Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 209–217
DOI: 10.15439/2014F249
ACSIS, Vol. 2

Improving the performance of machine learning classifiers for Breast Cancer diagnosis based on feature selection

Noel Pérez*, Miguel A. Guevara†, Augusto Silva†, Isabel Ramos‡ and Joana Loureiro‡
*Institute of Mechanical Engineering and Industrial Management (INEGI)
Campus da FEUP, 4200-465 Porto, Portugal
Email: nperez@inegi.up.pt
†Institute of Electronics and Telematics Engineering of Aveiro (IEETA)
Campus Universitário de Santiago, 3810-193 Aveiro, Portugal
Email: {mguevaral, augusto.silva}@ua.pt
‡Faculty of Medicine - Centro Hospitalar São João (FMUP-HSJ)
Al. Prof. Hernâni Monteiro, 4200-319 Porto, Portugal
Email: radiologia.hs@mail.telepac.pt; joanaploureiro@gmail.com

***Abstract*—This paper proposed a comprehensive algorithm for building machine learning classifiers for Breast Cancer diagnosis based on the suitable combination of feature selection methods that provide high performance over the Area Under receiver operating characteristic Curve (AUC). The new developed method allows both for exploring and ranking search spaces of image-based features, and selecting subsets of optimal features for feeding Machine Learning Classifiers (MLCs). The method was evaluated using six mammography-based datasets (containing calcifications and masses lesions) with different configurations extracted from two public Breast Cancer databases. According to the Wilcoxon Statistical Test, the proposed method demonstrated to provide competitive Breast Cancer classification schemes reducing the number of employed features for each experimental dataset.**

I. Introduction

BREAST CANCER is a major concern and the second-most common and leading cause of cancer deaths among women [1]. According to published statistics, Breast Cancer has become a major health problem in both developed and developing countries over the past 50 years. Its incidence has increased recently with an estimated of 1,152,161 new cases in which 411,093 women die each year [2]. At present, there are no effective ways to prevent it, because its cause remains unknown. However, an efficient diagnosis in its early stages can give a woman a better chance of full recovery [2]. Therefore, its early detection can play an important role in reducing the associated morbidity and mortality rates.

Breast Cancer screening has proved to be the best way to detect cancer early. A useful and suggested approach is the double reading of mammograms (two radiologists read the same mammograms) [3], which has been advocated to reduce the proportion of missed cancers, but the workload and cost associated are high. With the support of Breast Cancer Computer-Aided Diagnosis (CADx) systems only one radiologist is needed to read each mammogram rather than two.

There is good evidence in the literature that Breast Cancer CADx systems can improve the AUC performance of radiologists [4] [5] [6] [7] [8], e.g. in [9], it was presented an evaluation of the variation of performance in terms of sensitivity and specificity of two radiologists with different experience in mammography (6 and 2 years respectively), with and without the assistance of two different CADx systems (SecondLook and CALMA). The evaluation was made on a dataset composed by 70 images of patients with cancer (biopsy proven) and 120 images of healthy breasts (with a three years follow up). The results showed that the use of a CADx allows for a substantial increment in sensitivity (up to 15.6%) and a less pronounced decrement in specificity, which was more significant for the least experienced of the radiologists.

However, the performance of current and future commercial CADx systems still needs to be improved so that they can meet the requirements of clinics and screening centers [10] [11].

In this work, we proposed a new method supported on the combination of five Feature Selection Methods (FSMs) and MLCs respectively, for building Breast Cancer classification schemes (i.e. calcifications and mases) that provide the high performance over the AUC curve. The selected FSMs are filter-based methods, which use heuristics (statistics) based on general characteristics of the data rather than a MLC (as wrapper or embedded paradigm) to evaluate the merit of features [12] [13] [14] [15]. As an optimal subset of features is always relative to a certain evaluation function [16], it was used different FSMs with different evaluation function: the traditional CHI-Square Discretization (CHI2) [17] based on the chi-square statistic function, Information Gain (IG) [18] based on the information measure, One-Rule (1Rule) [19] based on rules as a principal evaluation function and Re-

This work was supported by the Breast Cancer Digital Repository Consortium (BCDR - http://bcdr.inegi.up.pt))

lief [20] based on the distance measure. Also, it was used an algorithm developed in previous work [21] named RMean based on a voting function for indexing relevant features (see Algorithm 2, which is revisited here). The proposed method dynamically form subsets of features extracted from resultant rankings (one by each FSM applied) for feeding five machine learning models: Feed Forward Back-Propagation (FFBP) neural network [22], Support Vectors Machine (SVM) [23], Naive Bayes (NB) [24], Linear Discriminant Analysis (LDA) [25] and k-Nearest Neighbors (kNN) [26] respectively. Finally, the selection of the best classification scheme is based on the Wilcoxon Statistical Test [27] [28] following two criteria: (1) the higher obtained AUC value and (2) if there is classification performances tied, the one using less number of employed features is preferred. The method was evaluated on six datasets containing calcifications and masses lesions (with differents configurations) extracted from two public Breast Cancer databases and and it is included a statistical comparison of achieved results.

The remainder of the work is ordered as follows: the Materials and Methods section, overviews the employed databases, FSMs and MLCs. Also, describes in detail the proposed method and the experimental setup design for its evaluation. The Results and Discussion section presents an exploratory comparison based on the obtained AUC scores using the Wilcoxon statistical test [27] [28] to assess the meaningfulness of differences between classification schemes. Finally, Conclusions and Future work are drawn in the last section.

II. Materials and Methods

A. Databases

This work is supported on two public databases: the Breast Cancer Digital Repository (BCDR), which is the first wide-ranging annotated Portuguese Breast Cancer database, with anonymous cases from medical historical archives supplied by Faculty of Medicine - Centro Hospitalar de São João at University of Porto, Portugal [29] and the Digital Database for Screening Mammography (DDSM). For convenience, the DDSM images used in this study were obtained from the Image Retrieval in Medical Applications (IRMA) project (courtesy of TM Deserno, Dept. of Medical Informatics, RWTH Aachen, Germany) where the original LJPEG images of DDSM were converted to 16 bits Portable Network Graphics (PNG) format [30] [31].

