Cancer Control
Volume 26: 1-8
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1073274819876598
journals.sagepub.com/home/ccx
SAGE

Decision Support for Breast Cancer Detection: Classification Improvement Through Feature Selection

Flavio S. Fogliatto, PhD[1], Michel J. Anzanello, PhD[1], Felipe Soares, MSC[1], and Priscila G. Brust-Renck, PhD[1]

Abstract

Several statistical-based approaches have been developed to support medical personnel in early breast cancer detection. This article presents a method for feature selection aimed at classifying cases into categories based on patients' breast tissue measures and protein microarray. The effectiveness of this feature selection strategy was evaluated against the commonly used Wisconsin Breast Cancer Database—WBCD (with several patients and fewer features) and a new protein microarray data set (with several features and fewer patients). Features were ranked according to a feature importance index that combines parameters emerging from the unsupervised method of principal component analysis and the supervised method of Bhattacharyya distance. Observations of a training set were iteratively categorized into malignant and benign cases through 3 classification techniques: k-Nearest Neighbor, linear discriminant analysis, and probabilistic neural network. After each classification, the feature with the smallest importance index was removed, and a new categorization was carried out until there was only one feature left. The subset yielding maximum accuracy was used to classify observations in the testing set. Our method yielded average 99.17% accurate classifications in the testing set while retaining average 4.61 out of 9 features in the WBCD, which is comparable to the best results reported by the literature on that data set, with the advantage of relying on simple and widely available multivariate techniques. When applied to the microarray data, the method yielded average accuracy of 98.30% while retaining average 2.17% of the original features. Our results can aid health-care professionals during early diagnosis of breast cancer.

Keywords

decision support, breast cancer diagnosis, feature selection, *k*-nearest neighbor, probabilistic neural network, linear discriminant analysis

Received January 08, 2019. Received revised April 18, 2019. Accepted for publication July 31, 2019.

Introduction

Breast cancer (BC) is one of the most reported types of cancer around the world and the third leading cause of death among women, exceeded only by lung cancer and heart diseases.[1] As most types of cancers, BC genetic drivers are not yet completely defined, leading to a lack of effective medical diagnosis strategies.[2] Nevertheless, early diagnosis is the best way to achieve higher survival rates among patients.[3,4]

Different systems to classify samples have been proposed, aimed to aid technical personnel during the processing of data gathered from BC laboratory examinations. Some of the first classifier systems developed were proposed by Street et al[5] and Fogel et al.[6] A classifier can be defined as an expert system capable of reaching a conclusion by analyzing some type of

[1] Industrial Engineering Department, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

Corresponding Author:
Flavio S. Fogliatto, Industrial Engineering Department, Federal University of Rio Grande do Sul, Av. Osvaldo Aranha, 99 - 5o andar, Porto Alegre, RS 90035-190, Brazil.
Email: ffogliatto@producao.ufrgs.br

data. In the case of BC, this system may analyze images or laboratorial examinations and classify the given sample as either malignant or benign. Several rationales may guide the classification, although the majority are based on artificial neural networks (eg, Marcano-Cedeño et al[7]) or fuzzy theory (FT; eg, Abonyi and Szeifert[8]). Classifiers support analysts' decisions leading to more accurate and uniform evaluations, particularly when fatigue takes place.

Two main factors can be pointed out as crucial to determine the efficiency of a classifier system: relevance and accuracy. First, the data presented to the classifier should be relevant for the classification problem being analyzed. In light of that, data should usually be preprocessed to enhance classification accuracy, using techniques such as standardizing or scaling the input features and/or eliminating the irrelevant or redundant information (ie, feature selection [FS]). Second, the classification algorithm should be able to produce reliable and accurate results assessed by different performance indicators, such as sensitivity and specificity. In this study, both relevance and accuracy are discussed.

Several frameworks relying on techniques of different complexity and performance have been proposed in recent years to select the most relevant features of BC data. One of the benchmarking data sets for FS in this field is the Wisconsin Breast Cancer Dataset (WBCD; UCI Repository of Machine Learning Databases). In such data set, 9 features were analyzed on images from fine-needle breast tissue aspirates obtained from 699 individuals, for which a diagnosis was also provided.

For example, Marcano-Cedeño et al[7] proposed a neural network-based classifier that simulates the biological property of metaplasticity on a multilayer perceptron algorithm with backpropagation using 60% of the entries in the WBCD as training set and running 100 experiments with varying network parameters with 100 repetitions each. Metaplasticity was defined as the induction of synaptic changes also depending on prior synaptic activity. Their best classification accuracy was 99.26%. In a similar study,[9] they have achieved an accuracy of 99.63% in the same data set.

