J Med Syst (2012) 36:2259–2269
DOI 10.1007/s10916-011-9693-2

ORIGINAL PAPER

Discovering Mammography-based Machine Learning Classifiers for Breast Cancer Diagnosis

Raúl Ramos-Pollán · Miguel Angel Guevara-López · Cesar Suárez-Ortega · Guillermo Díaz-Herrero · Jose Miguel Franco-Valiente · Manuel Rubio-del-Solar · Naimy González-de-Posada · Mario Augusto Pires Vaz · Joana Loureiro · Isabel Ramos

Received: 16 December 2010 / Accepted: 28 March 2011 / Published online: 9 April 2011

Abstract This work explores the design of mammography-based machine learning classifiers (MLC) and proposes a new method to build MLC for breast cancer diagnosis. We massively evaluated MLC configurations to classify features vectors extracted from segmented regions (pathological lesion or normal tissue) on *craniocaudal* (CC) and/or *mediolateral oblique* (MLO) mammography image views, providing BI-RADS diagnosis. Previously, appropriate combinations of image processing and normalization techniques were applied to reduce image artifacts and increase mammograms details. The method can be used under different data acquisition circumstances and exploits computer clusters to select well performing MLC configurations. We evaluated 286 cases extracted from the repository owned by HSJ-FMUP, where specialized radiologists segmented regions on CC and/or MLO images (biopsies provided the golden standard). Around 20,000 MLC configurations were evaluated, obtaining classifiers achieving an area under the ROC curve of 0.996 when combining features vectors extracted from CC and MLO views of the same case.

Keywords Breast cancer CAD · Machine learning classifiers · Mammography classifiers

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

R. Ramos-Pollán (✉) · C. Suárez-Ortega · G. Díaz-Herrero · J. M. Franco-Valiente · M. Rubio-del-Solar
CETA-CIEMAT Center of Extremadura for Advanced Technologies,
Calle Sola 1,
10200 Trujillo, Spain
e-mail: raul.ramos@ciemat.es

C. Suárez-Ortega
e-mail: cesar.suarez@ciemat.es

G. Díaz-Herrero
e-mail: guillermo.diaz@ciemat.es

J. M. Franco-Valiente
e-mail: josemiguel.franco@ciemat.es

M. Rubio-del-Solar
e-mail: manuel.rubio@ciemat.es

M. A. Guevara-López · N. González-de-Posada · M. A. P. Vaz
INEGI-FEUP Institute of Mechanical Engineering and Industrial Management, Faculty of Engineering, University of Porto,
Rua Dr. Roberto Frias 400,
4200–465 Porto, Portugal

M. A. Guevara-López
e-mail: mguevaral@inegi.up.pt

N. González-de-Posada
e-mail: nposada@inegi.up.pt

M. A. P. Vaz
e-mail: gmavaz@inegi.up.pt

J. Loureiro · I. Ramos
HSJ-FMUP Hospital de São João - Faculty of Medicine,
University of Porto,
Al. Prof. Hernani Monteiro,
4200–319 Porto, Portugal

J. Loureiro
e-mail: joanaploureiro@gmail.com

I. Ramos
e-mail: radiologia.hsj@mail.telepac.pt

developing countries over the past 50 years, and its incidence has increased recently. In Portugal, each year, 4,500 new cases of breast cancer are diagnosed and 1,600 women are estimated to die from this disease [2]. At present, there are no effective ways to prevent breast cancer, because its cause remains unknown. However, efficient diagnosis of breast cancer in its early stages can give a woman a better chance of full recovery. Therefore, early detection of breast cancer can play an important role in reducing the associated morbidity and mortality rates. Screening mammography is the primary imaging modality for early detection of breast cancer because it is the only method of breast imaging that consistently has been found to decrease breast cancer-related mortality. Mammography may detect cancer one and a half to four years before a cancer becomes clinically evident [3].

Double reading of mammograms (two radiologists read the same mammograms) [4] has been advocated to reduce the proportion of missed cancers. But the workload and cost associated with double reading are high. Instead, Computer-Aided Diagnosis/Detection (CAD) systems can assist one single radiologist when reading mammograms providing support to their diagnosis. These systems, which use computer technologies to detect abnormalities in mammograms (calcifications, masses, architectural distortions, etc.) used as second opinion criteria by radiologists, can play a key role in the early detection of breast cancer and help to reduce the death rate among women with breast cancer in a cost-effective manner [5]. Thus, in the past years, CAD methods and related techniques have attracted the attention of both research scientists and radiologists. For research scientists, there are several interesting research topics in cancer detection and diagnosis systems, such as high-efficiency, high-accuracy lesion detection algorithms, including the detection of calcifications and masses, architectural distortions, bilateral asymmetries, etc. [6, 7]. Radiologists, on the other hand, are paying more attention to the effectiveness of clinical applications of CAD methods.

This work proposes and evaluates a method to design mammography-based machine learning classifiers (MLC) for breast cancer diagnosis. Additionally it aimed at validating the contents of the first Portuguese "Breast Cancer Digital Repository" (BCDR) to support the method and obtain well performing classifiers.

Materials and methods

Breast cancer digital repository (BCDR)

We built the BCDR, the first Portuguese Breast Cancer database, with anonymous cases from medical historical archives (complying with current privacy regulations as they are also used to teach regular and postgraduate medical students) supplied by Hospital de São João - Faculty of Medicine at University of Porto, Portugal referred to as "*the Hospital*" in this text. BCDR is supported and hosted on the Digital Repositories Infrastructure (DRI) platform developed by the authors. Our previous work using DRI to create and host the BCDR repository can be revisited in [8–11]. DRI simplifies hosting of digital repositories (such as for medical imaging) integrating local and distributed storage resources across different locations (hospitals, university, computer centers, etc.). The BCDR data model, hosted at our DRI infrastructure, is a subset of the Digital Imaging and Communications in Medicine (DICOM) file format [12] customized by radiologists at the Hospital for storing and managing specific case information related to digital mammography images (see data model in Fig. 1). This work complements recent results in managing DICOM objects within Grid environments (see the TRENCADIS middleware [13]) by applying the DICOM standard at the Hospital and integrating it with the full machine learning classifiers development lifecycle (Fig. 1), where (1) mammography images of the BCDR are preprocessed through a graphical workstation, (2) specialized radiologists mark and classify biopsied cases which are then sotred in the BCDR, (3) data features are extracted from the stored annotations, (4) MLC configurations are massively searched and (5) selected MLC are integrated back into the workstation providing automated second opinion diagnosis to doctors. At the time of writing, BCDR includes samples of all Breast Image Reporting and Data System (BI-RADS) classes and it is composed of 950 cases, each one with the associated proven biopsy that constitutes our golden standard. Its data model supports each patient undergoing one or more studies, each study composed of one or more images (such as digitized film screen mammography images) and one or more lesions. Each image may have one or more segmentations (for different lesions) and each lesion can be associated to some segmentations, typically in mediolateral oblique (MLO) and craniocaudal (CC) images of the same breast. Moreover, each lesion can be also linked to many classifications (by different specialists, automatic classifiers, etc.). Currently, for each segmented region 18 features are automatically computed and stored forming a features vector, which is representative of the image region statistics, shape and texture. Then, the features vector is assigned to a certain class by an expert radiologist or a machine learning classifier, such as the ones based on Artificial Neural Networks (ANN) or Support Vector Machines (SVM). The BCDR model supports the possibility to assign to the same features vector several classifications by different clinicians and MLC under different class families. In this work we consider only the BI-RADS class family [14]. BCDR allows