The BCDR is composed of 1734 patient cases with mammography and ultrasound images, clinical history, lesion segmentation and selected pre-computed image-based descriptors; each case may have one or more Region of Interest (ROI) with associated Pathological Lesion (PL) segmentations (for different PLs), typically in Mediolateral Oblique (MLO) and Craniocaudal (CC) images of the same breast.

On the other hand, the DDSM database is composed by 2620 patient cases divided into three categories: normal cases (12 volumes), cancer cases (15 volumes) and benign cases (14 volumes); like in the BCDR, each case may have one or more associated PL segmentations, usually in MLO and CC image views of the same breast.

B. Feature Selection Methods

Several types of extracted features (e.g. intensity statistics, shape and texture) from mammograms have been combined to form subsets of features, which extensively provided significant information for lesions classification [32] [33] [34] [35]. However, selecting the most appropriate subset of features is still a very difficult task; usually a satisfactory instead of the optimal feature subset is searched.

The selected methods were all derived from the filter paradigm, because its execution is a one step process without any data exploration (search) involved and are also independent of classifiers [36].

1) CHI2 Discretization: This method consists on a justified heuristic for supervised discretization [17]. Numerical features are initially sorted by placing each observed value into its own interval. Then the chi-square statistic (x^2) is used to determine whether the relative frequencies of the classes in adjacent intervals are similar enough to justify merging. The extent of the merging process is controlled by an automatically set x^2 threshold. The threshold is determined through attempting to maintain the fidelity of the original data.

2) IG method: The IG measurement normalized with the symmetrical uncertainty coefficient [18] is a symmetrical measure in which the amount of information gained about Y after observing X is equal to the amount of information gained about X after observing Y (a measure of feature-feature intercorrelation). This model is used to estimate the value of an attribute Y for a novel sample (drawn from the same distribution as the training data) and compensates for information gain bias toward attributes with more values.

3) 1Rule: This method estimates the predictive accuracy of individual features building rules based on a single feature (can be thought of as single level decision trees) [19]. As it is used training and test datasets, it is possible to calculate a classification accuracy for each rule and hence each feature. Then, from classification scores, a ranked list of features is obtained. Experiments with choosing a selected number of the highest ranked features and using them with common machine learning algorithms showed that, on average, the top three or more features are as accurate as using the original set. This approach is unusual due to the fact that no search is conducted.

4) Relief: This method uses instance based learning to assign a relevance weight to each feature [20]. Each feature weight reflects its ability to distinguish among the class values. The feature weight is updated according to how well its values distinguish the sampled instance from its nearest hit (instance of the same class) and nearest miss (instance of opposite class). The feature will receive a high weight if it differentiates between instances from different classes and has the same value for instances of the same class. For nominal features it is defined as either 1 (the values are different) or 0 (the values are the same), while for numeric features the difference is the actual difference normalized to the interval $[0..1]$.

C. Machine Learning Models

The discrimination between samples of two classes may be formulated as a supervised learning problem, which is defined as the prediction of the value of a function for any valid input after training a learner using examples of input and target output pairs [37]. For the problem at hand, the function has only two discrete values: benign or malignant. Hence the problem can be modeled as a two-class classification problem. A variety of MLCs have been applied in CADx approaches for Breast Cancer detection/classification. The Artificial Neural Networks (ANN) [14] [23] [38] [39], SVM [14] [22] [23] [37] [39] [40] and LDA [41] [42] seem to be the most commonly used type of classifiers. Other less popular, but perform very well are NB [43] [44] [45] and kNN [25] [43] classifiers respectively.

A brief description of these MLCs is given here:

1) FFBP Neural Network Classifier: The FFBP neural network is a particular model of ANN, which provides a nonlinear mapping between its input and output according to the back-propagation error learning algorithm. This model has demonstrated to be capable of approximating an arbitrarily complex mapping within a finite support using only a sufficient number of neurons in few hidden layers (all layers using a sigmoid function as kernel type) [22].

2) SVM Classifier: SVMs are based on the definition of an optimal hyperplane, which linearly separates the training data. In comparison with other classification methods, a SVM aims to minimize the empirical risk and maximize the distances (geometric margin) of the data points from the corresponding linear decision boundary [23].

3) NB Classifier: The NB classifier is based on probabilistic models with strong (Naive) independence assumptions [24]. It assumes that c is a class variable depending on n input features: $x_1, x_2, \cdots, x_n$. The prediction of c can be described by the following conditional model: $p(c|x_1, x_2, \cdots, x_n)$ and according to the Bayes' theorem:

$$p(c|x_1, x_2, \cdots, x_n) = \frac{p(c)p(x_1, x_2, \cdots, x_n|c)}{p(x_1, x_2, \cdots, x_n)}$$

where $p(c)$ is the prior probability of c, $p(x_1, x_2, \cdots, x_n|c)$ is the conditional probability depending on c, and $p(x_1, x_2, \cdots, x_n)$ is the probability of input features. If each feature x_i is conditionally dependent, as the denominator $p(x_1, x_2, \cdots, x_n)$ does not depend on c, which is actually a constant when features are given; the conditional probability over the class variable c can be expressed as:

$$p(c|x_1, x_2, \cdots, x_n) = \frac{1}{z}p(c)\prod_{i=1}^{n} p(x_i|c)$$

where z is a normalization constant. The above NB classifier can be trained based on the relative frequencies shown in the training set to get an estimation of the class priors and feature probability distributions. For a test sample, the decision rule will be picking the most probable hypothesis (value of c) which is known as the maximum a posteriori decision rule using the above model.