Onan[10] proposed a classification model based on 3 phases, integrating a fuzzy-rough approach to a reranking algorithm. The classification procedure relies on a fuzzy-rough nearest neighbor algorithm; the proposed method yielded 99.71% accurate classifications. Kong et al[11] proposed a new FS and discriminant analysis (DA) for sparse subspace learning and applied on the WBCD data set. The novel method, called Jointly Sparse Discriminant Analysis, achieved a maximum accuracy of 93.85% and also allowed the investigation of the most relevant factors for BC identification.

Table 1 summarizes the accuracy performance of several frameworks for WBCD classification. The classifiers presented below may be categorized according to the theoretical foundations on which they are based: statistics/support vector machine (S/SVM), decision trees/linear programming, neural network (NN), FT, FS, and DA.

In this study, we propose a method for FS and classification of cases into benign or malignant categories by deriving a feature importance index from the combination of principal component analysis (PCA) parameters (ie, weights and variance explained by each retained component) with the Bhattacharyya Distance (BD) method. Including the BD, which is a supervised method, along with PCA (which is unsupervised), information about classes of the training data is also included in the analysis, enhancing the selection of the most relevant features.[22] The current study also investigates the potential integration of 2 additional classification tools in the framework originally proposed in Fogliatto and Anzanello[22]: probabilistic NN (PNN) and linear DA (LDA).

Table 1. Classification Accuracies Obtained in the Wisconsin Breast Cancer Database With Propositions From the Literature.

Source	Method	Accuracy (%)
Kong et al[11]	FS and DA	93.85
Quinlan[12]	DT/LP	94.74
Nauck and Kruse[13]	FS and NN	95.06
Lee et al[14]	FS	95.14
Abonyi and Szeifert[8]	FS	95.57
Verikas and Bacauskiene[15]	NN	96.44
Setiono[16]	NN	96.58
Setiono[17]	NN	96.70
Street et al[5]	DT/LP	97.30
Peña-Reyes and Sipper[18]	FS	97.80
Fogel et al[6]	NN	98.05
Abbass[19]	NN	98.10
Polat and Günes[20]	S/SVM	98.53
Albrecht et al[21]	DT/LP	98.80
Marcano-Cedeño et al[7]	NN	99.26
Akay[2]	S/SVM	99.51
Marcano-Cedeño et al[9]	NN	99.63
Onan[10]	FT	99.71

Abbreviations: DA, discriminant analysis; DT/LP, Decision Trees/Linear Programming; FS, feature selection; FT, fuzzy theory; NN, neural network; S/SVM, statistics/support vector machine.

The index intended to merge the PCA ability to spot the features responsible for explaining most of the data variability with the BD skill to identify the features whose distributions mostly differ in terms of inserting observations into classes. The observations (ie, cases) of a training set were classified into 2 classes (benign or malignant) using a series of classification techniques: k-nearest neighbor (KNN), PNN, and LDA. The sets of features leading to the maximum accuracy were chosen and used to classify observations in the testing set. Measures of sensitivity, specificity, positive, and negative predicted values were also used for comparison.

The effectiveness of this FS strategy was evaluated against the commonly used WBCD and a new protein microarray data set.[23] The WBCD was used as a baseline for comparison to the literature because of its known results of FS in machine learning methods. Another data set from a study of Syed et al[23] was included for further comparison regarding the use of protein microarray analysis to evaluate the sera from patients with BC (malignant and benign) and control patients. Microarray data are usually comprised of hundreds or thousands of high correlated features, leading to data sets where the number of features

Table 2. Code and Description of Features in the Wisconsin Breast Cancer Database.

Code	Description	Code	Description
F_1	Clump thickness	F_6	Bare nuclei
F_2	Uniformity of cell size	F_7	Bland chromatin
F_3	Uniformity of cell shape	F_8	Normal nucleoli
F_4	Marginal adhesion	F_9	Mitosis
F_5	Single epithelial cell size		

is extremely larger than the number of samples (contrary to the WBCD in which the number of features is smaller than the number of patients). In both data sets, however, the selection of the most discriminant features plays an essential role and provides clinical information about the potential biomarkers that can aid in early cancer detection.

