Int J CARS
DOI 10.1007/s11548-013-0838-2

ORIGINAL ARTICLE

An evaluation of image descriptors combined with clinical data for breast cancer diagnosis

Daniel C. Moura · Miguel A. Guevara López

Received: 12 December 2012 / Accepted: 22 March 2013

Abstract
Purpose Breast cancer computer-aided diagnosis (CADx) may utilize image descriptors, demographics, clinical observations, or a combination. CADx performance was compared for several image features, clinical descriptors (e.g. age and radiologist's observations), and combinations of both kinds of data. A novel descriptor invariant to rotation, histograms of gradient divergence (HGD), was developed to deal with round-shaped objects, such as masses. HGD was compared with conventional CADx features.
Method HGD and 11 conventional image descriptors were evaluated using cases from two publicly available mammography data sets, the digital database for screening mammography (DDSM) and the breast cancer digital repository (BCDR), with 1,762 and 362 instances, respectively. Three experiments were done for each data set according to the type of lesion (i.e., all lesions, masses, and calcifications), resulting in six scenarios. For each scenario, 100 training and test sets were generated via resampling without replacement and five machine learning classifiers were used to assess the diagnostic performance of the descriptors.
Results Clinical descriptors outperformed image descriptors in the DDSM sample (three out of six scenarios), and combining the two kind of descriptors was advantageous in five out of six scenarios. HGD was the best descriptor (or comparable to best) in 8 out of 12 scenarios, demonstrating promising capabilities to describe masses.
Conclusions The combination of clinical data and image descriptors was advantageous in most mammography CADx scenarios. A new descriptor based on the divergence of the gradient (HGD) was demonstrated to be a feasible predictor of breast masses' diagnosis.

Keywords Breast cancer · Image descriptors · Clinical data · Machine learning classifiers · Computer-aided diagnosis (CADx) · Histograms of gradient divergence (HGD)

Introduction

According to the World Health Organization, breast cancer is the second most common form of cancer in the world, with over 1.5 million predicted diagnoses in 2010 and causing more than half a million deaths per year [1]. In the European Union, it is responsible for one in every six deaths from cancer in women [2]. Breast cancer has a known asymptomatic phase that can be detected with mammography [3], and therefore, mammography is the primary imaging modality for screening.

Double-reading (two radiologists independently read the same mammograms) has been advocated to reduce the proportion of missed cancers and it is currently included in most of the screening programs [4]. However, double-reading incurs in additional workload and costs. Alternatively, computer-aided detection/diagnosis (CADe/CADx) systems may assist a single radiologist reading mammograms

Electronic supplementary material The online version of this article (doi:10.1007/s11548-013-0838-2) contains supplementary material, which is available to authorized users.

D. C. Moura (✉) · M. A. Guevara López
Instituto de Engenharia Mecânica e Gestão Industrial, Universidade do Porto, Rua Dr Roberto Frias, 400, 4200-465 Porto, Portugal
e-mail: daniel.moura@fe.up.pt; dmoura@inegi.up.pt

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

D. C. Moura
Faculdade Engenharia, Departamento de Engenharia Informática, Universidade do Porto, Porto, Portugal

Published online: 13 April 2013

providing support to her/his decisions [5,6]. CADe systems focus on the detection of suspicious lesions, while CADx systems aim at classifying lesions identified by the radiologist. In this work, we will focus on CADx.

CADx systems typically rely on machine learning classifiers (MLC) to provide diagnosis [7,8]. In order to train a MLC for breast cancer diagnosis, a set of predictors is required describing the observation. Ideally, predictors should have high discriminant power that allows inferring if a given observation is from a malignant finding or not. This is, however, a changeling topic that has gathered the focus of research of several sciences, from medicine to computer vision. Thus, several types of predictors may be used for inferring the diagnosis. Here, we focus on two particular types of predictors: (1) clinical data: information about the patient (e.g. age, gender) and observations of radiologists about the mammograms (e.g. breast density, abnormalities); and (2) image descriptors: a set of statistics computed from the mammograms that may help characterizing lesions. Image descriptors can be further divided into two categories: (1) general, if they describe features that are transversal to the different kinds of lesions, and (2) lesion-specific, if they only make sense for a given type of lesion, such as the regularity of the contour of a mass [8], or the number of microcalcifications inside a cluster [7].

In this study, we focus on the combination of general image descriptors and clinical data. General image descriptors have the advantage of being applicable to all kinds of lesion and not requiring rigorous contours outlining the lesions, as opposed to shape descriptors that are often used to describe masses [8]. Since general image descriptors only require a region containing the lesion, they are convenient to use in clinical settings where radiologists have very limited amount of time for analyzing cases (e.g. screening). Several works have explored different kinds of general image descriptors for characterizing breast lesions. These descriptors typically describe the region of the mammogram by their distribution of grey levels or by features related to texture. In [9,10], statistics over grey levels are used to discriminate between normal and abnormal tissue. In [11,12], Zernike moments are computed to describe masses. Several works [13–19] use Haralick features [20] to classify calcifications and masses trough texture. Other texture descriptors that are often used include the grey-level run length analysis [17,18,21] and features from the grey-level difference matrix [19,21]. Multi-scale approaches have also shown promising results and include the use of Gabor filter banks [15,22], Wavelets [23–26], and Curvelets [27]. In previous work, we have also explored the combination of intensity and texture descriptors [5]. The reader is addressed to [7,8] for a comprehensive review of the area. To the best of our knowledge, descriptors based on the spatial distribution of the gradient such as histograms of oriented gradients (HOG) [28] have not been applied to breast cancer.

Despite the large amount of research on image descriptors for breast lesion classification, the main focus is typically the evaluation of performance of the descriptors in a standalone point of view, i.e., the image descriptors are the only predictors used to train MLC. While these studies are important for understanding the discriminative power of the image descriptors, we believe that it is essential to know how MLC based in these descriptors behave in the presence of relevant clinical data and if these two distinct types of data can complement each other to provide more accurate diagnosis.

This work presents a comparative study of four groups of image descriptors (intensity, texture, multi-scale texture and spatial distribution of the gradient). These descriptors are computed from rectangular regions of interest (patches) of mammographic images containing a lesion and combined with patients' clinical data to train machine learning classifiers (MLC) for breast cancer diagnosis (Fig. 1). Within the last group of descriptors, we tested the popular HOG descriptor [28] and we propose a novel descriptor that is especially designed for round-shaped objects, such as masses. The proposed descriptor, histograms of gradient divergence (HGD), enables to capture patterns of the gradient that are invariant to rotation.

The main goals of this study are to compare the performance of image descriptors in the presence and absence of clinical data, and understanding the contribution of these two types of data in the performance of MLC. In particular, we test the hypotheses that (1) combining image descriptors and clinical data enables achieving better results than the standalone clinical data or the standalone image descriptors, (2) the relative performance of the image descriptors is not necessarily the same when clinical data are also fed to the MLC, and (3) the suitability of a given descriptor dependents on the type of lesion. In addition, we propose and evaluate a new image descriptor, the HGD.

Materials and methods

This section describes the evaluation methodology of image descriptors for breast cancer diagnosis in the presence and absence of clinical features. This evaluation is done within the context of training machine learning classifiers to predict the diagnostic of a lesion based on a set of features computed from a region of the mammogram containing the lesion (Fig. 1). This section starts by describing the data sets that were utilized on the experiments, followed by a brief explanation of the image descriptors that were evaluated, and ends with the definition of the experimental study.