4) LDA Classifier: LDA is a traditional method for classification [25]. The basic idea is to try to find an optimal projection (decision boundaries optimized by the error criterion), which can maximize the distances between samples from different classes and minimize the distances between samples from the same class. For the binary classification, observations are classified by the following linear function:

$$g_i(x) = W_i^T x - c_i \quad 1 \leq i \leq 2$$

where W_i^T is the transpose of a coefficient vector, x is a feature vector and c_i is a constant as the threshold. The values of W_i^T and c_i are determined through the analysis of a training set. Once these values are determined, they can be used to classify the new observations (smallest $g_i(x)$ is preferred).

5) kNN Classifier: The kNN classifier is a nonparametric technique called a "lazy learning" because little effort goes into building the classifier and most of the work is performed at the time of classification. The kNN assigns a test sample to the class of the majority of its k-neighbors; that is, assuming that the number of voting neighbors is $k = k_1 + k_2 + k_3$ (where k_i is the number of samples from class i in the k-sample neighborhood of the test sample, usually computed using the Euclidean distance), the test sample is assigned to class m if $k_m = max(k_i), i = 1, 2, 3$ [26].

D. Proposed Method

The proposed method is supported on the combination of five FSMs and MLCs respectively, for building Breast Cancer classification schemes that provide the high performance over the AUC curve.

The employed FSMs are filter methods, which use heuristics (statistics) based on general characteristics of the data rather than a MLC (as wrapper or embedded paradigm) to evaluate the merit of features [12] [13]. As an optimal subset of features is always relative to a certain evaluation function [16], the selected FSMs were: CHI2 discretization [17] based on the chi-square statistic function, IG [18] based on the information measure, 1Rule [19] based on rules as a principal evaluation function, Relief [20] based on the distance measure and the recently developed RMean method [21] based on a voting function (averaging each feature position) for indexing relevant features (see Algorithm 2, which revisited here).

As it is shown in the Algorithm 1, the dataset D and the total of features in the initial subset nS constituted the starting point of the proposed method. Once, this method is a multistep modelling procedure, the application of the k-fold Cross Validation (CV) method [46] to the entire sequence of modelling steps guarantee reliable results [47]. Thus, it was applied 10 times 10-CV before features ranking to avoid giving an unfair advantage to predictors, and before classification step to prevent overfitting of classifiers to the training set [46] (see Algorithm 1 step 3 and 13). The application of FSMs on the processed dataset S_{cv} produced five different ranking of features (see Algorithm 1 step 4 to 8). Then, from each ranking of features, it were dynamically built ranked subset of features with different size S_{ini}.

Algorithm 1 Proposed method

Require: $D[f_1, f_2, f_3, ..., f_n] : n \geq 2$;
$nS \leftarrow$ maximum number of features in the initial subset;
Ensure: $C_{(best)}$; Best classification scheme
1: $C_{(best)} = []; C_{(aux)} = []; S_{ini} = []; S_{cv} = []; D_{cv} = []; R_{CHI2} = []; R_{IG} = []; R_{1R} = []; R_{Rel} = []; L = []$;
2: $nF \leftarrow nS$; Initializing nF
3: $D_{cv} \leftarrow$10-$CV(D)$; Applying 10 times 10-CV
4: $R_{CHI2} = eval(CHI2, D_{cv})$; Ranking by CHI2
5: $R_{IG} = eval(IG, D_{cv})$; Ranking by IG
6: $R_{1R} = eval(1R, D_{cv})$; Ranking by 1R
7: $R_{Rel} = eval(Relief, D_{cv})$; Ranking by Relief
8: $R_{RMean} = eval(RMean, D_{cv})$; Ranking by RMean
9: $L = [R_{CHI2}, R_{IG}, R_{1R}, R_{Rel}, R_{RMean}]$; List of ranking
10: **for** $i = 1 : length(L)$ **do**
11: **for** $j = 1 : trunc(L_i/nS)$ **do**
12: $S_{ini} \leftarrow extract(nF, L_i)$; Extract the first nF features from L_i
13: $S_{cv} \leftarrow$10-$CV(S_{ini})$; Applying 10 times 10-CV
14: $C_{(i,j,FFBP)} \leftarrow eval(FFBP, S_{cv})$ Applying the FFBP
15: $C_{(i,j,SVM)} \leftarrow eval(SVM, S_{cv})$ Applying the SVM
16: $C_{(i,j,NB)} \leftarrow eval(NB, S_{cv})$ Applying the NB
17: $C_{(i,j,LDA)} \leftarrow eval(LDA, S_{cv})$ Applying the LDA
18: $C_{(i,j,kNN)} \leftarrow eval(kNN, S_{cv})$ Applying the kNN
19: $nF \leftarrow nF + nS$; Updating the number of features nF
20: **end for**
21: $C_{(aux)} \leftarrow C_{(aux)} + max(C)$; Higher statistically
22: **end for**
23: $C_{(best)} \leftarrow max(C_{(aux)})$; Higher statistically

These ranked subsets of features were processed by the 10-CV method before feeding the FFBP neural network [22], SVM [23], NB [24], LDA [25] and kNN [26] classifiers respectively (see Algorithm 1 step 13 to 18). In the last step, two important criteria are evaluated in order to select the best classification scheme: (1) the higher obtained AUC value and (2) if there is classification performances tied, the one using less number of employed features is preferred. Both criteria were conducted using the Wilcoxon Statistical Test, i.e. a non-parametric alternative test to the paired t-test, which ranks the differences in performances of two classifiers [27] [28]. This test provided a fairly comparison among all obtained AUC performances, and therefore a reasonable selection of $C_{(best)}$.
The implementation of the proposed method was in JAVA language and the source code of all employed FSMs and MLCs are available in the WEKA data mining software version 3.6 [48].

E. Experimental Setup

This section outlines the experimental evaluation design of the proposed method using two public Breast Cancer databases. That involves the datasets creation and machine learning models configurations are important aspects to be described here.