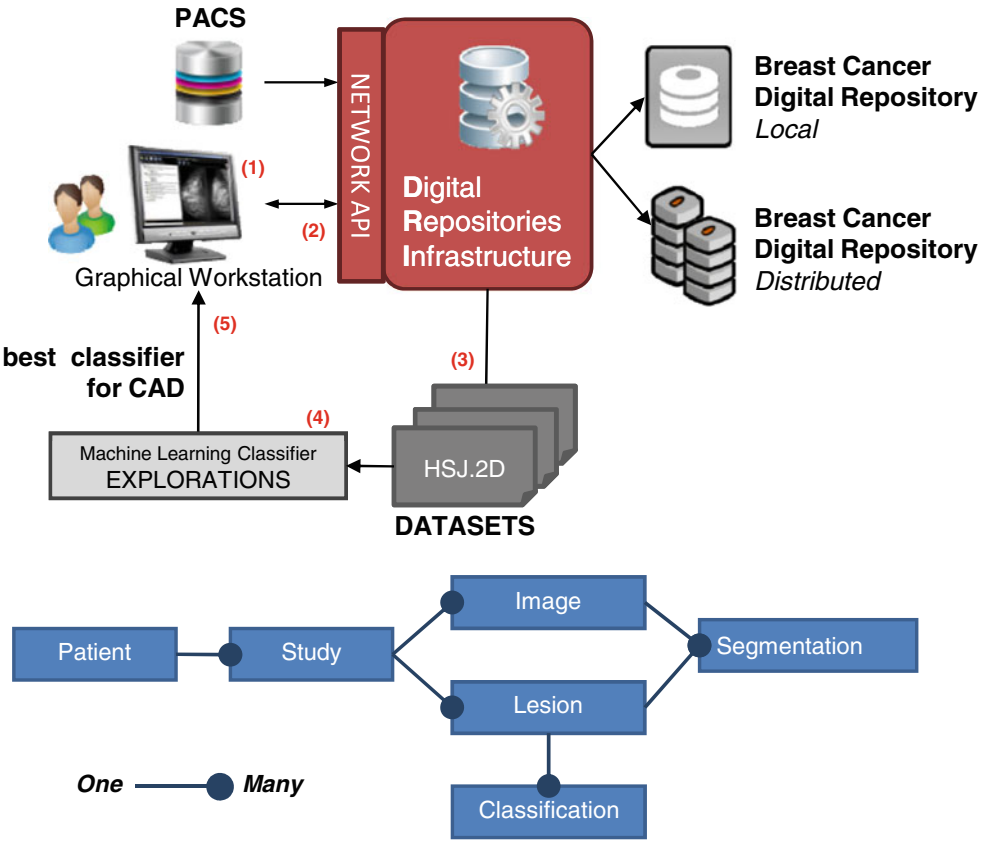

Fig. 1 BCDR DICOM-based data model and MLC Development lifecycle consisting of the following steps: (1) mammography images are preprocessed through a graphical workstation, (2) specialized radiologists mark and classify biopsied cases, (3) data is extracted from the annotations, (4) MLC configurations are massively searched and (5) selected MLC are integrated back into the workstation providing automated second opinion diagnosis to doctors

also the storage of a variety of sets of experiments of classification runs, performed both by human experts and automatic classifiers, so that later they become available for statistical analysis.

Mammograms image analysis and diagnosis

A specialized workstation system prototype (see Fig. 2) was developed providing a graphical user interface to interact with the contents of the BCDR as stored on the DRI platform. From the functional point of view, it allows the storage, retrieval and management of cases information in the BCDR. Additionally, it provides functionality to process, analyze and diagnose mammography images by implementing distinctive groups of image processing and analysis algorithms, such as classical pre-processing filters, mathematical morphological operators, image enhancing filters, thresholding, online segmentation, features extraction techniques, integration of custom made MLC, etc. The process by which a mammography image is diagnosed with the help of the workstation goes through, but not limited to, the following steps (see Fig. 2):

1. **Region of Interest (ROI) selection:** The user selects the specific image region where the lesion or abnormality is suspected to be.
2. **Image Preprocessing:** The ROI pixels are enhanced so that, in general, noise is reduced and image details are enhanced.
3. **Segmentation:** The suspected lesion is marked and separated from the rest of the ROI by identifying its contour. Segmentation is semi-automatic, where the user segments the region assisted by the computer through an interactive technique based on deformable models (snakes, active shape model, etc.) [15] and intelligent scissors (livewire) [16].
4. **Features Extraction:** Quantitative measures (features) of different nature are extracted out from the segmented region to produce a features vector which is representative of the segmented region. The graphical workstation currently extracts 18 features including statistics *(skewness, kurtosis, perimeter, area, standard deviation, minimum, maximum, mode and mean)*, shape *(elongation, roughness, form, circularity)* and texture *(correlation, angular second moment, contrast, inverse difference moment, entropy)* [6, 17, 18].
5. **Classification:** The features vector is assigned to a certain class, corresponding to a lesion type and/or a benignancy/malignancy status. The developed workstation supports this process in two modes:

   a) *Dataset construction mode:* The user performs steps 1 to 4, and then provides a classification