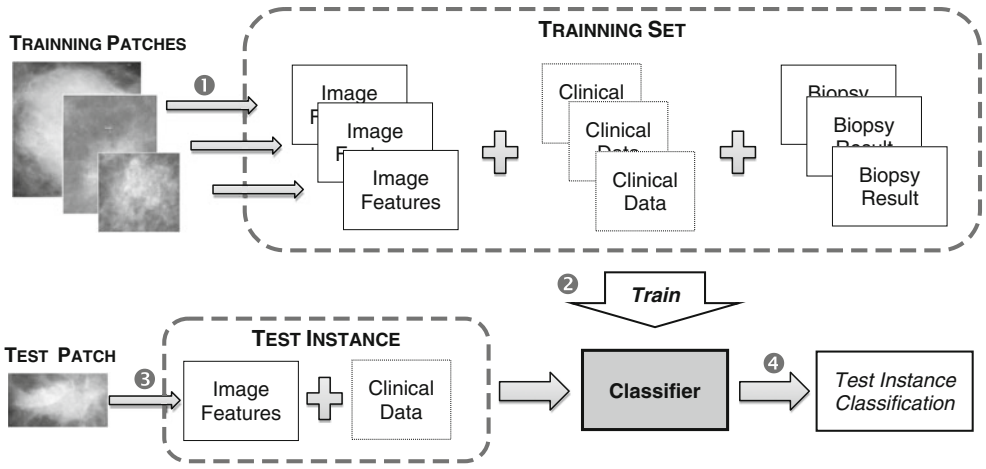

Fig. 1 Illustration of the training and classification processes. For training a classifier, first, image features are computed from a set of patches (1). Then, the computed features (optionally coupled with clinical data) together with the biopsy results of the radiographed lesions are used to train the classifier (2). Having a classifier properly trained, the diagnosis of a new lesion's patch (optionally coupled with clinical data) (steps 3 and 4)

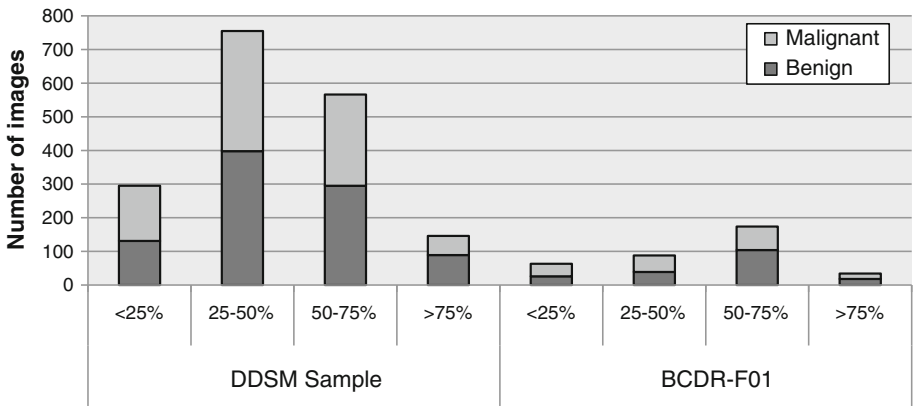

Fig. 2 Distribution of breast densities in the data sets

Data sets

Two data sets were used in this study, namely a sample of the digital database for screening mammography (DDSM) [29] and the BCDR-F01 data set of the breast cancer digital repository (BCDR) [5,30]. These public repositories[1] were selected because they provide the highest number of annotated mammograms with biopsy-proven diagnostic. Both data sets include, for each case, the age of the patient, the density of the breast (BI-RADS scale) (Fig. 2), and the contour of the lesions in one or two mammograms per breast, mediolateral oblique (MLO) and craniocaudal (CC). Both data sets were originality built from film mammograms that were scanned to produce digital images. The two data sets differ in the resolution of the images, the number of grey levels, the number of cases and in the included observations of radiologists, and therefore, they are described separately. Figure 3 shows statistics for the two data sets regarding the number of malignant and benign cases for the most common types of lesions: masses and calcifications.

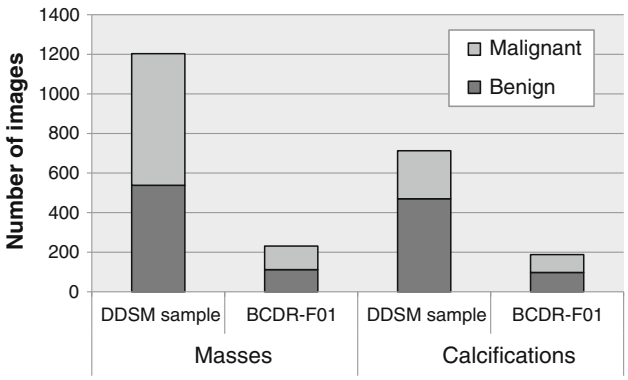

Fig. 3 Distribution of the most common abnormalities in the data sets

[1] BCDR-F01 from BCDR is now available for download at http://bcdr.inegi.up.pt.

DDSM

A total of 1,762 segmentations were extracted from DDSM from the benign volumes 1, 5, 6, and 13 and cancer volumes 1, 2, 5, 9, and 15. These volumes correspond to all the volumes of the scanner LUMISYS with the exception of volume 'Benign 14', which was left out for achieving a balanced number of benign and malignant cases. Thus, the DDSM sample for this study is composed by 913 segmentations of benign findings and 849 of malignant findings. The LUMISYS scanner is the scanner with the highest resolution of DDSM (50 microns), producing images with an average size of 3,118 × 5,001 pixels and 3,600 grey levels. For convenience, the images used in this study were obtained from the 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 PNG format [31,32]. The average patient age at the time of the study is 57.7 years old, ranging from 31 to 89. In addition to the age of the patient and density of the breast, the DDSM data set also includes the subtlety of the lesion (an integer number ranging between 1 and 5). Regarding the observations of the radiologists about the lesions, DDSM stores if there are masses or calcifications and characterizes the shape and margins of masses as well as the type and distribution of calcifications using keywords of the BI-RADS glossary. For this study, this information was encoded by creating two binary attributes indicating the presence/absence of masses and calcifications, as well as an additional binary attribute for each possible keyword. This sparse representation allows for multiple findings per lesion (e.g. a mass with calcifications) and multiple keywords characterizing an aspect of a lesion (e.g. the margins of mass can be simultaneously obscured, ill defined, and spiculated). Summarizing, the clinical data for each instance of the data set built from DDSM include a total of 35 attributes: 32 binary attributes describing the lesions, two ordinal attributes for breast density and lesion subtlety, and one numerical attribute for storing the age of the patient at the time of the study.

BCDR (BCDR-F01)

BCDR-F01 is the first data set being released to public of the Breast Cancer Digital Repository (BCDR) [5,30]. This data set is composed by cases of Portuguese female patients with mean age of 54.4 years old, ranging from 28 to 82. The mammograms of this data set were digitized with lower resolution than DDSM resulting in images of 720 × 1,167 pixels with 256 grey levels. BCDR-F01 has a total of 362 segmentations from which 187 are from benign findings and the remainder 175 from malignant findings. In addition to the patient age and breast density, the data set includes a set of selected binary attributes for indicating abnormalities observed by radiologists, namely masses, microcalcifications, calcifications (other than microcalcifications), axillary adenopathies, architectural distortions, and stroma distortions. Thus, the clinical data for each instance of the BCDR-F01 data set include a total of eight attributes per instance: six binary attributes related to observed abnormalities, an ordinal attribute for breast density, and a numerical attribute that contains the patient age at the time of the study.