Algorithm 2 RMean

Require: $D[f_1, f_2, f_3, ..., f_n] : n \geq 2$;
Ensure: R_{Mean};
1: $R_{Mean} = []; R_{CHI2} = []; R_{IG} = []; R_{1R} = []; R_{Rel} = []; D_{cv} = []$;
2: $D_{cv} \leftarrow$10-$CV(D)$; Applying 10 times 10-CV
3: $R_{CHI2} \leftarrow eval(CHI2, D_{cv})$; Ranking by CHI2
4: $R_{IG} \leftarrow eval(IG, D_{cv})$; Ranking by IG
5: $R_{1R} \leftarrow eval(1R, D_{cv})$; Ranking by 1R
6: $R_{Rel} \leftarrow eval(Relief, D_{cv})$; Ranking by Relief
7: $R_{Mean} \leftarrow (R_{CHI2} + R_{IG} + R_{1R} + R_{Rel})/4$; Averaging the features position throughout resultant rankings from steps 3,4,5 and 6.
8: $R_{Mean} \leftarrow sort(R_{Mean}, 'ascendant')$; Sorting in ascendant way the resultant ranking from the step 6.

1) Datasets Creation: A set of 23 image-based descriptors (features) were extracted from the BCDR and DDSM databases to be used in this work. Selected descriptors included intensity statistics, shape and texture features, computed from segmented calcifications and masses in both MLO and CC mammography views. The intensity statistics and shape descriptors were selected according to the radiologists experience (similar to the clinician procedure) and the American College of Radiology (BIRADS-Mammography atlas) [49], which described in detail how to detect/classify pathological lesions. Additionally, texture descriptors were the Halarick's descriptors extracted from the grey-level co-occurrence matrices [50]. An overview of the mathematical formulation for computing features is presented below:

- *Skewness*:
$$f_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^3}$$
with x_i being the i^{th}-value and $\bar{x}$ the sample mean.
- *Kurtosis*:
$$f_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^2} - 3$$
with x_i being the i^{th}-value and $\bar{x}$ the sample mean.
- *Area Fraction* (f_3): is the percentage of non-zero pixels in the image or selection.
- *Circularity*:
$$f_4 = 4\pi\frac{area}{perimeter^2}$$
- *Perimeter*: $f_5 = length(E)$ with $E \subset O$ being the edge pixels.
- *Elongation*: $f_6 = \left(\frac{m}{M}\right)$ with m being the minor axis and M the major axis of the ellipse that has the same normalized second central moments as the region surrounded by the contour.

- *Standard Deviation*:

$$f_7 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

with x_i being the grey level intensity of the i^{th}-pixel and $\bar{x}$ the mean of intensity.

- *Roughness*:

$$f_8 = \frac{perimeter^2}{4\pi * area}$$

- *Minimum* (f_9) and *Maximum* (f_{10}): the minimum and maximum intensity value in the region surrounded by the contour.
- *Shape*:

$$f_{11} = \frac{perimeter * elongation}{8 * area}$$

- *X Centroid*:

$$f_{12} = \frac{min(x) + max(x)}{2}$$

with x being the set of X coordinates of the object's contour.

- *Entropy*:

$$f_{13} = \sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)log(p(i,j))$$

with *L* being the number of grey-levels, and $p(i,j)$ being the probability of pixels with grey-level i occur together to pixels with grey-level j.

- *X Center Mass* (f_{14}): normalized X coordinates of the center of mass of O.
- *Angular Second Moment*:

$$f_{15} = \sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)^2$$

with L being the number of grey-levels, and $p(i,j)$ being the probability of pixels with grey-level i occur together to pixels with grey-level j.

- *Median*:

$$f_{16} = \begin{cases} \frac{n+1}{2} & \text{if } length(X) \text{ is odd} \\ \frac{X(\frac{n}{2}) + X(\frac{n}{2}+1)}{2} & \text{if } length(X) \text{ is even} \end{cases}$$

with X being the set of intensities.

- *Contrast*:

$$f_{17} = \sum_{i}\sum_{j} p(i,j)(i-j)^2$$

with $p(i,j)$ being the probability of pixels with grey-level i occur together to pixels with grey-level j.

- *Correlation*:

$$f_{18} = \frac{\sum_i \sum_j [ijp(i,j)] - \mu_x\mu_y}{\sigma_x\sigma_y}$$

with μ_x, μ_y, σ_x and σ_y being the means and standard deviations of the marginal distribution associated with $p(i,j)$.

- *Mean*:

$$f_{19} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

with n being the number of pixels inside the region delimited by the contour and x_i being the grey level intensity of the i^{th} pixel inside the contour.

- *Inverse Difference Moment*:

$$f_{20} = \sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j)$$

with $p(i,j)$ being the probability of pixels with grey-level i occur together to pixels with grey-level j.

- *Y Center Mass* (f_{21}): normalized Y coordinates of the center of mass of O.
- *Area*: $f_{22} = |O|$ with O being the set of pixels that belong to the segmented lesion.
- *Y Centroid*:

$$f_{23} = \frac{min(y) + max(y)}{2}$$

with y being the set of Y coordinates of the object's contour.

Conformable to the number of patient cases of used databases, it were created six datasets containing calcifications and masses lesions with different configurations: (1) two balanced datasets (same quantity of benign and malignant instances), (2) two unbalanced datasets containing more benign than malignant instances and (3) two unbalanced datasets holding more malignant than benign instances, representatives of BCDR and DDSM respectively. The BCDR supplies several datasets for scientific purposes (Availaible on http://bcdr.inegi.up.pt), we used the BCDR-F01 distribution to form the BCDR1 dataset holding 374 features vectors; BCDR2 and BCDR3 datasets with a total of 287 features vectors respectively.