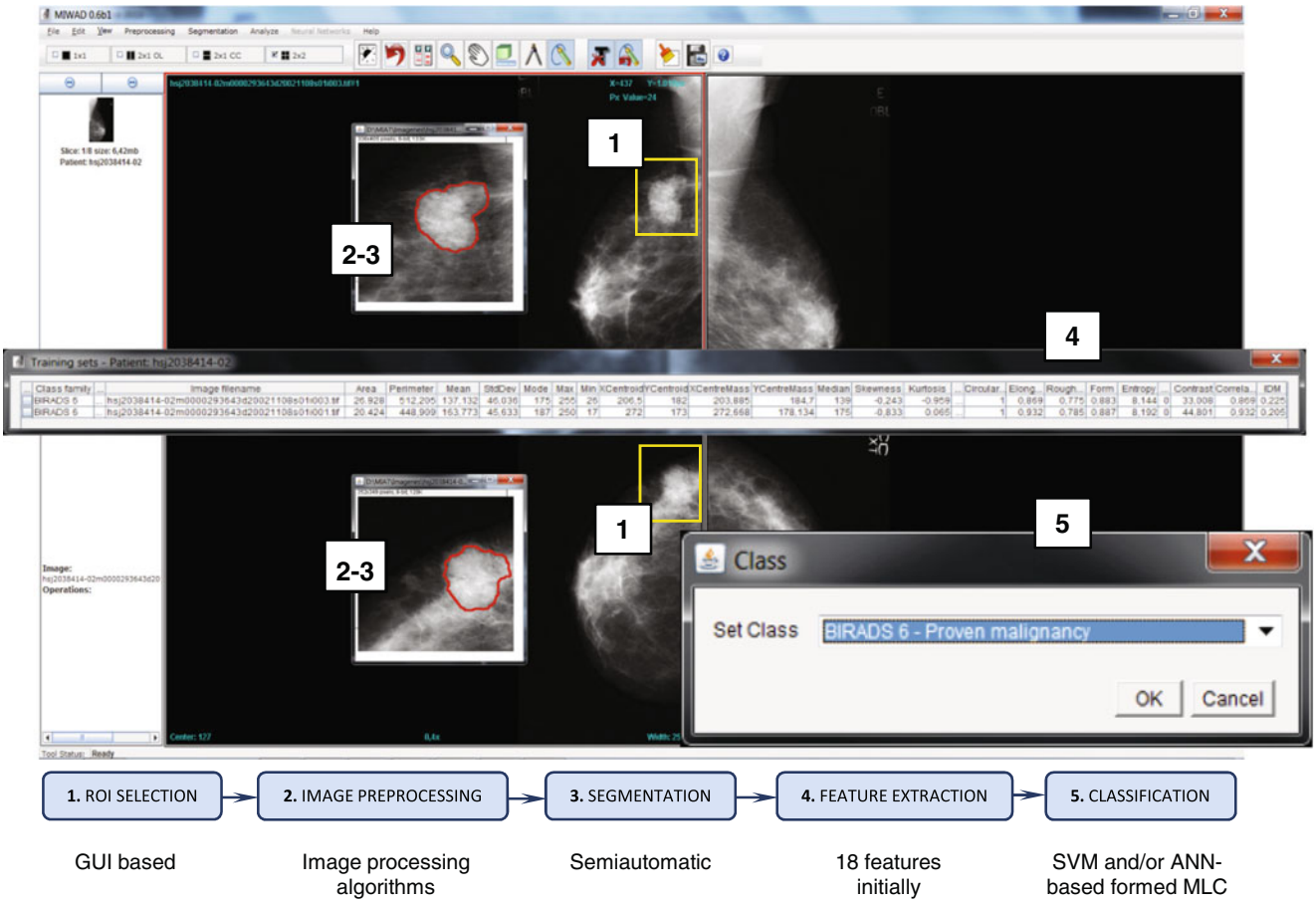

Fig. 2 Mammography image graphical workstation and functionality workflow. Observe how the same lesion is segmented on both the MLO and CC mammogram image views

based on his own knowledge and expertise. In our experiment, specialized radiologists use all available information to provide a BI-RADS classification including, most importantly, the biopsies which provide our golden standard. With this, we build datasets composed of features vectors classified by specialists and use them to train MLC.

b) *CAD mode:* The user performs steps 1 through 4, and then he uses a previously trained MLC to obtain an automatic classification of the extracted features vector to use it as second opinion in his diagnosis and patient management decisions.

Dataset preparation

From BCDR, two specialized radiologists at the Hospital used the workstation to evaluate and BI-RADS classify 286 cases. Only cases having both CC and MLO mammography images of left and right breasts were selected, including associated critical information such as lesion type, biopsies results, etc. A few cases were normal ones (no lesion) and the rest showed one of the following lesions: microcalcifications, calcifications, masses, architectural distortions and asymmetries. Several image processing operations were applied and validated on all selected images to improve ROIs details. The goal was to find fast and simple image preprocessing operations for denoising and enhancing possible pathological lesions or normal tissue image regions. This validation included suitable combinations of preprocessing filters, mathematic morphology, thresholding and edge detection among others techniques. However, the most common defect of mammography images was the poor contrast resulting from a reduced, and perhaps nonlinear, image amplitude range. Then, we found that in a first preprocessing step, ROI details can be in general improved by adjusting image intensities (a conventional contrast enhancement technique based on amplitude rescaling of each pixel). To enhance images contrast we mapped gray scale intensity values of input CC and MLO mammography images to new values such that 1% of data is saturated at low and high intensities to produce a new image in which the contrast is increased.

Special effort was made by specialists trying to locate possible ROIs and segment the same lesion in both CC and MLO associated mammography images for each case. This double-segmentation was successfully performed in 126 cases producing each one two features vectors (one for each CC and MLO image), whereas in the remaining 160 cases only one ROI was segmented, either in the CC or in the MLO image, producing one single features vector. This was attributable to various reasons, including technical issues, difficulties in ROI identification in both CC and MLO images or casual contingencies. For each segmentation (in MLO and/or CC images) a features vectors was extracted formed by 18 features including shape, texture and statistical features as described above.

From this raw data three primary datasets were constructed, as shown in Fig. 3. Dataset HSJ.2D holds all 412 features vectors extracted from the 286 cases, where the 126 double-segmentation cases produced two vectors each (252 vectors) and the 160 single-segmentation cases produced one vector each. Dataset HSJ.3DSINGLE contains only the 252 vectors produced by the double-segmentation cases. Finally, for each features vector pair from double-segmentation cases, we formed a single vector joining the 18 features segmented from the MLO image and the 18 features segmented from the CC image. This resulted in the HSJ.3DJOIN dataset, containing 126 vectors with 36 features each. With this, we aimed at understanding (as has been reported [19]) if relating two mammograms views (MLO and CC) segmentations of the same lesion could be exploited to gain classification accuracy.