Image descriptors

Intensity descriptors

Intensity statistics This descriptor is a set of statistics calculated directly over the grey levels of the pixels belonging to the patch. Previous work on detection and classification of breast cancer has used the mean value and the standard deviation as descriptors (e.g. [9,10]), but higher order statistics have been explored such as the skewness and kurtosis (e.g. [9]). Here, we include these statistics (mean, standard deviation, skewness and kurtosis), together with the minimum and maximum intensity value of the patch, making a total of six features. This combination of features was successfully explored in previous work [5].

Histogram measures Gonzales et al. [33] describe six measures based on statistical moments that are calculated from the grey-level histogram of the patch. These measures are the average intensity, contrast, smoothness, skewness, uniformity and entropy. In comparison with the previous descriptor (Intensity statistics), these measures are calculated from the histogram of the patch (instead of directly from the grey levels of the patch) and also differ by including a measure of uniformity of the histogram and a measure of randomness (entropy), while not including neither kurtosis nor the grey-level limits (minimum and maximum). This descriptor was explored in [34] for breast tissue classification and in [35] for content-based retrieval of mammograms.

Invariant moments Hu [36] proposed a set of seven features based on statistical moments that are invariant to translation, scale, and orientation of the observation. Invariant moments have been explored by [35] for content-based retrieval of mammograms. As suggested by [33], a logarithmic function was used to decrease the range of each moment.

Zernike moments Zernike moments [37] are constructed using a set of complex polynomials that describe a unitary disc (radius = 1). A descriptor of a circular patch may, thus, be defined by the coefficients of the polynomials. In contrast to statistical moments and invariant moments, Zernike moments have an orthogonal basis guaranteeing independent

coefficients. In addition, they remain invariant to translation, rotation, and scale. The first polynomial (order 0) has only one term with coefficient equal to the average pixel intensity, while higher order polynomials add detail to the description of the patch. Zernike moments have been previously used for classifying breast masses (e.g. [11,12]) and for retrieval of similar masses from a database (e.g. [38]).

Texture descriptors

Haralick features Haralick features [20] describe the texture of a patch and are computed from the grey-level co-occurrence matrix (GLCM). The GLCM is a 2D histogram measuring the joint probability of two grey levels (g_1, g_2) occurring at a given distance d and at a given direction θ. Thus, the element $GLCM(g_1, g_2)$ represents the number of times a pixel with intensity g_1 appears together a pixel with intensity g_2, with d and θ being fixed parameters. Typically, θ is 0°, 45°, 90°, or 135° and the d is a city block distance $\geq$1 pixel. Additionally, intensities may be grouped in B bins. From this matrix, a set of 14 features is computed. Haralick et al. proposed computing these features for the four directions and averaging the results in order to achieve some invariance to rotation. Several studies have included GLCM features for classifying microcalcifications (e.g. [13–16]) and masses (e.g. [15–19]).

GLRL Grey-level run length (GLRL) analysis [39] computes the occurrence of sets of consecutive collinear pixels with given length (l) and direction (θ) for a given grey level (g). Occurrences are stored in a GLRL matrix with the element $GLRL(l, g)$ representing the number of times sequences of pixels with length l is associated with the grey level g. Grey levels are grouped in B bins and GLRL matrices are computed for four directions ($\theta = 0°, 45°, 90°$, and 135°). A set of 11 features is calculated for each direction, rendering 44 features. In [21], GLRL features were utilized for classifying microcalcifications and in [17,18] for masses.

GLDM The grey-level difference matrix (GLDM) stores the occurrences of absolute differences between pairs of grey levels (Δg) separated by a given distance (d) in a given direction (θ), with the element $GLDM(d, \Delta g)$ being the number of times the grey-level difference Δg is observed at a distance d. Grey levels are grouped in B bins and GLDM matrices are computed for four directions ($\theta = 0°$, 45°, 90°, and 135°). A set of five features (mean, contrast, entropy, angular second moment, and inverse difference moment) is calculated for each matrix, rendering twenty features. GLDM features were used in [21] for classifying microcalcifications and in [19] for masses.

Multi-scale texture descriptors

Gabor filter banks Considering the spatial domain, Gabor filters can be described as a Gaussian kernel modulated by a sinusoidal plane wave [40]. These filters are often used for edge detection as they allow to detect edges in a given orientation (θ) and at a given frequency (λ). In addition, by adjusting the standard deviation of the Gaussian envelope (σ), it is possible to adjust the degree of blurring. Several approaches have been explored in the literature for applying Gabor filters to the classification of both masses and microcalcifications (e.g. [15,22]). Here, a set of descriptors was produced by calculating the mean, standard deviation, energy and entropy of the magnitude of the complex response of a set of Gabor filters with different orientations (θ), frequencies (λ), and scales (σ).

Wavelets In signal theory, a discrete wavelet transform enables to decompose a discretized signal in two sets of coefficients: approximation and detail [33]. While approximation coefficients are the result of a low-pass filter that provides a coarse approximation of the original signal, detail coefficients result from a high-pass filter that extract local variations of the signal. By repeating the discrete wavelet transform over the approximation coefficients, one is able to extract multi-scale representations of the signal. Regarding the 2D discrete wavelet transform, the decomposition of an image originates an approximation image and three detail images (horizontal, vertical, and diagonal), all with half the width and height of the original image. Iterating over the approximation images enables computing representations at multiple scales, from which features can be calculated. Computing wavelet representations requires the definition of the filters and the number of levels of decomposition (L). Wavelets have been used previously by either selecting the highest coefficients of each level (e.g. [23,24]) or by computing statistics of the coefficients of each level (e.g. [25,26]). The first option is typically employed in data sets where all the patches have the same size, and therefore, the number of coefficients remains the same for all patches. Here, the second option was chosen since it enables using the original patches of the lesions independently of their sizes. The same statistics used on Gabor filters (i.e. mean, standard deviation, energy and entropy) were applied to each sub-image that resulted from the wavelet transform.

Curvelets The curvelet transform is a higher dimensional generalization of the wavelet transform designed to represent images at different scales and different angles [41]. It was proposed to cope with some of the limitations of wavelets, and in fact, when compared to wavelets, it offers additional advantages such as optimal sparse representation of objects with edges, optimal image reconstruction in

severely ill-posed problems, and optimal sparse representation of wave propagators [42]. In [27], the authors have shown that curvelets outperform wavelets when classifying radiographed patches of the breast. Curvelets require the definition of two parameters: the number of scales and the number of angles. At the end, for each scale, several sub-images are available depending on the chosen number of angles. Here, for each sub-image, the mean, standard deviation, energy and entropy were calculated, like in the previous multi-scale textures descriptors.