Materials and Methods

Databases

Our study did not require an ethical board approval because it was based on publicly available information from 2 online databases. The first was the WBCD, which is comprised of 699 samples (16 of which with missing values) obtained from fine-needle aspirates of human breast tissues. The fine-needle aspirates test allows investigating malignancy in breast masses in a rather cost-effective, noninvasive manner.[21] Nine features were measured in each sample, and observed values were written using an integer value scale of 10 points, in which 1 denotes the closest to benign situation; the features are listed in Table 2. A class label (benign or malignant) is associated with each sample. Considering the 683 complete samples, there are 339 malignant and 444 benign cases. The data set is available online (http://www.ics.uci.edu/~mlearn/MLRepository). Features are assessed by physicians upon analysis of 10 features related to the size, shape, and texture of cell nuclei (a complete description of these features is available in Street et al[5]). Three outcomes are automatically obtained for each feature: their mean and largest values and their standard deviations. The 30 resulting outcomes support the decision on the 9 features in Table 2. The evaluation process is highly subjective and dependent on the physician's skill and experience.

The second database was the protein microarray database, which is comprised of 642 features regarding the levels of protein expression analyzed by means of microarray chips[23] and is available online at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE34555. Out of the total features, 284 clones were used in which 181 proteins were identified. The study assessed 60 serum samples from patients and healthy women: 24 were related to malignant tumors, 16 to benign tumors, and 20 to healthy controls. The detailed experimental procedures and data acquisition are covered in Syed et al.[23] The FS procedure and classification used led to a classification accuracy of 93% using 50 features. They also reported 100% sensitivity and 85% specificity.

Data Analysis Techniques

Two methods were used to generate a feature importance index aimed at guiding feature removal: PCA and BD. The PCA is a widely used method of dimension reduction. This technique combines the original J features described in a matrix $\mathbf{X}$ into new uncorrelated features called principal components[24]; each new principal component explains as much as possible variance in the original data. Rencher[24] points out that the number of principal components extracted from the original data can be defined by the amount of explained variance. Two parameters derived from the PCA are of typical interest: the weight associated with feature j, w_{jr}, which is determined in a way that the variance between the components is maximized, and the percentage of variance explained by each retained component r $(r = 1, \ldots, R)$, λ_r.[25]

Bhattacharyya distance is a supervised technique that measures the similarity of 2 probability distributions,[26] in which each distribution refers to 1 of the 2 classes. The distance between class b (ie, denoting a benign case) and m (ie, denoting a malignant case) can be expressed by the sum of the similarity of each feature j at a given time, $B_j(b, m)$.[27] Equation 1 shows how these factors are related, considering variance σ^2 and mean μ of the statistical distributions of the jth feature for classes b and m. According to Jung et al,[28] the larger the BD of a feature, the farther away the distributions that describe the classes are. Aligned with the FS purpose of this study, features giving rise to large $B_j(b, m)$ are deemed relevant for correct diagnosis (ie, correct case assignment).

$$B_j(b,m) = \frac{1}{4} + \ln\left(\frac{1}{4}\left(\frac{\sigma_{bj}^2}{\sigma_{mj}^2} + \frac{\sigma_{mj}^2}{\sigma_{bj}^2} + 2\right)\right) + \frac{1}{4}\left(\frac{(\mu_{bj} - \mu_{mj})^2}{\sigma_{bj}^2 + \sigma_{mj}^2}\right) \tag{1}$$

Three classification techniques were selected to test our hypotheses to improve discrimination of the FS methods: KNN, (LDA, and PNN. k-Nearest Neighbor is a technique that assigns a sample to the same class of the closest KNN. The algorithm computes the (Euclidean) distance from the current sample to the already trained observations and assigns it to the most frequent class of the nearest k training observations. The number of k neighbors can be defined via cross-validation on the training set.[29]

Linear discriminant analysis is a method used to find linear combinations of the original features to define new features with similar characteristics, also named components. The goal is to maximize the discriminability in each component by maximizing the Fisher ratio, which is equivalent to minimizing intraclass variability and maximizing interclass variability at the same time. The new observations are transformed according the dimensions of the components of LDA.[24] In other words, each observed value is compared to a trained threshold and then assigned to the correspondent class of events.