The distribution of BI-RADS classes on the HSJ.2D dataset was as follows: 8 features vectors classified in class 1, 172 in class 2, 75 in class 3, 55 in class 4, 26 in class 5 and 76 features vectors classified in class 6. However, as some BI-RADS classes were rather scarce, classes were grouped as benign (BI-RADS 1, 2 and 3) and malign (BI-RADS 4, 5 and 6) making all datasets binary and producing the following distribution on each dataset: HSJ.2D 255 benign/157 malign, HSJ.3DSINGLE 182 benign/70 malign, HSJ.3DJOIN 91 benign/35 malign. This BI-RADS distribution included normal cases and all lesions mentioned above (microcalcifications, calcifications, masses, architectural distortions and asymmetries).

Dataset normalization is required before feeding data to any MLC, since different features of the same vector usually take values over ranges of different sizes and nature. This affects MLC performance and in this study each of the three datasets just described was normalized using three different techniques, seeking to understand how to better preprocess them, producing a total of nine datasets to explore as shown in Fig. 3. The three normalization procedures used were **euclidian, range to [0,1]** and **principal component analysis**.

Euclidean normalization was calculated by $V' = V/||V||$where $V = (v_1, v_2, \ldots, v_n)$ represents the original features vector and $V' = (v_1', v_2', \ldots, v_n')$ the normalized resulting vector. $||V||$ is the vector norm defined as $||V|| = \sqrt{V \cdot V} = \sqrt{v_1^2 + \cdots + v_n^2}$

Range normalization [0,1] processes individually each vector's feature v_i to guarantee they all fall within the [0,1] interval and it was calculated as $v_i' = \frac{v_i}{Max_i - Min_i}$ where v_i and v_i' are the original and normalized feature values respectively and Min_i and Max_i are the minimum and maximum values of feature i in all elements of the dataset.

Principal Component Analysis (PCA) was done using the WEKA toolkit [20], reducing the dimensionality of each dataset to account for 99% of its variability.

Summarizing, starting with three base datasets: HSJ.2D (412 features vectors, 18 features), HSJ.3DSINGLE (252 features vectors, 18 features) and HSJ.3DJOIN (126

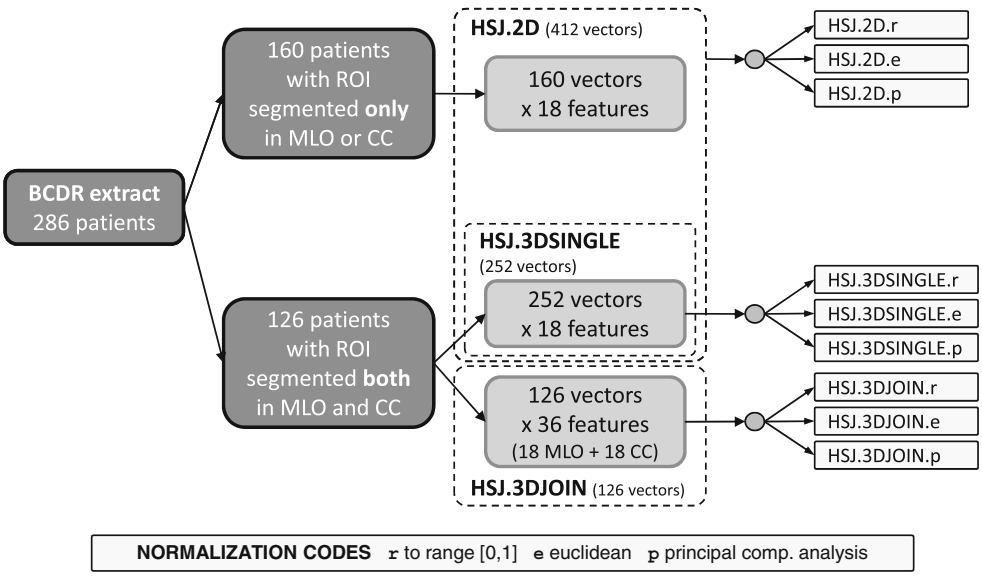

Fig. 3 Datasets built for MLC exploration

features vectors, 36 features, 18 from each original vector), we created 9 working datasets that were produced after normalizing each dataset with **range [0,1]**, **euclidian** and **PCA** normalization procedures (named HSJ.2D.r, HSJ.2D.e, HSJ.2D.p, HSJ.3DSINGLE.r, etc.).

Machine learning classifiers explorations

Machine Learning Classifier (MLC) approaches have been reported in the past few years for mammography images analysis and classification with different degrees of success [6, 7, 19, 21–30] (see discussion in Section 3). In this work we consider SVM and ANN based MLC. There are different kinds of ANNs and SVMs and, in addition to the choice of ANN network structure (layers and neurons) or SVM kernel each one can be tuned by a number of parameters. We consider an MLC configuration as a certain combination of ANN/SVM type, structure arrangement and parameters. In general, MLC design amounts to choosing well performing configurations within the "MLC search space" composed by all possible MLC configurations [31]. MLC search spaces are vast and highly dimensional and their exploration remains mostly a heuristic task, largely dependent on the experience of the designer, the nature of the classification task in hand and the availability of computing resources. Training a single MLC configuration is already computationally expensive (taking from a few seconds to several hours) and therefore exploring regions of the MLC search space is far more computationally demanding.

In order to train, validate and integrate MLC into our graphical workstation for computer assisted second opinion diagnosis, we performed a massive exploration of MLC configurations using ANN and SVM from the Encog [32] and libsvm [33] publicly available toolkits. The exploration was done over the public Grid computer cluster [34, 35] at CETA-CIEMAT consisting of 50 CPU cores. Feedforward ANNs as provided by Encog can be trained with back-propagation (denoted by *ffbp*), simulated annealing (*ffsa*) and genetic algorithms (*ffga*), see [32] for details. In addition, the authors extended Encog to support ROC optimization based backpropagation (*ffbproc*) and simulated annealing (*ffsaroc*) [36]. From libsvm we use cost based optimization SVMs (denoted by *csvm*). Each MLC requires several configuration parameters to be set depending on the underlying engine used (such as *learning-rate* for *ffbp*, *start-temperature* for *ffsa*, *population-size* for *ffga*, *kernel-type* for *csvm*, etc.). An exploration, made of many of MLC configurations, is therefore composed of many engine and parameter combinations and conforms a (usually small) region of the MLC search space.