Spatial distribution of the gradient

Histograms of oriented gradient Histograms of oriented gradients (HOG) describe patches through the distribution of the gradient [28]. Patches are divided into a grid of blocks (e.g. 3×3), and each block is described by a histogram of the orientation of the gradient. Each histogram has a predefined number of bins dividing the range of possible orientations (from 0 to 2π radians, or from 0 to π radians), and the value of each bin is calculated by summing the magnitude of the gradient of the pixels which have gradient direction within the limits of the bin. Finally, histograms may be normalized, with the most common option being the L1 and L2 norm [28]. HOG is a very popular descriptor in Computer Vision based on the local descriptor of the Scale Invariant Feature Transform (SIFT) [43]. It has been successfully used, for instance, in human detection and face recognition (e.g. [28,44]). However, to the best of our knowledge, it has never been used before for describing breast lesions. Here, we propose fitting an HOG to the patch of a lesion to describe the lesion. This makes the descriptor dependent on the orientation of the object. Invariance to rotation could be implemented by calculating the dominant orientation of the gradient and by rotating the rectangular patch before calculating the descriptor. However, as in [28], we choose to use patches in its original orientation because, for several cases, part of the rotated patches would fall out of the image, or bounding boxes of the rotated patches were much larger than the original bounding box. Nevertheless, HOG is expected to perform well describing masses since they are especially suited for describing shape. Both normalized and non-normalized variants are tested with different number of blocks as well as histograms with different number of bins.

Histograms of gradient divergence In this work, we propose a novel image descriptor that introduces the concept of gradient divergence to measure shape regularity invariantly of rotation. The descriptor, named histograms of gradient divergence (HGD), is based on the principle that round-shaped objects with regular and continuous border, such as a circles and ellipses, have the gradient of their boundaries pointing to the centre of the object (assuming light filled objects on a dark background). For such objects, we may say that the gradient converges to the centre.

Assuming that the object is centred on the patch, we propose measuring the gradient divergence of a pixel P as the angle between the vector of the intensity gradient on P and a vector with origin on P pointing to the centre of the patch (Fig. 4). In addition to account for divergence of the gradient, our descriptor also considers the distance of the pixel to the centre through the use of regions. For allowing compact representations, R concentric regions are created, with each region being described by a histogram of gradient divergence with B orientation bins. Invariance to rotation is naturally achieved by using circular concentric regions and by storing the divergence of the gradient instead of the orientation of the gradient.

In this work, all the regions of an HGD descriptor have equal number of pixels and do not overlap, which simplifies parameterization since only the number of regions needs to be known to determine their limits. Nevertheless, the inner and outer radius of each region could be manually specified.

To emphasize strong variations of intensity, the contribution of every divergence angle to a histogram is given by the magnitude of the gradient. At the end, and like in HOG, the histograms may be normalized. In the experiments reported here, three modalities were tested: no normalization, L2 norm, and division by the maximum bin of all histograms of the descriptor.

This descriptor is especially suited for masses and aims at describing the regularity of their shape. Two or more regions enable to capture variations of divergence at the border of masses and at the core, which allows describing regularity of the border and detecting spiculations (Fig. 4).

Evaluation

Descriptors were compared based on their classification performance using several machine learning classifiers available on Weka version 3.6 [45], namely Support Vector Machines (SVM), Random Forests (RF), Logistic Model Trees (LMT), K Nearest Neighbours (KNN), and Naive Bayes (NB). For all classifiers with the exception of NB (which is parameterless), threefold cross-validation was performed on the training set for optimizing the classifiers parameters. Linear SVM was chosen for simplicity and speed with regulation parameter C ranging from 10^{-2} to 10^{3}. The number of trees of RF was optimized between 50 and 400, with each tree having $\log_2(A) + 1$ randomly selected attributes, where A is the number of attributes available in the current data set. On LMT, the number of boosting iterations was also optimized. Finally, the number of neighbours (K) of KNN varied from 1 to 20, and the contribution of each neighbour was always weighted by the distance to the instance being classified. The WEKA configurations of the classifiers are available on