Finally, PNN is an adaptation of NNs for classification developed by Specht.[30] Neural networks consist of input units that are linearly combined in the hidden layers, which are later integrated to an activation function (eg, logistic). The result is then summed to produce an output. In PNN, the activation function is replaced by a statistically derived function asymptotically approaching the Bayes optimal decision surface.[31]

Additional classification performance measures were evaluated for the testing set: sensitivity, specificity, positive predictive value, and negative predictive value. Sensitivity (Equation 2) is the probability of a sample to be classified as positive when it is truly malignant, and specificity (Equation 3) is the probability of a truly benign sample to be classified as negative. Positive and negative predictive values (Equations 4 and 5, respectively) represent a true positive and a true negative result, that is, the correct classification of the malignant and the benign cases.

$$Sensitivity = \frac{TP}{TP + FN}. \quad (2)$$

$$Specificity = \frac{TN}{TN + FP}. \quad (3)$$

$$Positive\,predictive\,value = \frac{TP}{TP + FP}. \quad (4)$$

$$Negative\,predictive\,value = \frac{FN}{FN + TN}. \quad (5)$$

Procedure for Data Analysis

The method to select features that best classify the data sets observations into 2 classes relies on 4 operational steps: (1) split original data set into training and testing sets and apply both PCA and BD to the training set; (2) generate attribute importance indices based on parameters emerging from PCA and BD; (3) classify the data set using each of the tested classification techniques and compute the classification accuracy and eliminate the feature with the lowest importance index, classify the data set again, and recompute the accuracy and continue the iterative process until 1 feature is left; and (4) choose the subset of features yielding the maximum classification accuracy and classify the testing set based on those features. These operational steps are detailed as follows.

In step (1), the data set was separated into training and testing sets. The training set was used to select key features, and the testing set was used to represent the new observations to be classified based on the selected features. Different proportions of training and testing sets were tested (see step 4). Three proportions of training to testing sets were used: 0.60 (ie, 60% of observations were kept in the training set and 40% in the testing set), 0.75, and 0.90. For each proportion, 1000 replicates were run on different training and testing sets obtained by randomly shuffling observations in the data sets. (Averaging performance metrics tend to give rise to more reliable measurement of the method's performance than a sole estimation obtained from a single partition of the data set into training and testing sets, which might lead to unreliable results.)

We then applied both PCA and BD methods to the training sets. The PCA's outputs of interest included the component weights (w_{jr}) and the percentage of variance λ_r explained by each retained component r ($r = 1, \ldots, R$). The BD's method was applied to a matrix comprised of the features of the original data sets and the class each sample belonged to. The method devised a parameter $B_j(b, m)$ proportional to the distance (or overlap) between the distributions for feature j in each class. The parameter was used to identify features whose distributions contributed the most in the categorization procedure.

In step 2, a feature importance index was generated to guide the removal of irrelevant features. Feature j's index was denoted by $v_j, j = 1, \ldots, J$, and was generated based on PCA weights w_{jr} (percentage of variance λ_r explained by each retained component) and on the BD output $B_j(b, m)$, as detailed in Equation 6. Features with large w_{jr}, λ_r, and $B_j(b, m)$ were preferred, since they represent high variability that enable better discrimination of observations into classes.[19] That is, the higher the v_j, the more important feature j was in categorizing observations into classes.

$$v_j = B_j(b,m) \times \sum_{r=1}^{R} |w_{jr}|\lambda_r,\ j = 1, \ldots, J. \quad (6)$$

In step 3, training observations were classified into 2 classes considering all J features using each of the 3 testing methods for categorization (KNN, LDA, and PNN). The feature with the smallest v_j was identified and removed, and a new classification was done using the $J - 1$ remaining features. The classification accuracy was computed again, and this procedure was repeated removing the next attribute with the smallest v_j until a single feature was left.

In step 4, the subset of features yielding the maximum accuracy was selected. In case there was more than one subset with identical accuracy values, the one with the smallest number of retained features was selected for parsimony. Next, the testing set was classified using the selected features and resulting accuracy was computed. We also computed sensitivity, specificity, positive-predictive value, and negative-predictive value for validation of the models.

In order to evaluate the consistency of the proposed method, steps 1 to 4 were conducted on different proportions of training and testing sets to assess the method's performance on small and large training data sets. The different data sets were generated by random shuffling and splitting of the data set being analyzed, certifying that all observations appeared at least once in the training set. The average classification criteria and number of retained features for each proportion was computed for final analysis, in which the subset of features yielding the best performance was identified. A value close to 1 in sensitivity, specificity, positive-predictive value, and negative-predictive value is preferred, since these metrics are related to classification performance. A smaller value is preferable when evaluating the number of retained features.

Results and Discussion

The WBCD

Table 3 depicts the average classification performance in the testing set and number of retained features for different classification techniques and data set partitions. The highest average

Table 3. Average Performance and Standard Deviation of Proposed Method for Different Data Set Partitions and Classification Techniques in the Wisconsin Breast Cancer Database.