Due to the wide range of information in the DDSM database, it were considered only two volumes of cancer and benign cases (random selection) to form the DDSM1 dataset holding 582 features vectors; DDSM2 and DDSM3 datasets with a total of 491 features vectors respectively. Figure 1 shows a detailed description of the datasets creation workflow.

2) MLCs Configuration: For all MLCs with the exception of the NB (which is parameterless), 10-CV method [46] was performed on the training set for optimizing the classifiers parameters.

The FFBP neural network was used with a total of hidden layers determined according to the equation (*attributes* + *number of classes*)/2; one output layer associated with the binary classification (benign or malignant); transfer function for all layers based on the sigmoid function and the number of iterations (epochs) were optimized in the range of 100 to 1000 epochs (with an interval increment of 100 units).

The SVM classifier was used with the regularization parameter *C* (cost) optimized in the range of 10^{-3} to 10^3 and the kernel

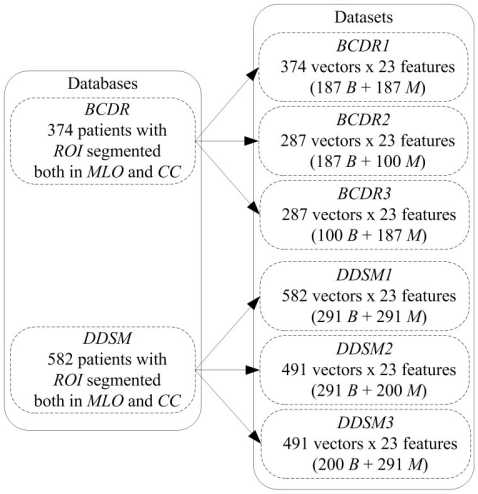

Fig. 1. Datasets creation flowchart; *B* and *M* represent Benign and Malignant class instances.

type based on a linear function, which provided better results respect to others kernel such as: radial basis, polynomial and sigmoid function (from our experimental experience).
The kNN classifier included the estimation of an optimal value k for the size of the neighborhood varying from 1 to 20, and the contribution of each neighbor was always weighted by the distance to the instance being classified.

III. Results and Discussion

According to the experimental setup section, a total of 750 ranked subsets of features containing image-based features extracted from segmented calcifications and masses lesions were analyzed using the proposed method and the straightforward statistical comparison based on the mean of AUC performances over 100 runs highlighted interesting results for balanced and unbalanced datasets respectively.

3) Performance on Balanced Datasets: The higher AUC value obtained in the BCDR1 dataset was formed by the SVM classifier and the RMean method using a total of 15 features (AUC value of 0.8365). This result was statistically superior to the majority of the remaining classification schemes. However, there were others schemes with similar performances statistically at p=0.05 (see Table I). From these results, it is possible to select the FFBP neural network in combination with the CHI2 discretization and RMean method with 5 features as the most appropriate classification schemes in this dataset. They reached an AUC value of 0.8264 and 0.8272 using the minimum number of features respectively.
For DDSM1 dataset, the higher AUC value was obtained by the combination of the LDA classifier and the RMean method using 20 features (AUC value of 0.807). However, this result did not provide statistical evidence to be better than others combinations (see Table I). Similar to the DDSM1 dataset, the combinations formed by the FFBP neural network in conjunction with the Relief and RMean methods using 10 features provided similar performances

TABLE I
Classification schemes with nonsignificant difference in AUC performances for BCDR1 and DDSM1 balanced datasets

Dataset	Best Scheme	Other Scheme	AUC	p=0.05
BCDR1	SVM+RMean+15F (0.8365)	FFBP+CHI2+5F	0.8264	p=0.574
		FFBP+RMean+5F	0.8272	p=0.494
		FFBP+CHI2+10F	0.8219	p=0.359
		SVM+Relief+15F	0.8831	p=0.603
		LDA+RMean+15F	0.8284	p=0.295
		LDA+RMean+20F	0.8279	p=0.286
DDSM1	LDA+RMean+20F (0.807)	FFBP+Relief+10F	0.8061	p=0.592
		FFBP+RMean+10F	0.8056	p=0.475
		SVM+RMean+20F	0.7939	p=0.139

TABLE II
Classification schemes with nonsignificant difference in AUC performances for BCDR2 and DDSM2 unbalanced datasets

Dataset	Best Scheme	Other Scheme	AUC	p=0.05
BCDR2	SVM+RMean+10F (0.8389)	FFBP+RMean+10F	0.8352	p=0.573
		LDA+CHI2+10F	0.8278	p=0.365
		LDA+RMean+10F	0.8284	p=0.403
DDSM2	FFBP+RMean+5F (0.8406)	FFBP+IG+10F	0.8405	p=0.682

statistically (AUC value of 0.8061 and 0.8056 respectively). These results were obtained using a less number of employed features. Thus, both classification schemes were selected as the most appropriated classification schemes in this dataset.

4) Performance on Unbalanced Datasets: The higher AUC performance for BCDR2 dataset was formed by the SVM classifier and the RMean method using 10 features, reaching an AUC value of 0.8389. This result was not statistically superior to obtained results by others classification schemes

TABLE III
Classification schemes with nonsignificant difference in AUC performances for BCDR3 and DDSM3 unbalanced datasets

Dataset	Best Scheme	Other Scheme	AUC	p=0.05
BCDR3	LDA+RMean+5F (0.8611)	FFBP+RMean+5F	0.8562	p=0.592
DDSM3	LDA+RMean+20F (0.807)	FFBP+Relief+10F	0.8061	p=0.592
		FFBP+RMean+5F	0.78	p=0.094
		SVM+Relief+10F	0.7879	p=0.139
		SVM+Relief+15F	0.7853	p=0.126
		SVM+RMean+10F	0.786	p=0.129
		SVM+RMean+15F	0.783	p=0.128
		NB+Relief+10F	0.785	p=0.125
		NB+RMean+10F	0.783	p=0.118
		LDA+Relief+10F	0.7883	p=0.145
		LDA+Relief+15F	0.7877	p=0.139
		LDA+1R+20F	0.7845	p=0.12
		LDA+RMean+10F	0.789	p=0.153
		LDA+RMean+15F	0.7861	p=0.135

using the same number of employed features(see Table II). Therefore, the four combinations presented in the Table II could be considered as the most appropriated schemes for lesions classification in the BCDR2 dataset.