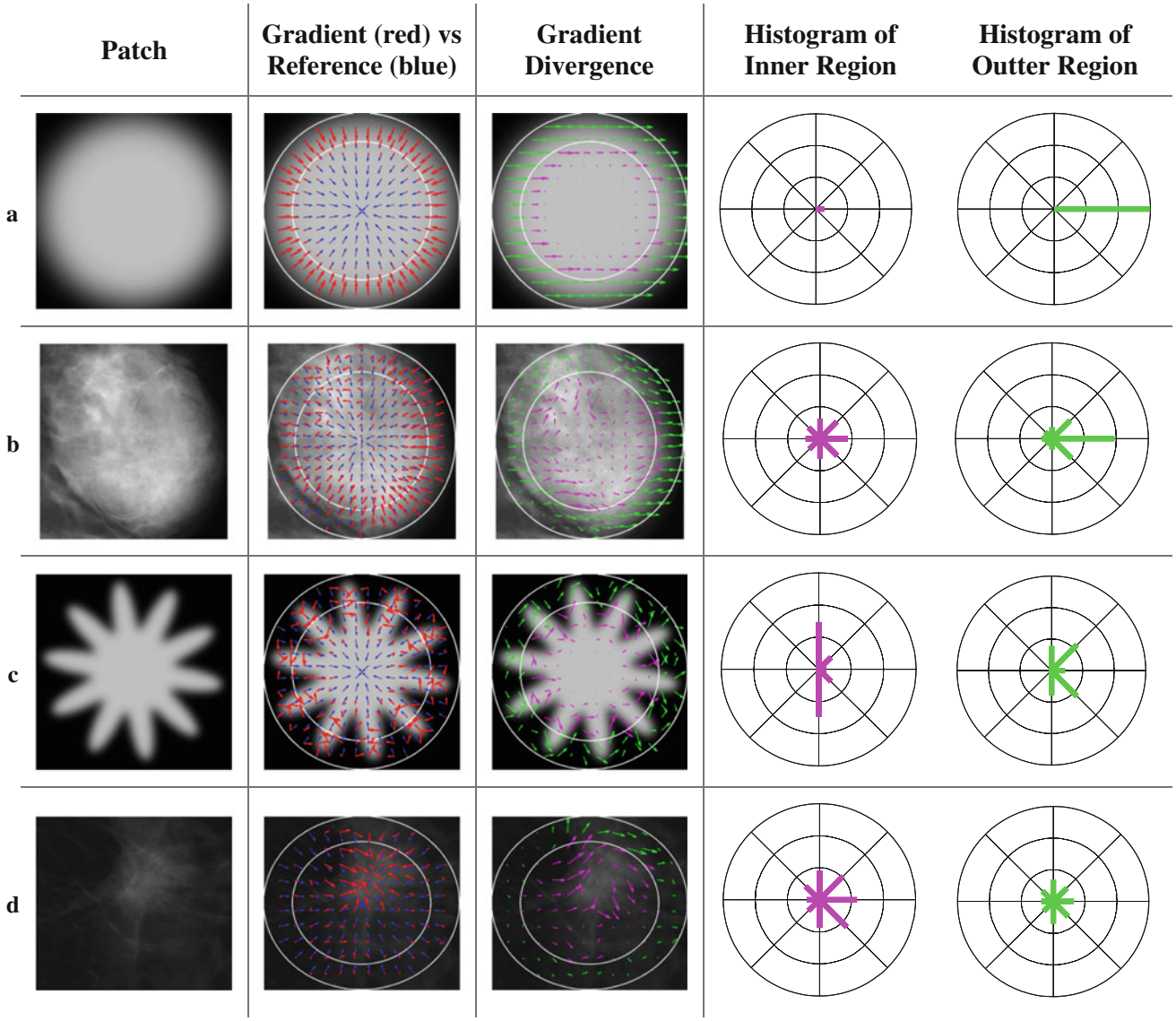

Fig. 4 Illustration of the proposed descriptor (HGD), with 2 regions and 8 bins for patches of a blurred disk (*a*); a mass with well-defined borders from BCDR-F01 (*b*); an artificially generated spiculated shape (*c*); and a spiculated mass from BCDR-F01 (*d*). The *second column* shows a sparse representation of the gradient (*red arrows*) and the reference (convergence) vectors (*blue arrows*). The *third column* shows the gradient divergence vectors, which have magnitude equal to the gradient and orientation equal to the angular difference between the gradient and reference vectors (horizontal, left to right vectors means zero divergence). The *last column* shows the histograms of the two regions of the descriptor (L2-norm normalization). Histograms have 8 bins, with the first (zero divergence) pointing to the right, and the remaining following anti-clockwise. The descriptor is formed by grouping the values of each of the 16 bins in a single vector

Online Resource 1. For all classifiers, attribute range normalization [0, 1] was performed as pre-processing with the minimum and maximum values of the attributes found in the training set and then applied to both training and test sets.

For computing the image descriptors, rectangular patches of the lesions were created by extracting the part of the mammogram within the bounding box of the outlines provided by both data sets. For all descriptors, with the exception of Gabor filters and Zernike moments, the features were computed using the patch on its original size. Due to computational requirements, the patches used to compute Gabor filters were resized so that the larger dimension would be of exactly 128 pixels, while keeping the aspect ratio. As for Zernike moments, the patch was resized to 128 $\times$ 128 also due to computational requirements and because this descriptor requires the patch to be of equal width and height.

Three scenarios were evaluated concerning the input of the classifiers: (1) standalone clinical data (i.e. 35 attributes for the DDSM data set and 8 attributes for BCDR-F01),

Table 1 Descriptors' parameters explored in the experiments

Descriptor	Parameter	Values
Zernike moments	Polynomial order	All polynomials from order 0 to 2, 4, 6, 8, 10
Haralick features	Number of grey-level bins (B)	8, 16, 32
	Distance between pixels (d)	1, 2, 4, 8, 16, 32
GLRL	Number of grey-level bins (B)	64, 128, 256
GLDM	Distance between pixels (d)	1, 2, 4, 8, 16, 32
	Number of grey-level bins (B)	64, 128. 256
Gabor filters	Standard deviation of the Gaussian (σ)	1, 2, 4, 8, 16
	Orientation (θ)	$0, \frac{\pi}{8}, \frac{\pi}{4}, \frac{3\pi}{8}, \frac{\pi}{2}, \frac{5\pi}{4}, \frac{3\pi}{4}, \frac{7\pi}{8}$
	Frequency (λ)	$2^{-2}, 2^{\frac{-5}{2}}, 2^{-3}, 2^{\frac{-7}{2}}, 2^{-4}, 2^{\frac{-9}{2}}, 2^{-5}$
Wavelets	Number of scales	1, 2, 4, 8, 16
	Wavelet filters	Haar, db8, sym8, bior3.7
Curvelets	Number of scales	2, 4, 6
	Number of angles at the 2nd coarsest level	8, 16, 32
HOG	Number of blocks (width × height)	$3 \times 3, 5 \times 5$
	Number of orientation bins	8, 16
	Normalization	None, L2-norm
HGD	Number of regions	2, 4, 8
	Number of orientation bins	8, 16, 32
	Normalization	None, L2-norm, Maximum

without computing any image descriptors from the patches, (2) standalone image descriptors, where the image descriptors are the only predictors of the classifier, and (3) the image descriptors together with clinical data. The evaluation measure was the Area Under the Curve of the Receiver Operator Characteristic (AUC). Resampling without replacing was performed 50 times for each view (MLO and CC) resulting in 100 runs per experiment to provide different splits across training and test sets, with 80 % of the cases randomly selected for training the classifier, and the remaining 20 % used for test. The two views were trained and tested independently to prevent biasing results and finally the AUCs from both views were merged resulting in a total of 100 evaluations per experiment. This experiment was done for both DDSM and BCDR-F01 data sets, for all descriptors and for all classifiers.

An additional experiment was performed where the descriptors were evaluated in two subsets of the original data sets: one with only the lesions that included masses, and another with only the lesions that contained calcifications. For compensating the subsets' unbalanced number of benign and malignant cases, instances were reweighted in the training stage according to the ratio between the less and the most represented class, guaranteeing equal contribution of each class when training the classifier.

For each descriptor, several parameter combinations were explored. Table 1 details the values tested for each parameter, which were based on the literature and on empirical tests.

When comparing descriptors, the best combination of parameters' values and classifier was used. Comparisons were based on the median AUC of the 100 runs (mAUC) and were supported by Wilcoxon signed rank tests to determine whether differences have statistical evidence ($p < 0.05$). A nonparametric test was preferred to a parametric, as suggested by [46], since nonparametric tests do not assume normal distributions or homogeneity of variance.

Results

Comparison of descriptors by type of lesion

Results are first presented for the data sets containing all types of lesions and then for masses and calcifications subsets. Table 2 shows results for the median run of all experiments. Boxplots with the results of the 100 runs are available on Online Resource 2.

All types of lesion

The standalone clinical data significantly outperformed all the standalone image descriptors on the DDSM data set ($p < 0.001$), scoring mAUC = 0.853. The best standalone image descriptor was GLRL (mAUC = 0.743) with no

Table 2 Classification performance (median AUC) of the standalone clinical data and of the image descriptors (standalone and combined with clinical data)

Data set	Standalone clinical data	Combined with clinical data	Image descriptors											
			IS	HM	IM	Zer	Har	GLRL	GLDM	Gab	Wav	Curv	HOG	HGD
All Lesions														
DDSM sample	0.853	No	0.715	0.691	0.667	0.691	0.736	**0.743**	0.683	0.725	0.731	0.712	0.729	0.736
		Yes	**0.868**	0.860	0.859	0.864	0.857	0.862	0.845	0.854	0.860	0.865	0.848	0.851
BCDR F01	0.712	No	0.637	0.614	0.691	0.648	0.710	0.654	0.641	0.712	0.719	0.705	0.739	**0.825**
		Yes	0.766	0.765	0.770	0.754	0.784	0.713	0.743	0.788	0.776	0.781	0.765	**0.817**
Masses														
DDSM sample	0.867	No	0.707	0.667	0.647	0.675	0.718	**0.733**	0.683	0.711	0.720	0.703	0.707	0.732
		Yes	**0.890**	0.882	0.887	**0.890**	0.885	0.879	0.880	0.878	0.884	0.887	0.877	0.883
BCDR F01	0.829	No	0.670	0.648	0.681	0.740	0.765	0.688	0.695	0.764	0.768	0.712	0.788	**0.860**
		Yes	0.844	0.830	0.841	0.833	0.876	0.799	0.823	0.848	0.849	0.843	0.841	**0.894**
Calcifications														
DDSM sample	0.807	No	0.733	0.754	0.700	0.718	**0.774**	0.764	0.695	0.766	0.773	0.729	0.717	0.706
		Yes	0.799	0.779	0.787	0.791	0.797	0.787	0.769	**0.803**	0.792	0.783	0.764	0.777
BCDR F01	0.725	No	0.711	0.704	0.728	0.617	**0.793**	0.694	0.683	0.790	0.765	0.756	0.710	0.778
		Yes	0.790	0.768	0.783	0.741	**0.815**	0.737	0.728	**0.815**	0.801	0.800	0.747	0.783
Number of wins			2	0	0	1	3	2	0	3	2	0	0	**8**