		Classification Technique					
		KNN		LDA		PNN	
Data Set Partitions (% Training–% Testing)	Average Performance Criteria on Testing Set	Mean	SD	Mean	SD	Mean	SD
60%–40%	Accuracy	0.9717	0.0089	0.9649	0.0101	0.9742	0.0080
	Sensitivity	0.9637	0.0221	0.9328	0.0247	0.9768	0.0151
	Specificity	0.9766	0.0096	0.9830	0.0075	0.9731	0.0104
	Positive predictive value	0.9565	0.0186	0.9674	0.0148	0.9507	0.0200
	Negative predictive value	0.9798	0.0133	0.9634	0.0152	0.9872	0.0088
	Retained features	6.1050	1.8167	6.3250	1.7506	6.4450	1.7413
75%–25%	Accuracy	0.9745	0.0118	0.9663	0.0142	0.9802	0.0097
	Sensitivity	0.9713	0.0256	0.9388	0.0335	0.9838	0.0197
	Specificity	0.9769	0.0114	0.9820	0.0113	0.9789	0.0123
	Positive predictive value	0.9575	0.0219	0.9652	0.0224	0.9563	0.0235
	Negative predictive value	0.9836	0.0156	0.9666	0.0204	0.9929	0.0121
	Retained features	5.5650	1.9089	5.9400	1.8394	5.9750	1.8740
90%–10%	Accuracy	0.9777	0.0178	0.9680	0.0208	0.9917	0.0168
	Sensitivity	0.9808	0.0289	0.9402	0.0490	0.9921	0.0268
	Specificity	0.9777	0.0214	0.9848	0.0155	0.9924	0.0192
	Positive predictive value	0.9578	0.0421	0.9688	0.0325	0.9770	0.0362
	Negative predictive value	0.9877	0.0204	0.9665	0.0314	0.9990	0.0177
	Retained features	4.4450	1.8667	4.6800	2.1028	4.6100	1.9431

Abbreviations: KNN, k-Nearest Neighbor; LDA, linear discriminant analysis; PNN, Probabilistic Neural Network; SD, standard deviation.

accuracy (99.17%) was yielded by the PNN technique when applied to a proportion training to testing of 90% to 10%. This accuracy level was obtained when 4.61 features (out of the 9 original ones) were retained on average. This technique ranks among the best frameworks intended to classify WBCD observations, mostly due to the feature importance index based on PCA and BD parameters. Although our method does not outperform the benchmarking accuracy (99.71%[10]), the fact that it relies on straightforward and widely available multivariate techniques favors its use by researchers and practitioners.

When assessing the average performance of the tested techniques (regardless of data set partitions), PNN yielded average 98.20% of correct classifications, followed by KNN (97.46%) and LDA (96.64%). Probabilistic NN was also the best choice in terms of average sensitivity, specificity, and negative-predictive value results. In other words, taking all observations into account for the classification procedure (as proposed in the PNN fundamentals) led to better results than using a limited number of nearest observations (ie, KNN) or finding a linear combination of features (ie, LDA). Regarding the average number of retained features, there were no substantial differences on the performance of the tested techniques: Results ranged from 4.44 to 4.68 features when the 90% to 10% partition was considered.

As for the effect of different data set partitions upon the assessed metrics, results suggested that larger training sets provided the model with additional information to build a classifier aimed to rank features' importance in a more consistent way. Table 3 also presents the average standard deviation for different data set partitions and classification techniques. Probabilistic NN presents smaller deviations in most of the assessed metrics, suggesting that the technique is less impacted by different partitions of training and testing sets as replications were carried out. In light of the aforementioned results, PNN appears to be the best classification technique, with superior performance, intuitive fundamentals, and large availability in statistical packages.

Table 4 depicts the performance criteria when the 9 original features are used for classification. Overall, average accuracy using the selected subsets increased 0.3% when compared to categorization on all the original features. Although that increment in accuracy is not substantial, we emphasize that slight improvements are expected when classification performance is around 100%. In addition, the fact that almost 5 of the 9 original features for proportion 90%/10% are not relevant for classification significantly reduces the number of features to be measured on each patient, yielding simpler and cheaper data collection. That corroborates the FS procedure from a practical perspective.

Protein Microarray Database

In this data set, the 60 samples were divided based on breast-nodule positive and control.[23] The breast-nodule positive class contained 40 samples of malignant and benign tumors, and the control contains the 20 healthy samples. Table 5 presents the average classification performance and the average number of retained features for the protein microarray database. Different from the previous data set, KNN led to the highest average

Table 4. Average Performance and Standard Deviation for Different Data Set Partitions and Classification Techniques in the Wisconsin Breast Cancer Database Consisting of the 9 Original Features.