In addition, classifier evaluation was made through classical accuracy measures (percentage of dataset elements correctly classified) and through plotting ROC (Receiver Operating Characteristic) curves and computing their area under the ROC curve (ROC Az) [37, 38], using the bi-normal distribution method as provided by JLABROC4 [39] and the Mann–Whitney statistic provided by WEKA [20]. As described, the MLC development lifecycle uses the datasets created with the graphical workstation and stored in the database as inputs to the MLC exploration process which makes extensive use of computer clusters to evaluate each MLC configuration. Selected MLC are then integrated back into the graphical workstation for automatic second opinion diagnosis, closing the cycle.

The datasets described above were used for exploring SVM and ANN based classifiers search spaces with the goals of finding well performing MLC configurations for each dataset and understanding what normalization procedure helps us obtain better MLC classifiers. Both for SVMs and ANNs the strategy was first to make general explorations with a wide range of parameters and then performing more fine grained explorations around the MLC configurations yielding better classification performance. For validation, a variant of the bootstrap method [40, 41] was used, where before training each MLC configuration, 40% of the dataset was labeled randomly for testing and the rest for training (referred to in our experiments as *testpct40*) preserving the class ratio. This process was repeated 5 times for each configuration to allow statistical smoothing. In addition, for SVMs we also used leave-one-out validation (see [42], denoted in our experiments as *one2all*) where a dataset of size n is trained n times, each one with n-1 elements used for training, the one left out for testing and averaging the results.

ANNs are much more computationally expensive to train than SVMs, depending mostly on the number of iterations (epochs) and dataset sizes. This makes ANNs unfeasible to be trained with *one2all* validation and we only used *testpct40* with all available Encog engines (ffbp, ffbproc, ffsa, ffsaroc, ffga), following the same strategy for each engine. First, an exploration over a wide range of parameter values was made, using a reduced number of iterations. Then, additional explorations with increased iterations were made around the best configurations or those showing room for convergence (where error rates did not stall).

Well performing classifiers resulting from exploring the 9 working datasets would become the targets to be integrated into CAD systems, with the corresponding image and data preprocessing. In total, for each dataset we setup around 4,350 MLC configurations:

- Six hundred SVM configurations with one2all validation (480 configurations for a general exploration and the rest for finer ones).

- Three thousand SVM configurations with testpct 40 validation (the same 600 configurations as with one2all validation, trained 5 times each one with testpct 40).
- 150×5 ANN configurations with testpct 40 validation . This is, 30 configurations for each of the 5 ANN engines, trained 5 times each one with testpct 40 (20 configurations for a general exploration and 10 configurations for the finer ones).

Datasets were explored in the following order:

1. **HSJ.2D**: Normalized datasets including all (412) features vectors and all (18) features (HSJ.2D.r, HSJ.2D.e and HSJ.2D.p).
2. **HSJ.3DSINGLE.p**: Dataset including only 256 selected features vectors and all (18) features with **PCA** normalization. Only the PCA normalized dataset was explored since, in the previous step with dataset HSJ.2D, we observed that both PCA and range [0,1] normalization gave satisfactory results, whereas Euclidean normalization behaved consistently worse.
3. **HSJ.3DJOIN.p**: Dataset including only 126 selected features vectors and 36 features with **PCA** normalization. Only the PCA normalized dataset was explored for the same reason as above.

In total 5 out of the 9 original datasets were explored, training 21,750 MLC configurations. This took around 200 days of CPU time, requiring over four physical days on the public Grid computer cluster at CETA-CIEMAT with 50 CPU cores. After training and validation, each MLC configuration produced two measures for classifier performance, on the test part of the dataset: **TESTPCT**, the percentage of elements correctly classified (referred to as accuracy in most of the literature) and **TESTAZ,** the area under the ROC curve (ROC Az).

All data was gathered from anonymous cases in medical historical archives at the Hospital and processed complying with current privacy regulations as they are also used to teach regular and postgraduate medical students.

Results and discussion

Table 1 summarizes the explorations on HSJ.2D datasets and Table 2 details the best MLC obtained on the HSJ.3DSINGLE.p and HSJ.3DJOIN.p datasets. ROC curves for best TESTAZs marked in gray in both tables are plotted in Fig. 4. It is important to remark that, although for simplicity results from *one2all* and *testpct40* validation methods are shown together, their comparative interpretation must be undertaken with care. In *one2all* a dataset of size n is trained n times, each one labeling n-1 elements for

Table 1 Summary of explorations over HSJ.2D datasets for SVM configurations (with testpct40 and one2all validation) and ANN configurations (only with testpct40 validation). For ANNs, only results for datasets normalized with Principal Component Analysis and Range-[0,1] normalization are shown, since Eucludian normalization produced worse results in all cases

Dataset	Engine	Configs trained	MAX TESTPCT	AVG TESTPCT	STDDEV TESTPCT	MAX TESTAZ	AVG TESTAZ	STDDEV TESTAZ
Support vector machines (with testpct40 validation)								
HSJ.2D.r	libsvm.csvc	600×5	0,750	0,654	0,031	0,778	0,660	0,053
HSJ.2D.p	libsvm.csvc	600×5	0,733	0,655	0,030	0,770	0,664	0,055
HSJ.2D.e	libsvm.csvc	600×5	0,709	0,613	0,021	0,733	0,511	0,104
Support vector machines (with one2all validation)								
HSJ.2D.p	libsvm.csvc	600	0,689	0,667	0,010	0,683	0,656	0,011
HSJ.2D.r	libsvm.csvc	600	0,687	0,662	0,013	0,677	0,654	0,010
HSJ.2D.e	libsvm.csvc	600	0,653	0,621	0,018	0,632	0,498	0,063
Artificial neural networks (with testpct40 validation)								
HSJ.2D.p	encog.ffsa	30×5	0,709	0,656	0,025	0,788	0,704	0,044
HSJ.2D.r	encog.ffsa	30×5	0,733	0,659	0,046	0,771	0,700	0,043
HSJ.2D.p	encog.ffbpiroc	30×5	0,721	0,640	0,054	0,758	0,679	0,032
HSJ.2D.r	encog.ffsaroc	30×5	0,726	0,650	0,038	0,752	0,683	0,044
HSJ.2D.p	encog.ffbp	30×5	0,709	0,635	0,039	0,744	0,662	0,045
HSJ.2D.p	encog.ffga	30×5	0,709	0,618	0,052	0,737	0,622	0,060
HSJ.2D.r	encog.ffbpiroc	30×5	0,673	0,631	0,021	0,735	0,669	0,038
HSJ.2D.p	encog.ffsaroc	30×5	0,695	0,637	0,036	0,733	0,675	0,039
HSJ.2D.r	encog.ffbp	30×5	0,691	0,635	0,028	0,726	0,659	0,043
HSJ.2D.r	encog.ffga	30×5	0,701	0,620	0,034	0,665	0,595	0,030