Besides, for DDSM2 dataset the best classification performance was obtained by the combination of the FFBP neural network classifier and the RMean method using 5 features; reaching AUC value of 0.8406. However, this result did not statistically outperform the obtained result by the combination of the FFBP neural network classifier and the IG method with 10 features, attainment an AUC value of 0.8405 (see Table II). Despite the small and insignificant difference in term of AUC performances, the first combination was selected as the most appropriated classification scheme because it reached this results using a less number of features.

The best classification performance for BCDR3 dataset was provided by the combination of the LDA classifier and the RMean method with 5 features, accomplishment an AUC score of 0.8611 (see Table III). This result was not statistically superior to the obtained result by the combination of the FFBP neural network classifier and the RMean method with 5 features, which achieved an AUC score of 0.8562. As both classification schemes reached these results using the minimum number of employed features could be considered as the most appropriated classification schemes for BCDR3 dataset.

In the DDSM3 dataset, the higher AUC value was obtained by the combination of LDA classifier and the RMean method with a total of 10 features (AUC value of 0.7889). This classification result showed nonsignificant difference respect to others combinations, which reached similar AUC performances (see Table III). Despite the several combinations which can be used as a good classification scheme for this dataset. Only the scheme formed by the FFBP neural network and the RMean method stretched the result using the minimum number of features (5). Thus, it was considered as the most appropriated classification scheme in the DDSM3 dataset (see Table III).

Regarding of the classifiers performance, results show that the selection of the most appropriated classifier is dependent on the dataset and the FSM. From Table I, II and III, it possible to read that the best MLC was the FFBP neural network classifier, appearing consistently on every appropriated classification scheme for all datasets. These results were expected since this classifier has demonstrated to be capable of generalizing decision boundary in a more complex features space [25]. Meanwhile the best FSM was the RMean method (see Algorithm 2), which appeared consistently on every successful classification scheme, providing in most cases the minimal subset of features.

IV. Conclusions and Future Work

In this work, it is made a statistical exploration of different classification schemes within the context of Breast Cancer classification. The main contribution it was developed a new and robust method for building machine learning classifiers that combines suitably several feature selection methods with different evaluation function. This method was effective in providing competitive classification schemes for balanced and unbalanced datasets: the FFBP neural network and the RMean method using 5 features was the best scheme for BCDR1, BCDR3, DDSM2 and DDSM3 datasets, attainment an AUC value of 0.8264, 0.8562, 0.8406 and 0.78 respectively. Also, the FFBP neural network and the RMean method with 10 features in the DDSM1 dataset, reaching an AUC value of 0.8056, and the SVM classifier with the RMean method using 10 features for the BCDR2, stretching an AUC value of 0.8399. Regarding to MLCs and FSMs, the FFBP neural network classifier and the RMean method were the best, appearing consistently in the majority of successful schemes. In future work, we plan to assess the performance using others benchmarking datasets with different experimental setup: including clinical and more image-based features to evaluate the sensibility and generalization of the proposed method. Also, it's further integration in a real Breast Cancer CADx system.

Acknowledgment

MSc. Pérez acknowledges "Fundação para a Ciência e a Tecnologia (grant SFRH/BD/48896/2008)" for financial support. Prof. Guevara acknowledges the Cloud Thinking project (CENTRO-07-ST24-FEDER-002031), co-funded by QREN, "Mais Centro" program. The institutions participating in the Breast Cancer Digital Repository Consortium express their gratitude for the support of the European Regional Development Fund.

References

[1] M. D. Althuis, J. M. Dozier, W. F. Anderson, S. S. Devesa, and L. A. Brinton, "Global trends in breast cancer incidence and mortality 1973-1997", Int. J. Epidemiol., vol. 34, pp. 405-412, April 1, 2005, http://dx.doi.org/10.1093/ije/dyh414.

[2] F. Kamangar, G. M. Dores, and W. F. Anderson, "Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world", Journal of clinical oncology, vol. 24, pp. 2137-2150, 2006, http://dx.doi.org/10.1200/JCO.2005.05.2308.

[3] J. Brown, S. Bryan, and R. Warren, "Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms", BMJ (Clinical research ed.), vol. 312, pp. 809-812, 1996, http://dx.doi.org/10.1136/bmj.312.7034.809.

[4] Z. Huo, M. L. Giger, C. J. Vyborny, and C. E. Metz, "Breast cancer: effectiveness of computer-aided diagnosis observer study with independent database of mammograms", Radiology, vol. 224, pp. 560-8, Aug 2002, http://dx.doi.org/10.1148/radiol.2242010703.

[5] L. Hadjiiski, H. P. Chan, B. Sahiner, M. A. Helvie, M. A. Roubidoux, C. Blane, et al., "Improvement in radiologists' characterization of malignant and benign breast masses on serial mammograms with computer-aided diagnosis: an ROC study", Radiology, vol. 233, pp. 255-65, Oct 2004, http://dx.doi.org/10.1148/radiol.2331030432.

[6] L. Hadjiiski, B. Sahiner, M. A. Helvie, H. P. Chan, M. A. Roubidoux, C. Paramagul, et al., "Breast masses: computer-aided diagnosis with serial mammograms", Radiology, vol. 240, pp. 343-56, Aug 2006, http://dx.doi.org/10.1148/radiol.2401042099.

[7] K. Horsch, M. L. Giger, C. J. Vyborny, L. Lan, E. B. Mendelson, and R. E. Hendrick, "Classification of breast lesions with multimodality computer-aided diagnosis: observer study results on an independent clinical data set", Radiology, vol. 240, pp. 357-68, Aug 2006, http://dx.doi.org/10.1148/radiol.2401050208.