		Classification Technique					
		KNN		LDA		PNN	
Data Set Partitions (% Training–% Testing)	Average Performance Criteria on Testing Set	Mean	SD	Mean	SD	Mean	SD
60%–40%	Accuracy	0.9654	0.0073	0.9650	0.0089	0.9748	0.0074
	Sensitivity	0.9542	0.0196	0.9351	0.0199	0.9780	0.0132
	Specificity	0.9714	0.0103	0.9810	0.0078	0.9730	0.0091
	Positive predictive value	0.9476	0.0176	0.9637	0.0145	0.9514	0.0156
	Negative predictive value	0.9755	0.0102	0.9658	0.0102	0.9881	0.0071
	Retained features	9	0	9	0	9	0
75%–25%	Accuracy	0.9686	0.0104	0.9663	0.0118	0.9763	0.0102
	Sensitivity	0.9607	0.0218	0.9364	0.0253	0.9810	0.0171
	Specificity	0.9728	0.0134	0.9822	0.0109	0.9737	0.0125
	Positive predictive value	0.9501	0.0234	0.9658	0.0204	0.9525	0.0217
	Negative predictive value	0.9792	0.0113	0.9669	0.0128	0.9898	0.0091
	Retained features	9	0	9	0	9	0
90%–10%	Accuracy	0.9753	0.0168	0.9706	0.0196	0.9798	0.0165
	Sensitivity	0.9750	0.0312	0.9436	0.0456	0.9841	0.0260
	Specificity	0.9755	0.0213	0.9847	0.0170	0.9776	0.0201
	Positive predictive value	0.9557	0.0372	0.9708	0.0317	0.9594	0.0352
	Negative predictive value	0.9871	0.0159	0.9714	0.0226	0.9917	0.0135
	Retained features	9	0	9	0	9	0

Abbreviations: KNN, k-Nearest Neighbor; LDA, linear discriminant analysis; PNN, Probabilistic Neural Network; SD, standard deviation.

Table 5. Average Performance and Standard Deviation of Proposed Method for Different Data Set Partitions and Classification Techniques in the Protein Microarray Database.

		Classification Technique					
		KNN		LDA		PNN	
Data Set Partitions (% Training–% Testing)	Average Performance Criteria on Testing Set	Mean	SD	Mean	SD	Mean	SD
60%–40%	Accuracy	0.8610	0.0034	0.8330	0.0047	0.8707	0.0036
	Sensitivity	0.9210	0.0039	0.8808	0.0047	0.9323	0.0038
	Specificity	0.7410	0.0087	0.7375	0.0082	0.7475	0.0091
	Positive predictive value	0.8807	0.0035	0.8722	0.0038	0.8849	0.0037
	Negative predictive value	0.8392	0.0070	0.7656	0.0082	0.8606	0.0070
	Retained features	83.1591	6.4487	90.0245	10.0433	89.7315	8.4195
75%–25%	Accuracy	0.9037	0.0043	0.8629	0.0057	0.9027	0.0045
	Sensitivity	0.9540	0.0037	0.9128	0.0053	0.9732	0.0032
	Specificity	0.8032	0.0109	0.7632	0.0109	0.7616	0.0117
	Positive predictive value	0.9122	0.0045	0.8893	0.0049	0.8966	0.0047
	Negative predictive value	0.9100	0.0072	0.8282	0.0101	0.9431	0.0068
	Retained features	57.3426	5.3238	61.7002	8.2735	49.3899	5.7071
90%–10%	Accuracy	0.9830	0.0040	0.9580	0.0070	0.9730	0.0051
	Sensitivity	0.9867	0.0041	0.9587	0.0075	0.9907	0.0035
	Specificity	0.9720	0.0105	0.9560	0.0130	0.9200	0.0172
	Positive predictive value	0.9930	0.0026	0.9873	0.0038	0.9797	0.0044
	Negative predictive value	0.9520	0.0120	0.9067	0.0156	0.9080	0.0176
	Retained features	13.9635	2.2408	12.8601	2.3936	14.5373	3.3225

Abbreviations: KNN, k-Nearest Neighbor; LDA, linear discriminant analysis; PNN, Probabilistic Neural Network; SD, standard deviation.

accuracy when 90% of the data set was used in the training portion. The model achieved an average accuracy of 98.30% while retaining an average of 13.96 of 642 features. Syed et al[23] reported an accuracy of 93% retaining 50 features from the original data set. In that sense, our propositions outperformed the results originally obtained by such authors.