The highest score of each scenario is highlighted at bold, and scores with no evidence of differences to the highest ($p < 0.05$) are underlined. The last row shows the total number of times each descriptor achieved the highest (or comparable to highest) score

IS intensity statistics, *HM* histogram measures, *IM* invariant moments, *Zer* Zernike moments, *Har* Haralick features, *Gab* Gabor filter banks, *Wav* wavelets, *Curv* curvelets

evidence of significant differences to HGD (mAUC = 0.736, $p = 0.439$), while significantly outperforming the remainder descriptors ($p < 0.05$). When combining clinical data with image descriptors, 4 of the 12 descriptors did not show statistical evidence of outperforming the standalone clinical data, namely GLDM, Gabor Filters, HOG, and HGD. The highest result was achieved by intensity statistics combined with clinical data (mAUC = 0.868), significantly outperforming the remainder ($p < 0.002$).

On the BCDR-F01 data set, the standalone clinical data had a performance of mAUC = 0.712 and were significantly outperformed by HGD (mAUC = 0.825, $p < 0.001$) and HOG (mAUC = 0.739, $p < 0.001$). The HGD descriptor was clearly superior to the remainder ($p < 0.001$) with a difference on mAUC of 0.085 when compared to HOG, the second best descriptor. HGD was the only descriptor that did not significantly alter its standalone performance when combining with clinical data ($p = 1.000$) and remained to be the best, significantly outperforming the remainder descriptors ($p < 0.009$). GLRL combined with clinical data was not able to outperform standalone clinical data.

Overall, HGD was the only descriptor that scored best (or comparable to best) on both DDSM and BCDR-F01 data sets, with 3 wins out of 4.

Masses

Once again, the standalone clinical data significantly outperformed all the standalone image descriptors on the DDSM data set ($p < 0.001$), scoring mAUC = 0.867. As in the previous experiment, the best standalone image descriptor was GLRL (mAUC = 0.733), with no statistical evidence of differences to HGD (mAUC = 0.732, $p = 0.492$), while significantly outperforming the remainder ($p < 0.030$). When combining clinical data with image descriptors, all the image descriptors significantly outperformed the standalone clinical data ($p < 0.001$) and the highest score was achieved by Intensity statistics with mAUC = 0.890, with no statistical evidence of differences to Zernike moments (mAUC = 0.890, $p = 0.213$) and HGD (mAUC = 0.883, $p = 0.248$).

On the BCDR-F01 data set, the standalone clinical data had a performance of mAUC = 0.829 and were only outperformed by HGD (mAUC = 0.860, $p < 0.001$). When combining clinical data with image descriptors, four descriptors (intensity histograms, Zernike, GLRL and GLDM) did not show evidence of increasing the performance of standalone clinical data. All the remainder descriptors outperformed clinical data ($p < 0.028$), with HGD being the best (mAUC = 0.894). HGD performance was significantly superior to all the remainder ($p < 0.024$).

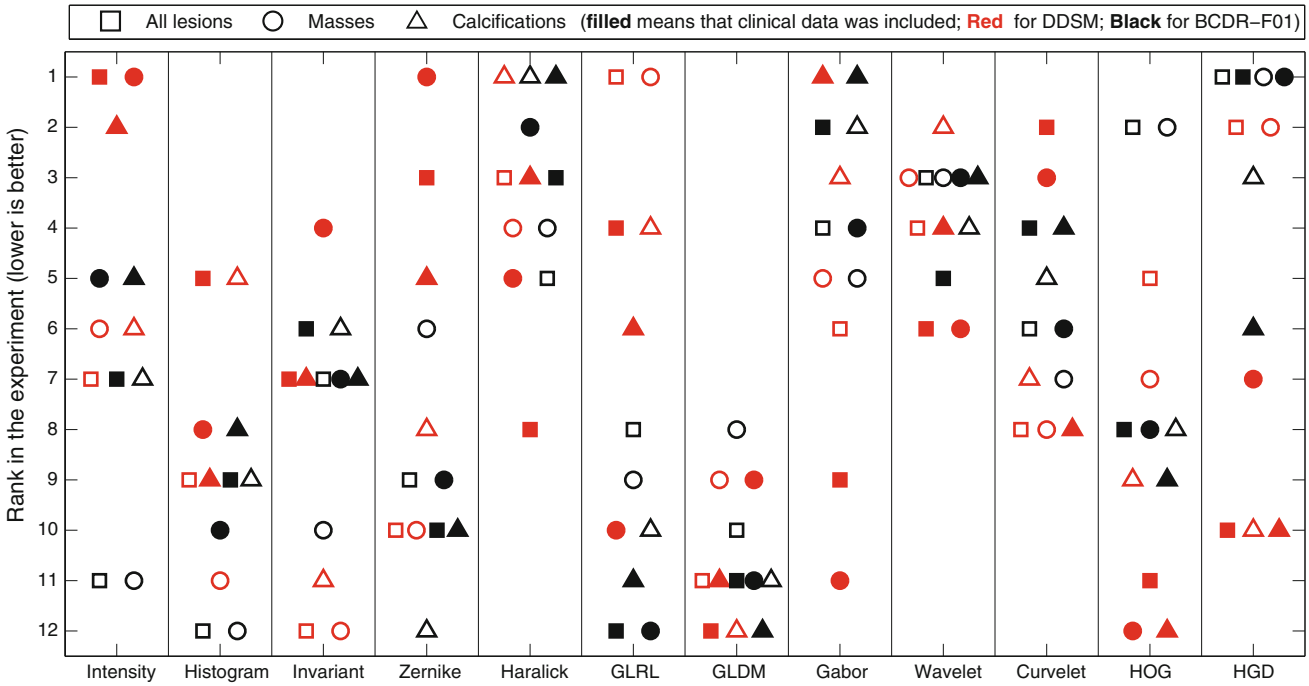

Fig. 5 Rankings of the descriptors for all the evaluated scenarios

Overall, HGD scored best (or comparable to best) on both DDSM and BCDR-F01 data sets on all experiments with masses (four out of four wins).

Calcifications

Like in the previous experiments, the standalone clinical data significantly outperformed all the standalone image descriptors on the DDSM data set ($p < 0.001$), scoring mAUC = 0.807. The best standalone image descriptor for classifying calcifications was Wavelets (mAUC = 0.774) with no significant differences to Haralick (mAUC = 0.773, $p = 0.635$). In contrast with the previous experiments, when clinical data were combined with image descriptors, none of the combinations was able to significantly outperform the standalone clinical data. In fact, with the exception of Gabor Filters, all the descriptors have significantly decreased the performance of standalone clinical data ($p < 0.022$).