Table 2 Best MLC obtained for HSJ.3DSINGLE.p and HSJ3DJOIN.p datasets (with Principal Component Analysis normalization)

Engine	Configuration parameters	Configuration values	VALDTN	TESTPCT	TESTAZ
HSJ.3DSINGLE.p					
libsvm.csvc	kernel degree gamma shrink coef0 cost weight probestimates	pol 2 0.0048 true 1.0 64.0 0.5 true	one2all	0,917	0,953
libsvm.csvc	kernel degree gamma shrink coef0 cost weight probestimates	pol 2 0.0010 true 0.6 512.0 1.0 true	one2all	0,933	0,950
libsvm.csvc	kernel degree gamma shrink coef0 cost weight probestimates	sigm 2 0.0010 true 0.6 512.0 1.0 true	testpct 40	0,901	0,949
libsvm.csvc	kernel degree gamma shrink coef0 cost weight probestimates	pol 2 0.0010 true 0.1 1.0 0.1 true	testpct 40	0,885	0,946
encog.ffbp	layers and neurons learnrate momentum epochs	[18:27:14:7:2] 0.1 0.2 500	testpct 40	0,950	0,992
encog.ffsaroc	layers and neurons starttempendtemp cycles epochs	[18:27:14:7:2] 100.0 2.0 100 200	testpct 40	0,920	0,966
encog.ffbpiroc	layers and neurons learnrate momentum epochs	[18:27:14:7:2] 0.1 0.2 500	testpct 40	0,931	0,956
encog.ffsa	layers and neurons starttempendtemp cycles epochs	[18:27:14:7:2] 100.0 2.0 100 200	testpct 40	0,911	0,951
HSJ.3DJOIN.p					
libsvm.csvc	kernel degree gamma shrink coef0 cost weight probestimates	pol 2 0.0048 true 1.0 64.0 0.5 true	one2all	0,937	0,984
libsvm.csvc	kernel degree gamma shrink coef0 cost weight probestimates	pol 2 0.0010 true 0.1 1.0 0.1 true	testpct 40	0,937	0,982
libsvm.csvc	kernel degree gamma shrink coef0 cost weight probestimates	sigm 2 0.0010 true 0.1 1.0 0.1 true	one2all	0,921	0,981
libsvm.csvc	kernel degree gamma shrink coef0 cost weight probestimates	rbf 2 0.01 true 0.6 1.0 1.0 true	testpct 40	0,913	0,981
encog.ffbp	layers and neurons learnrate momentum epochs	[36:54:27:14:2] 0.1 0.2 500	testpct 40	0,941	0,996
encog.ffsa	layers and neurons starttempendtemp cycles epochs	[36:54:27:14:2] 100.0 2.0 100 200	testpct 40	0,940	0,983
encog.ffbpiroc	layers and neurons learnrate momentum epochs	[36:54:27:14:2] 0.1 0.2 500	testpct 40	0,902	0,964
encog.ffsaroc	layers and neurons starttempendtemp cycles epochs	[36:54:27:14:2] 100.0 2.0 100 200	testpct 40	0,902	0,960

training and one for testing. Each time, the accuracy of the test part of the dataset (only one element) is either 0% or 100%, but then it is averaged over all elements of the dataset once the n training processes are completed. With this, TESTPCT and TESTAZ measure classifier performance on the whole dataset (since each element is used for testing once). In particular, TESTPCT represents more a proper probability rather than an averaged classifier score. On the other hand, with *testpct40* classifier performance (TESTPCT and TESTAZ) refers only to the 40% selected as test instances. Moreover, it is more subject to outliers, since random selection of the test elements may eventually favor those easier to classify.

Classification performance of HSJ.2D datasets Table 1 summarizes the performance of obtained classifiers for the HSJ.2D datasets per classifier engine and validation type. Each line summarizes the number of MLC configurations trained showing the maximum, the average and the standard deviation for TESTAZ and TESTPCT. SVM-based classifiers using the *one2all* validation method yielded a maximum TESTAZ of 0,683 and 0,778 with *testpc40*. For ANN engines, only results with range [0,1] and PCA normalization are shown, since Euclidean normalization always produced lower scores. For ANNs, the best TESTAZ was 0,788. These results are marked in gray in Table 1 and their ROC curves are plotted in Fig. 4 (left). This allows us to consider that: (1) range [0,1] and PCA normalization are more suitable for HSJ.2D than Euclidean normalization (2) there is a non negligible difference between SVM results using *one2all* and *testpct40* validation, recalling the remarks made above; and (3) similar classification performance results were obtained in SVM (with *one2all*) and ANN (with *testpct40*) based classifiers.