The average performance of the 3 classification tools was 91.59% for KNN, closely followed by PNN, with average accuracy of 91.54%. The LDA presented the worst performance in all assessed partitions, yielding average accuracy of 88.46%. Although KNN performed better regarding accuracy, PNN presented higher sensitivity (96.54% compared to 95.39% of KNN). As for the average number of retained features, both PNN and KNN performed similarly, with KNN retaining average 0.2688 more features. For this metric, LDA was outperformed by the other classification techniques: It retained average 54.86 features, while KNN retained 51.48. The only scenario in which LDA outperformed the other techniques was in the 90% to 10% data partition condition, retaining 12.86 features on average.

Table 5 also presents the variability in the results, measured as the standard deviation of the average performance criteria. k-Nearest Neighbor presented the best result for most performance metrics, only outperformed by PNN in sensitivity. It is noteworthy that the standard deviation of the average retained features decreased as the training proportion increased. Considering these facts, we recommend using the KNN algorithm with as much data to train the model as possible. Results achieved by the recommended method outperformed the ones currently available in the literature for this data set, leading to a more accurate and parsimonious model.

Conclusions

Early identification of BC increases the chance of remission given that there is no effective way to prevent the disease. In light of that, correctly classifying patients based on different data (eg, tissues biopsy or gene expression) can aid health-care professionals during the diagnosis process. Thus, multivariate frameworks aimed at reducing data set dimension and enhancing classification performance are highly desirable.

We proposed a framework that selects the most relevant features for categorizing patients into malignant or benign cases. For that matter, we first ranked features using a new feature importance index based on PCA and BD parameters; the combination of parameters derived from such methods was used to highlight features with higher variance and discriminant power. Next, our method iteratively classified patient records into proper classes through 3 classification techniques (KNN, LDA, and PNN): The less important feature indicated by the index was removed and classification was performed on the remaining features until a single feature was left. This method correctly classified average 99.17% instances of the WBCD using average 4.61 of the 9 original features. For the protein microarray database, the method yielded an average accuracy of 98.30% retaining an average of 16.96 of 642 features of the initial data set. Alternative performance metrics, such as sensitivity and specificity, also increased by means of the FS. The obtained performance may be paired with that of more complex classifications schemes available in the literature.

Further developments include the testing of more robust multivariate techniques to identify the most relevant features, and its integration to alternative data mining tools for classification. The transformation of original data using Kernel techniques aimed at improving classification performance of data mining tools is also intended. All scripts used in this manuscript are available as Supplemental Material.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Flavio S. Fogliatto https://orcid.org/0000-0002-0323-8060
Priscila G. Brust-Renck https://orcid.org/0000-0001-9891-510X

Supplemental Material

Supplemental material for this article is available online.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. *CA Cancer J Clin*. 2017;67(1):7-30.
2. Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Syst Appl*. 2009;36(2): 3240-3247. doi:10.1016/j.eswa.2008.01.009.
3. Shapiro S, Venet W, Strax P, Venet L, Roeser R. Ten- to fourteen-year effect of screening on breast cancer mortality. *J Natl Cancer Inst*. 1982;69(2):349-355. doi:10.1093/jnci/69.2.349.
4. Humphrey LL, Helfand M, Chan BK, Woolf SH. Breast cancer screening: a summary of the evidence for the U.S. Preventive Services Task Force. *Ann Intern Med*. 2002;137(5 pt 1):347-360. doi: 10.7326/0003-4819-137-5_Part_1-200209030-00012.
5. Street WN, Wolberg WH, Mangasarian OL. Nuclear feature extraction for breast tumor diagnosis. Paper presented at: IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology. 1993; San Jose, California. Volume 1905, 861-870.
6. Fogel DB, Wasson EC III, Boughton EM. Evolving neural networks for detecting breast cancer. *Cancer Lett*. 1995;96(1):49-53. doi:10.1016/0304-3835(95)03916-K.
7. Marcano-Cedeno A, Quintanilla-Dominguez J, Andina D. WBCD breast cancer database classification applying artificial metaplasticity neural network. *Expert Syst Appl*. 2011;38(8):9573-9579. doi:10.1016/j.eswa.2011.01.167.
8. Abonyi J, Szeifert F. Supervised fuzzy clustering for the identification of fuzzy classifiers. *Pattern Recognit Lett*. 2003;24(14): 2195-2207. doi:10.1016/S0167-8655(03)00047-3.
9. Marcano-Cedemo A, Quintanilla-Dominguez J, Andina D. Breast cancer classification applying artificial metaplasticity algorithm. *Neurocomputing*. 2011;74(8):1243-1250. doi:10.1016/j.neucom.2010.07.019.
10. Onan A. A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. *Expert Syst Appl*. 2015; 42(20):6844-6852. doi:10.1016/j.eswa.2015.05.006.