[8] L. A. Meinel, A. H. Stolpen, K. S. Berbaum, L. L. Fajardo, and J. M. Reinhardt, "Breast MRI lesion classification: Improved performance of human readers with a backpropagation neural network computer-aided diagnosis (CAD) system", Journal of Magnetic Resonance Imaging, vol. 25, pp. 89-95, 2007, http://dx.doi.org/10.1002/jmri.20794.

[9] A. Lauria, M. E. Fantacci, U. Bottigli, P. Delogu, F. Fauci, B. Golosio, et al., "Diagnostic performance of radiologists with and without different CAD systems for mammography", in Medical Imaging 2003, 2003, pp. 51-56, http://dx.doi.org/10.1117/12.480079.

[10] E. D. Pisano, C. Gatsonis, E. Hendrick, M. Yaffe, J. K. Baum, S. Acharyya, et al., "Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening", N Engl J Med, vol. 353, pp. 1773-1783, October 27, 2005, http://dx.doi.org/10.1056/NEJMoa052911.

[11] S. Ciatto, N. Houssami, D. Gur, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, et al., "Computer-Aided Screening Mammography", N Engl J Med, vol. 357, pp. 83-85, July 5, 2007, http://dx.doi.org/10.1056/NEJMc071248.

[12] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection", J. Mach. Learn. Res., vol. 3, pp. 1157-1182, 2003.

[13] I. Guyon and A. Elisseeff, "An Introduction to Feature Extraction", in Feature Extraction. vol. 207, I. Guyon, M. Nikravesh, S. Gunn, and L. Zadeh, Eds., ed: Springer Berlin Heidelberg, 2006, pp. 1-25, http://dx.doi.org/10.1007/978-3-540-35488-8_1.

[14] N. Pérez, M. A. Guevara, and A. Silva, "Improving breast cancer classification with mammography, supported on an appropriate variable selection analysis", in SPIE Medical Imaging, 2013, pp. 867022-867022-14, http://dx.doi.org/10.1117/12.2007912.

[15] N. Pérez, M. A. Guevara, and A. Silva, "EVALUATION OF FEATURES SELECTION METHODS FOR BREAST CANCER CLASSIFICATION", Icem15: 15th International Conference on Experimental Mechanics, p. 10, 2012.

[16] M. Dash and H. Liu, "Feature Selection for Classification", Intelligent Data Analysis, vol. 1, pp. 131-156, Jan 1, 1997.

[17] H. Liu and R. Setiono, "Chi2: Feature Selection and Discretization of Numeric Attributes", 1995, pp. 388-388, http://dx.doi.org/10.1109/TAI.1995.479783.

[18] B. P. Flannery, W. H. Press, S. A. Teukolsky, and W. Vetterling, "Numerical recipes in C", Press Syndicate of the University of Cambridge, New York, 1992.

[19] R. Holte, "Very Simple Classification Rules Perform Well on Most Commonly Used Datasets", Machine Learning, vol. 11, pp. 63-90, 1993/04/01 1993, http://dx.doi.org/10.1023/A:1022631118932.

[20] K. Kira and L. A. Rendell, "A practical approach to feature selection", presented at the Proceedings of the ninth international workshop on Machine learning, Aberdeen, Scotland, United Kingdom, 1992.

[21] N. Pérez, M. A. Guevara, A. Silva, and I. Ramos, "Ensemble features selection method as tool for Breast Cancer classification", Computing and Informatics, 2013, unpublished (under review).

[22] Y. H. Hu and J.-N. Hwang, "Introduction to Neural Networks for Signal Processing", in Handbook of neural network signal processing, ed: CRC press, 2001.

[23] A. Papadopoulos, D. I. Fotiadis, and A. Likas, "Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines", Artificial Intelligence in Medicine, vol. 34, pp. 141-150, Jun 2005, http://dx.doi.org/10.1016/j.artmed.2004.10.001.

[24] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification (2nd Edition): Wiley-Interscience, 2000.

[25] M. A. Alolfe, A. M. Youssef, Y. M. Kadah, and A. S. Mohamed, "Computer-Aided Diagnostic System based on Wavelet Analysis for Microcalcification Detection in Digital Mammograms", in Biomedical Engineering Conference, 2008. CIBEC 2008. Cairo International, 2008, pp. 1-5, http://dx.doi.org/10.1109/CIBEC.2008.4786080.

[26] S. Wang and R. M. Summers, "Machine learning and radiology", Medical Image Analysis, vol. 16, pp. 933-951, 2012, http://dx.doi.org/10.1016/j.media.2012.02.005.

[27] J. Demšar, "Statistical comparisons of classifiers over multiple data sets", The Journal of Machine Learning Research, vol. 7, pp. 1-30, 2006.

[28] M. Hollander and D. A. Wolfe, Nonparametric statistical methods, 2nd Edition ed.: Wiley-Interscience, 1999.

[29] R. Ramos-Pollan, M. A. Guevara-Lopez, C. Suarez-Ortega, G. Diaz-Herrero, J. M. Franco-Valiente, M. Rubio-Del-Solar, et al., "Discovering mammography-based machine learning classifiers for breast cancer diagnosis", J Med Syst, vol. 36, pp. 2259-69, Aug 2012, http://dx.doi.org/10.1007/s10916-011-9693-2.

[30] J. E. de Oliveira, A. M. Machado, G. C. Chavez, A. P. Lopes, T. M. Deserno, and A. Araujo Ade, "MammoSys: A content-based image retrieval system using breast density patterns", Comput Methods Programs Biomed, vol. 99, pp. 289-97, Sep 2010, http://dx.doi.org/10.1016/j.cmpb.2010.01.005.