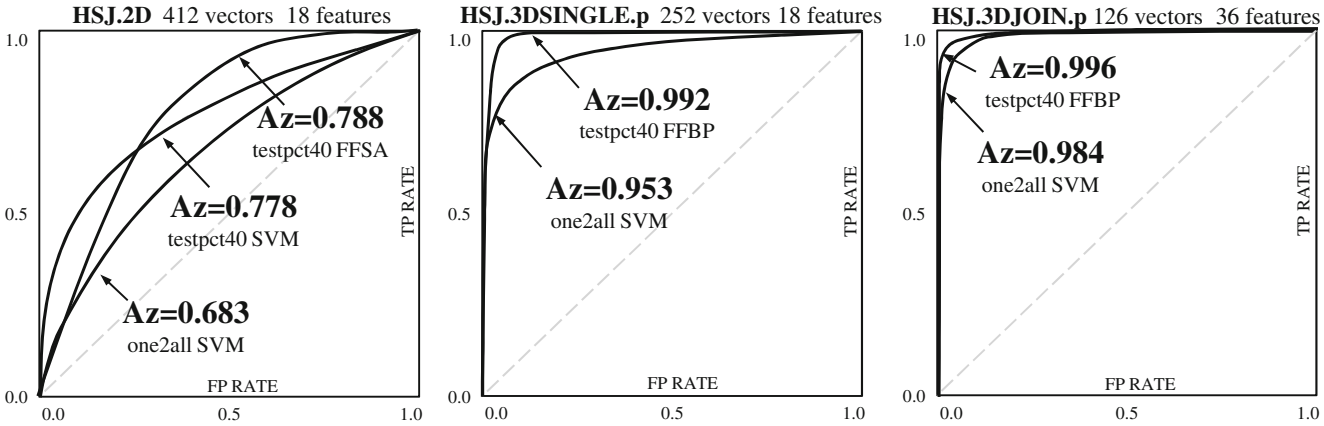

Fig. 4 ROC curves for best MLC obtained on each dataset and validation type

Classification performance of HSJ.3DSINGLE.p and HSJ.3DJOIN.p datasets Based on these results we determined exploring only the HSJ.3DSINGLE.p and HSJ.3DJOIN.p datasets, normalized by PCA and using both SVM and ANN based classifiers. Table 2 details the best classifiers respectively obtained for the HSJ.3DSINGLE.p and HSJ.3DJOIN.p datasets with SVM and ANN configurations, including the classifier parameters used. A significant increase of TESTPCT and TESTAZ values in classifiers for both datasets can be observed with respect to all HSJ.2D datasets. The highest TESTAZ values (0,996 in HSJ.3DJOIN.p and 0,992 in HSJ.3DSINGLE.p) were produced by ANN-based classifiers (with *testpct40* validation). However, SVM-based classifiers also produced high TESTAZ values (0,984 in HSJ.3DJOIN.p and 0,953 in HSJ.3DSINGLE.p). Figure 4 (center and right plots) shows the ROC curves corresponding to these TESTAZ values.

Validation method Results in [42] indicate that *one2all* (leave-one-out) validation gives a nearly unbiased estimator for classifier accuracy, but often with high variability, specially in small datasets. Our experiments showed little variability on all classifiers (both on accuracy and ROC Az), which encourages us to place stronger confidence in our results, at the expense of additional computing power required to train classifiers. Interestingly enough, results of *testpct40* and *one2all* validation differ very little in the HSJ.3DSINGLE.p and HSJ.3DJOIN.p datasets, which might suggest that *testpct40* could be used for these datasets instead of *one2all*, reducing significantly computing time. The fact that this is not exactly the case in HSJ.2D datasets, but still *one2all* and *testpct40* results are quite close, suggests that using a different percentage in *testpct* might lead to similar results, reducing as well the computing requirements to explore MLC for those datasets. This needs to be further explored, specially for biomedical datasets, since more data is being generated as our project at the Hospital continues, and statistically consistent results at reduced computing costs will be key to allow us place increasingly stronger confidence on the automatic diagnoses made by our system.

MLC-based detection/diagnosis methods for supporting semi-automated or automated breast cancer CAD systems have been developed with minor or major degree of success in the last two decades, seeking to modify the habitually qualitative diagnostic criteria into a more objective quantitative feature classification task. Results presented here are comparable and in some cases superior to prior reported approaches including: a naïve Bayes classification method [25] that can distinguish between the diffraction patterns of normal (20 instances) and cancerous tissue (22 instances) using as input a dataset of computed features vectors from X-ray mammography scatter images (above 90% accuracy); a comparative study of logistic regression and ANN-based classifiers [26] using as input a dataset of features vectors extracted from ultrasound images of 24 malignant and 30 benign masses (ANN-based classifier with 0.856 ROC Az, 95% sensitivity and 76.5% specificity); a supervised fuzzy clustering classification technique [27] validated with the Wisconsin breast cancer dataset (WBCD) composed by 699 features vectors to distinguish between benign and malignant cancers (95.57% accuracy); a method for rule extraction from ANN [28] validated on the WBCD (98.1% accuracy); an hybrid model integrating a case-based data clustering method and a fuzzy decision tree [29] validated on the WBCD (98.4% accuracy); a comparative study of different SVM training methods [30] that integrated: particle swarm optimization, quantum particle swarm optimization, quadratic programming, and the least square SVM method tested on the WBCD (93.52% accuracy). Interpretation of this comparison must be undertaken with care, accounting for the different datasets and experimental conditions with which the

different results are obtained across the literature sources. In order to provide a more objective ground, an experimental exploration to evaluate the method herewith proposed was performed on the WBCD, starting from the *dataset normalization* stage as described in previous section. We obtained 96.91% accuracy and 0.924 ROC Az with SVM based MLC and 97.14% accuracy and 0.933 ROC Az with ANN based MLC. This exploration consisted on 280 configurations requiring an additional 700 CPU hours.

Conclusions

In this paper, we presented a first evaluation of a method to design mammography-based machine learning classifiers (MLC) for breast cancer diagnosis, allowing to characterize breast lesions according to BI-RADS classes (grouped by benign and malignant). The results of our investigation confirm that: (1) BCDR contains critical information to create robust datasets of features vectors allowing massive exploration of classifiers search spaces, including the biopsies constituting the golden standard against which classifier performance is evaluated; (2) the *testpct40* validation method, or possibly some variation, could be considered to be used when *one2all* might not be computationally feasible, enlarging the exploration possibilities of MLC without reducing their statistical consistency; (3) using features vectors representing the same identified ROI (lesion) in both CC and MLO mammography images increased meaningfully classifiers performance; and (4) slightly better results obtained with the HSJ.3DJOIN dataset reinforce the idea that associating corresponding CC and MLO segmentations within the same features vectors might be useful. These results are in agreement with reported findings demonstrating that using only features vectors representing the same ROI (identified lesion) in both (two) views (CC and MLO) mammography images in the training step can be improved meaningfully classifiers accuracy and ROC Az.

Overall, the proposed method proved to be an appropriate framework for supporting the full lifecycle to build mammography-based MLC, which can be reproduced in other medical environments with reasonable cost.