11. Kong H, Lai Z, Wang X, Liu F. Breast cancer discriminant feature analysis for diagnosis via jointly sparse learning. *Neurocomputing*. 2016;177:198-205. doi:10.1016/j.neucom.2015.11.033.
12. Quinlan JR. *C4.5: Programs for Machine Learning*. San Francisco, CA: Morgan Kaufmann Publishers Inc; 1993.
13. Nauck D, Kruse R. Obtaining interpretable fuzzy classification rules from medical data. *Artif Intell Med*. 1999;16:149-169. doi:10.1016/S0933-3657(98)00070-0.
14. Lee HM, Chen CM, Chen JM, et al. An efficient fuzzy classifier with feature selection based on fuzzy entropy. *IEEE Trans Syst Man Cybern B Cybern*. 2001;31:426-432. doi:10.1109/3477.931536.
15. Verikas A, Bacauskiene M. Feature selection with neural networks. *Pattern Recognit Lett*. 2002;23:1323-1335. doi:10.1016/S0167-8655(02)00081-8.
16. Setiono R. Extracting rules from pruned neural networks for breast cancer diagnosis. *Artif Intell Med*. 1996;8:37-51. doi:10.1016/0933-3657(95)00019-4.
17. Setiono R. Generating concise and accurate classification rules for breast cancer diagnosis. *Artif Intell Med*. 2000;18(3):205-219. doi:10.1016/S0933-3657(99)00041-X.
18. Pena-Reyes CA, Sipper M. A fuzzy-genetic approach to breast cancer diagnosis. *Artif Intell Med*. 1999;17(2):131-155. doi:10.1016/S0933-3657(99)00019-6.
19. Abbass HA. An evolutionary artificial neural networks approach for breast cancer diagnosis. *Artif Intell Med*. 2002;25(3):265-281. doi:10.1016/S0933-3657(02)00028-3.
20. Polat K, Güneş S. Breast cancer diagnosis using least square support vector machine. *Digit Signal Process*. 2007;17:694-701. doi:10.1016/j.dsp.2006.10.008.
21. Albrecht A, Lappas G, Vinterbo S, et al. Two applications of the LSA machine. In: *9th International Conference on Neural Information Process* (pp. 184-189); Nov 18-22, 2002; Singapore. 2002.
22. Fogliatto FS, Anzanello MZ. A data mining method for breast cancer diagnosis based on selected features [abstract]. In: *4o Congresso Internacional dos Hospitais, Associação Portuguesa para o Desenvolvimento Hospitalar*; Nov 7-9, 2012; Lisbon, Portugal. 2012.
23. Syed P, Vierlinger K, Kriegner A, et al. Evaluation of autoantibody serum biomarkers for breast cancer screening and in silico analysis of sero-reactive proteins. *J Mol Biochem*. 2012;1:116-128. http://www.jmolbiochem.com/index.php/JmolBiochem/article/view/53. Accessed December 21, 2018.
24. Rencher AC.*Methods of Multivariate Analysis*. Hoboken, NJ: Wiley Interscience; 2002.
25. Anzanello MJ, Fogliatto FS, Rossini K. Data mining-based method for identifying discriminant attributes in sensory profiling. *Food Qual Prefer*. 2011;22:139-148. doi:10.1016/j.foodqual.2010.08.010.
26. Fukunaga K. *Introduction to Statistical Pattern Recognition*. Cambridge, MA: Academic Press; 1990.
27. Coleman GB, Andrews HC. Image segmentation by clustering. *Proceedings of the IEEE*. 1979;67(5):773-785. doi:10.1109/PROC.1979.11327.
28. Jung CR, Ortiz RS, Limberger R, et al. A new methodology for detection of counterfeit Viagra and Cialis tablets by image processing and statistical analysis. *Forensic Sci Int*. 2012;216:92-96. doi:10.1016/j.forsciint.2011.09.002.
29. Chaovalitwongse WA, Fan YJ, Sachdeo RC. On the time series K-Nearest Neighbor classification of abnormal brain activity. *IEEE Trans Syst Man Cybern Part A Syst Humans*. 2007;37(6):1005-1016. doi:10.1109/TSMCA.2007.897589.
30. Specht DF. Probabilistic neural networks. *Neural Networks*. 1990;3:109-118. doi:10.1016/0893-6080(90)90049-Q.
31. Duda RO, Hart PE, Stork DG. *Pattern Classification*. Hoboken, NJ: John Wiley & Sons. 2000.