[31] Júlia E. E. Oliveira, Mark O. Gueld, Arnaldo de A. Araújo, Bastian Ott, and T. M. Deserno., "Towards a Standard Reference Database for Computer-aided Mammography", in SPIE - Medical Imaging 2008: Computer-Aided Diagnosis, 69151Y, 2008, http://dx.doi.org/10.1117/12.770325.

[32] H. Soltanian-Zadeh, F. Rafiee-Rad, and S. Pourabdollah-Nejad D, "Comparison of multiwavelet, wavelet, Haralick, and shape features for microcalcification classification in mammograms", Pattern Recognition, vol. 37, pp. 1973-1986, Oct 2004, http://dx.doi.org/10.1016/j.patcog.2003.03.001.

[33] S.-K. Lee, P.-c. Chung, C.-I. Chang, C.-S. Lo, T. Lee, G.-C. Hsu, et al., "Classification of clustered microcalcifications using a Shape Cognitron neural network", Neural Networks, vol. 16, pp. 121-132, Jan 2003, http://dx.doi.org/10.1016/S0893-6080(02)00164-8.

[34] Y. López, Novoa, Andra., Guevara, Miguel., Silva, Augusto, "Breast Cancer Diagnosis Based on a Suitable Combination of Deformable Models and Artificial Neural Networks Techniques", in Progress in Pattern Recognition, Image Analysis and Applications. vol. Volume 4756/2008, ed: Springer Berlin / Heidelberg, 2008, pp. 803-811, http://dx.doi.org/10.1007/978-3-540-76725-1_83.

[35] Y. López, Novoa, Andra., Guevara, Miguel., Quintana, Nicolás., Silva, Augusto, "Computer Aided Diagnosis System to Detect Breast Cancer Pathological Lesions", in Progress in Pattern Recognition, Image Analysis and Applications. vol. Volume 5197/2008, ed: Springer Berlin / Heidelberg, 2008, pp. 453-460, http://dx.doi.org/10.1007/978-3-540-85920-8_56.

[36] L. Talavera, "An Evaluation of Filter and Wrapper Methods for Feature Selection in Categorical Clustering", in Advances in Intelligent Data Analysis VI. vol. 3646, A. F. Famili, J. Kok, J. Peña, A. Siebes, and A. Feelders, Eds., ed: Springer Berlin Heidelberg, 2005, pp. 440-451, http://dx.doi.org/10.1007/11552253_40.

[37] M. Elter and A. Horsch, "CADx of mammographic masses and clustered microcalcifications: a review", Medical physics, vol. 36, pp. 2052-2068, 2009, http://dx.doi.org/10.1118/1.3121511.

[38] Z. Ping, B. Verma, and K. Kuldeep, "A neural-genetic algorithm for feature selection and breast abnormality classification in digital mammography", in Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on, 2004, pp. 2303-2308 vol.3, http://dx.doi.org/10.1109/IJCNN.2004.1380985.

[39] M. E. Mavroforakis, H. V. Georgiou, N. Dimitropoulos, D. Cavouras, and S. Theodoridis, "Mammographic masses characterization based on localized texture and dataset fractal analysis using linear, neural and support vector machine classifiers", Artif Intell Med, vol. 37, pp. 145-62, Jun 2006, http://dx.doi.org/10.1016/j.artmed.2006.03.002.

[40] J. C. Fu, S. K. Lee, S. T. C. Wong, J. Y. Yeh, A. H. Wang, and H. K. Wu, "Image segmentation feature selection and pattern classification for mammographic microcalcifications", Computerized Medical Imaging and Graphics, vol. 29, pp. 419-429, Sep 2005, http://dx.doi.org/10.1016/j.compmedimag.2005.03.002.

[41] J. Shi, B. Sahiner, H. P. Chan, J. Ge, L. Hadjiiski, M. A. Helvie, et al., "Characterization of mammographic masses based on level set segmentation with new image features and patient information", Med Phys, vol. 35, pp. 280-90, Jan 2008.

[42] J. L. Jesneck, J. Y. Lo, and J. A. Baker, "Breast mass lesions: computer-aided diagnosis models with mammographic and sonographic descriptors", Radiology, vol. 244, pp. 390-8, Aug 2007, http://dx.doi.org/10.1148/radiol.2442060712.

[43] D. Moura and M. Guevara López, "An evaluation of image descriptors combined with clinical data for breast cancer diagnosis", International Journal of Computer Assisted Radiology and Surgery, vol. 8, pp. 561-574, Jul 2013, http://dx.doi.org/10.1007/s11548-013-0838-2.

[44] G. I. Salama, M. Abdelhalim, and M. A.-e. Zeid, "Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers", Breast Cancer (WDBC), vol. 32, p. 2, 2012, .

[45] A. Christobel, "An Empirical Comparison of Data mining Classification Methods", International Journal of Computer Information Systems, vol. 3, 2011.

[46] F. García López, M. García Torres, B. Melián Batista, J. A. Moreno Pérez, and J. M. Moreno-Vega, "Solving feature subset selection problem by a parallel scatter search", European Journal of Operational Research, vol. 169, pp. 477-489, 2006, http://dx.doi.org/10.1016/j.ejor.2004.08.010.

[47] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: data mining, inference and prediction", The Mathematical Intelligencer, vol. 27, pp. 83-85, 2005.

[48] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update", ACM SIGKDD explorations newsletter, vol. 11, pp. 10-18, 2009, http://dx.doi.org/10.1145/1656274.1656278.

[49] "American College of Radiology (ACR) ACR BIRADS - Mammography", in ACR Breast Imaging Reporting and Data System, Breast Imaging Atlas, Reston, VA, 2003.

[50] R. M. Haralick, Shanmuga.K, and I. Dinstein, "Textural Features for Image Classification", IEEE Transactions on Systems Man and Cybernetics, vol. Smc3, pp. 610-621, 1973, http://dx.doi.org/10.1109/Tsmc.1973.4309314.