Acknowledgements This work is part of the GRIDMED research collaboration project between INEGI (Portugal) and CETA-CIEMAT (Spain). Prof. Guevara acknowledges POPH - QREN-Tipologia 4.2 – Promotion of scientific employment funded by the ESF and MCTES, Portugal. CETA-CIEMAT acknowledges the support of the European Regional Development Fund

References

1. Althuis, M. D., et al., Global trends in breast cancer incidence and mortality 1973–1997. *Int. J. Epidemiol.* 34:405–412, 2005. April 1, 2005.
2. Veloso, V., "Cancro da mama mata 5 mulheres por dia em Portugal,". In: (Ed.) *CiênciaHoje.* Lisboa, Portugal, 2009
3. Tabár, L., et al., Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality. *Cancer* 91:1724–1731, 2001.
4. Brown, J., et al., Mammography screening: an incremental cost effectiveness analysis of double versus single reading of mammograms, *BMJ (Clinical research ed.)* 312:809–812, 1996.
5. Sampat, M. P., et al., Computer-Aided Detection and Diagnosis in Mammography. In: Al, B. (Ed.), *Handbook of Image and Video Processing*, Secondth edition. Academic, ed Burlington, pp. 1195–1217, 2005.
6. López, Y., et al., "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.
7. López, Y., et al., "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.
8. Ramos-Pollan, R. et al., "Exploiting eInfrastructures for medical image storage and analysis: A grid application for mammography CAD,". In: *The Seventh IASTED International Conference on Biomedical Engineering*, Innsbruck, Austria, 2010
9. Ramos-Pollan, R., et al., "Grid-based architecture to host multiple repositories: A mammography image analysis use case,". In: *3rd Iberian Grid Infrastructure Conference Proceedings*, Valencia, Spain, 2009, pp. 327–338
10. Ramos-Pollan, R., et al., "Building medical image repositories and CAD systems on grid infrastructures: A mammograms case,". In: *15th edition of the Portuguese Conference on Pattern Recognition.*, University of Aveiro. Aveiro, Portugal, 2009.
11. Ramos-Pollan, R., et al., "Grid computing for breast cancer CAD. A pilot experience in a medical environment,". In: *4th Iberian Grid Infrastructure Conference*, Minho, Portugal, 2010, pp. 307–318.
12. NEMA. (2010), *Digital Imaging and Communications in Medicine.* Available: http://dicom.nema.org/
13. Espert, I. B., et al., Content-based organisation of virtual repositories of DICOM objects. *Future Gener Comput. Syst.* 25:627–637, 2009.
14. D'Orsi, C. J., et al., Breast imaging reporting and data system: ACR BI-RADS-mammography, 4th Edition ed.: American College of Radiology, 2003.
15. Chenyang, X., and Prince, J. L., Snakes, shapes, and gradient vector flow. *Image Process. IEEE Trans.* 7:359–369, 1998.
16. Liang, J., et al., United snakes. *Med. Image Anal.* 10:215–233, 2006.
17. Rodenacker, K., A feature set for cytometry on digitized microscopic images. *Cell Pathol* 25:1–36, 2001.
18. Haralick, R., et al., Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* SMC-3:610–621, 1973.
19. Oliver, A., et al., A review of automatic mass detection and segmentation in mammographic images. *Med. Image Anal.* 14:87–110, 2010.
20. Mark Hall, et al., "The WEKA data mining software: an update," *SIGKDD Explorations,* vol. 11, 2009.
21. Park, S. C., et al., Improving performance of computer-aided detection scheme by combining results from two machine learning classifiers. *Acad. Radiol.* 16:266–274, 2009.

22. Verma, B., et al., Classification of benign and malignant patterns in digital mammograms for the diagnosis of breast cancer. *Expert Syst. Appl.* 37:3344–3351, 2010.
23. Mavroforakis, M. E., et al., Mammographic masses characterization based on localized texture and dataset fractal analysis using linear, neural and support vector machine classifiers. *Artif. Intell. Med.* 37:145–162, 2006.
24. Mavroforakis, M., et al., Significance analysis of qualitative mammographic features, using linear classifiers, neural networks and support vector machines. *Eur. J. Radiol.* 54:80–89, 2005.
25. Butler, S. M., et al., A case study in feature invention for breast cancer diagnosis using X-ray scatter images. In: Gedeon, T. D., and Fung, L. C. C. (Eds.), *AI 2003: Advances in Artificial Intelligence. vol. 2903*. Springer, Berlin/Heidelberg, pp. 677–685, 2003.
26. Song, J. H., et al., Comparative analysis of logistic regression and artificial neural network for computer-aided diagnosis of breast masses. *Acad. Radiol.* 12:487–495, 2005.
27. Abonyi, J., and Szeifert, F., Supervised fuzzy clustering for the identification of fuzzy classifiers. *Pattern Recognit. Lett.* 24:2195–2207, 2003.
28. Setiono, R., Generating concise and accurate classification rules for breast cancer diagnosis. *Artif. Intell. Med.* 18:205–219, 2000.
29. Fan, C.-Y., et al., A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. *Appl. Soft Comput.* 11:632–644, 2011.
30. Sweilam, N. H., et al., Support vector machine for diagnosis cancer disease: A comparative study. *Egypt. Inform. J.* 11:81–92, 2010.
31. Bishop, C. M., *Neural Networks for Pattern Recognition*: Oxford University Press, Inc., 1995.
32. Heaton, J., "*Programming Neural Networks with Encog 2 in Java*," ed: Heaton Research, Inc., 2010.
33. Chang, C-C., and LinC.-J., (2001, *LIBSVM: a library for support vector machines*. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm
34. Foster, I, and Kesselman, C., *The Grid 2, Second Edition: Blueprint for a New Computing Infrastructure*, 2nd ed.: Elsevier, 2004.
35. *The gLite middleware*. Available: http://glite.web.cern.ch
36. Ramos Pollan, R., et al., "Introducing ROC curves as error measure functions. A new approach to train ANN-based biomedical data classifiers,". In: *15th Iberoamerican Congress on Pattern Recognition*, Sao Paolo, Brasil, 2010.
37. Yoon, H. J., et al., Evaluating computer-aided detection algorithms. *Med. Phys.* 34:2024–2038, 2007.
38. Fawcett, T., An introduction to ROC analysis. *Pattern Recognit. Lett.* 27:861–874, 2006.
39. John Eng, M. D., (2006, March 7). *ROC analysis: Web-based calculator for ROC curves*. Available: http://www.jrocfit.org
40. Kim, J.-H., Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Comput. Stat. Data Anal.* 53:3735–3745, 2009.
41. Efron, B., and Gong, G., A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *Am. Stat.* 37:36–48, 1983.
42. Efron, B., Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Am. Stat. Assoc.* 78:316–331, 1983.