On the BCDR-F01 data set, the standalone clinical data had a performance of mAUC = 0.725 and were significantly outperformed by several standalone image descriptors, namely Haralick, Gabor Filters, HGD, Wavelets, and Curvelets ($p < 0.009$). Haralick scored highest (mAUC = 0.793), but with no evidence of statistical differences to Gabor Filters, HGD, and Wavelets. All the descriptors with the exception of GLDM and GLRL outperformed the standalone performance of clinical data when combined with it ($p < 0.002$). The descriptors that combined with clinical data scored highest were Gabor Filters and Haralick, both with mAUC = 0.815, outperforming the remainder ($p < 0.022$).

Overall, three descriptors performed best (or comparable to best) on both DDSM and BCDR-F01 data sets, namely Gabor Filters, Haralick features, and Wavelets with 2 to 3 wins out of 4.

Overall observations

In general, results for the DDSM data sets have lower dispersion than the results for BCDR-F01 (Online Resource 2), which was expected since the number of instances used to train classifiers in DDSM is about 5 times higher. In addition, the standalone performance of clinical data is higher in the DDSM sample, which was also predictable due to the higher number of instances and clinical attributes in DDSM.

On 96 % of the cases, image descriptors have significantly increased their performance when combined with clinical data. This combination besides boosting performance also alters the relative performance of the descriptors. This is particularly visible on intensity statistics (Fig. 5), which achieved top ranks when combined with clinical data on the experiments of the DDSM data set, despite its modest standalone performance. The opposite behaviour is also observable with, for instance, HGD classifying all lesions and GLRL classifying masses on the DDSM (top rankings when standing alone; outperformed when clinical data are available).

The proposed descriptor, HGD, has shown to be superior to the remainder descriptors tested here for mass classification. Its power is particularly observable in the BCDR-F01 data set where its performance stands out, while in the DDSM sample, the standalone performance of HGD is comparable to best. Inspection of the images of both data sets shows that the image quality of BCDR-F01 is superior to DDSM, making masses more distinct from the surrounding tissue, which is favourable to gradient-based descriptors, such as HGD and HOG. This is supported by the standalone performance of HOG when classifying masses, which achieves a strong second place on BCDR-F01, while dropping to seventh on DDSM (Fig. 5). Comparing the performances of HGD and HOG, it is visible that the use that HGD makes of the gradient is advantageous when classifying masses, with HGD always outperforming HOG on both data sets, despite HGD having a more compact representation.

Looking at the rankings distribution of the descriptors (Fig. 5), it is observable that both Wavelets and Haralick features are versatile descriptors, having the lowest dispersion while achieving high ranks. Haralick features were shown to be especially suitable for classifying calcifications, achieving top ranks on both data sets, followed by Gabor filters and then Wavelets.

Regarding the performance of the classifiers, results show that the selection of the classifier is dependent on the data set (Fig. 6). On the data sets of DDSM, the best classifiers for standalone descriptors were SVM and RF, scoring 58 and 44 % of wins, respectively. RF rises to first place when clinical data are available, scoring 56 % of the wins. On BCDR-F01, wins are more uniformly distributed by classifiers when not using clinical data, with RF, SVM, LMT, and NB scoring 61, 58, 53, and 42 % of wins, respectively. When clinical data are combined with the image descriptors on BCDR-F01, SVM and LMT clearly dominate with 83 and 72 % of wins. On average, there were 1.8 wins per experiment per descriptor.

Discussion

The importance of clinical data is well demonstrated in the experiments reported here. Different kinds of image descriptors were used to provide input to Machine Learning Classifiers (MLC) for classifying breast lesions, and on 96 % of the cases, there was statistical evidence that feeding clinical data together with the image descriptors improves classification results. Moreover, in the DDSM sample, MLC trained only with clinical data always provided better results than any MLC trained only with image descriptors. This is expected due to the descriptive and discriminant properties of the BI-RADS tags included in the clinical data of DDSM, and it was also observed in previous studies (e.g. [47]). Nevertheless, it was also shown that image descriptors can significantly improve the discriminant power of these tags, capturing additional features of the lesions. Increments of performance are $\sim$0.02 in the median AUC and are also accompanied by a decrease in the variance of performance. Only in the calcifications data set of DDSM was not possible to show that image descriptors may contribute to improve the classification performance of breast lesions.

The experiments in the BCDR-F01 sample, where clinical data are limited to age, breast density, and observed abnormalities, also show the importance of using such data, with the performance of MLC based in image descriptors being almost always significantly improved. Here, some image descriptors were capable of outperforming clinical data, and most of the descriptors were able to significantly improve the performance of clinical data when combined with it.

Results show that the relative performance of the standalone image descriptors changes when clinical data are added. Descriptors that have inferior performance when standing alone may be highly ranked when combined with clinical data. Therefore, a special caution is advised when generalizing conclusions about the standalone performance of image descriptors to scenarios where other data are available (e.g. clinical).

Results indicate that most of the descriptors are particularly suitable for a given type of lesion. HGD, the new descriptor proposed here, shows best performance on masses and when all types of lesions are present, on both DDSM and BCDR-F01. This was expected since HGD describes shape through the gradient of the image, and the shape of the lesion is a demonstrated predictor for diagnosis of masses. The new formulation proposed here based on concentric regions and on the concept of gradient divergence results in a compact descriptor that is naturally invariant to rotation and that effectively captures patterns related to the diagnosis of masses. The superiority of HGD to HOG was clearly demonstrated on all scenarios including masses, and in the BCDR-F01 (where masses have good visibility/definition), the difference to the remaining descriptors is clear.

Texture descriptors, namely Haralick features and Gabor filters, have shown to be more adequate for classifying calcifications. This agrees with a study in a different sample of DDSM (1,715 cases) focused on evaluating descriptors for classifying microcalcifications [47], where the authors also concluded that texture descriptors were the most suitable, with features based on the co-occurrence matrix having the most discriminative power scoring AUC = 0.776 for fatty tissues and AUC = 0.636 for dense tissues. Keeping in mind that the study reported here focus on all calcifications rather than only on microcalcifications, Haralick features scored 0.774 in the DDSM sample and 0.793 in the BCDR sample, for all breast densities, which compares well with [47]. Wavelets and Haralick features were

