive HER2-negative breast cancers. *PLoS ONE* **2017**, *12*, e0169962. [[CrossRef](#)] [[PubMed](#)]
49. Ahn, S.K.; Kim, M.K.; Kim, J.; Lee, E.; Yoo, T.K.; Lee, H.B.; Kang, Y.J.; Kim, J.; Moon, H.G.; Chang, J.M.; et al. Can we skip intraoperative evaluation of sentinel lymph nodes? Nomogram predicting involvement of three or more axillary lymph nodes before breast cancer surgery. *Cancer Res. Treat. Off. J. Korean Cancer Assoc.* **2017**, *49*, 1088–1096. [[CrossRef](#)]
50. Bevilacqua, J.L.B.; Kattan, M.W.; Fey, J.V.; Cody III, H.S.; Borgen, P.I.; Van Zee, K.J. Doctor, what are my chances of having a positive sentinel node? A validated nomogram for risk estimation. *J. Clin. Oncol.* **2007**, *25*, 3670–3679. [[CrossRef](#)]
51. Liu, J.; Sun, D.; Chen, L.; Fang, Z.; Song, W.; Guo, D.; Ni, T.; Liu, C.; Feng, L.; Xia, Y.; et al. Radiomics analysis of dynamic contrast-enhanced magnetic resonance imaging for the prediction of sentinel lymph node metastasis in breast cancer. *Front. Oncol.* **2019**, *9*, 980. [[CrossRef](#)]
52. Bove, S.; Comes, M.C.; Lorusso, V.; Cristofaro, C.; Didonna, V.; Gatta, G.; Giotta, F.; La Forgia, D.; Latorre, A.; Pastena, M.I.; et al. A ultrasound-based radiomic approach to predict the nodal status in clinically negative breast cancer patients. *Sci. Rep.* **2022**, *12*, 1–10. [[CrossRef](#)]
53. Ren, J.; He, T.; Li, Y.; Liu, S.; Du, Y.; Jiang, Y.; Wu, C. Network-based regularization for high dimensional SNP data in the case-control study of Type 2 diabetes. *BMC Genet.* **2017**, *18*, 1–12. [[CrossRef](#)] [[PubMed](#)]
54. Wu, C.; Ma, S. A selective review of robust variable selection with applications in bioinformatics. *Brief. Bioinform.* **2015**, *16*, 873–883. [[CrossRef](#)] [[PubMed](#)]