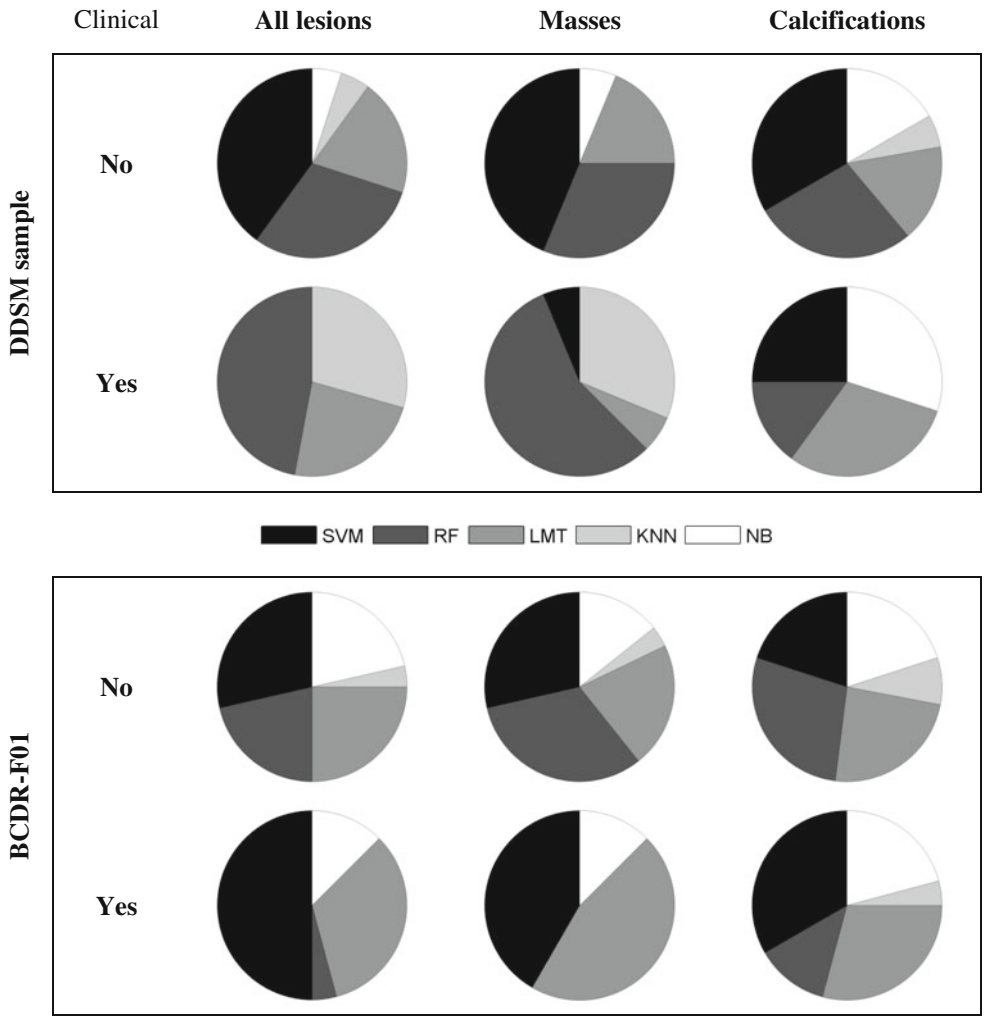

Fig. 6 Frequency of wins by classifier for each data set. A classifier gets one win for each descriptor where it performs best or when there is no statistical evidence ($p < 0.05$) of differences to the best classifier

the most versatile descriptors, showing to be suitable for both masses and calcifications on both DDSM and BCDR, although being far from the standalone performances of HGD when masses are present. Based on another study [27] where wavelets and curvelets are compared for breast lesion classification, it would be expected that curvelets would outperform wavelets. However, in [27], fixed-size regions were extracted that were defined big enough to enable using the same number of curvelet levels on all of them. In the experiments described here, the region over which descriptors are calculated depends on the size of the lesion, and for small lesions, it was not possible to compute the coefficients of the curvelets for all the levels, resulting in several missing values for most of the coefficients.

Descriptors describing the intensity were the group with the worst standalone performance, which was predictable since there is not a clear relation to properties of the lesions associated with breast cancer diagnosis. Nevertheless, Intensity statistics and Zernike moments achieved top rankings when combined with clinical data on the DDSM data set, showing that even simple descriptors can complement clinical data and significantly increase performance. However, this was not observed on BCDR-F01 where the number of clinical attributes is much lower, and thus, performance strongly relies on the capacity of the image descriptors to differentiate benign from malign lesions.

Study limitations

It is out of the scope of this study providing insight into the biophysical basis of the image features. Further investigation on this matter would help support and generalize conclusions regarding the performance of the descriptors on different data sets. Here, reliability on results and conclusions was accomplished by running a high number of runs per experiment, selecting parameters using cross-validation

and utilizing two very distinct data sets with different image resolution, pixel-depth, radiography equipment, number of instances, and clinical descriptors. Despite this care, it is not guaranteed that conclusions from this study may be generalized to other mammography data sets with different properties. Namely, conclusions are only valid for the sets of clinical and image descriptors explored here.

Conclusion

The contributions of this paper are twofold: a new image descriptor is proposed, histograms of gradient divergence (HGD), and the performance of several image descriptors diagnosing breast cancer is evaluated under the presence and absence of clinical data. This work demonstrates that combining image features and clinical data is advantageous. Moreover, it shows that the best standalone image descriptors do not necessarily remain the best when combined with clinical data. This study also shows that there are descriptors that have comparable to best performance on two distinct breast cancer mammography-based data sets, which differ on image resolution, clinical attributes and number of cases. In specific, HGD has shown promising results evaluating masses. The description of patches through deviations of the gradient to a convergence pattern allowed developing a compact descriptor that is naturally invariant to rotation and that can capture properties about the shape of the object contained in the patch. The success of HGD in the experiments reported here makes us believe that it may also be successful in other applications of medical imaging and computer vision in general.

Future work includes determining whether combining different image descriptors may improve classification performance, also in the presence and absence of clinical data. Possible research paths include the use of voting systems or combining attributes from several descriptors followed by a feature selection algorithm.

Acknowledgments The authors would like expressing their gratitude to the Department of Radiology at Hospital São João Porto, Portugal, who provided the data and assisted in the validation of the data sets used in this research. Prof. Guevara acknowledges POPH—QREN—Tipologia 4.2—Promotion of scientific employment funded by the ESF and MCTES, Portugal. Finally, the authors acknowledge TM Deserno, Dept. of Medical Informatics, RWTH Aachen, Germany, for providing the PNG images of the DDSM database.

Conflict of interest The authors declare that they have no conflict of interest.

Ethical standards All experiments were performed with public data from previous studies, and therefore, no ethical violations may result from the experiments reported here.

References

1. Matheus BR, Schiabel H (2011) Online mammographic images database for development and comparison of CAD schemes. J Digit Imaging 24(3):500–506. doi:10.1007/s10278-010-9297-2
2. Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS (2012) INbreast: toward a full-field digital mammographic database. Acad Radiol 19(2):236–248
3. Nelson HD, Tyne K, Naik A, Bougatsos C, Chan BK, Humphrey L (2009) Screening for breast cancer: systematic evidence review update for the US Preventive Services Task Force. Ann Intern Med 151(10):727
4. Tabar L, Vitak B, Chen THH, Yen AMF, Cohen A, Tot T, Chiu SYH, Chen SLS, Fann JCY, Rosell J, Fohlin H, Smith RA, Duffy SW (2011) Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. Radiology 260(3):658–663. doi:10.1148/radiol.11110469
5. Ramos-Pollán R, Guevara-López M, Suárez-Ortega C, Díaz-Herrero G, Franco-Valiente J, Rubio-del-Solar M, de Posada González N, Vaz M, Loureiro J, Ramos I (2011) Discovering mammography-based machine learning classifiers for breast cancer diagnosis. J Med Syst 1:11. doi:10.1007/s10916-011-9693-2
6. Warren Burhenne LJ, Wood SA, D'Orsi CJ, Feig SA, Kopans DB, O'Shaughnessy KF, Sickles EA, Tabar L, Vyborny CJ, Castellino RA (2000) Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 215(2):554–562
7. Cheng HD, Cai X, Chen X, Hu L, Lou X (2003) Computer-aided detection and classification of microcalcifications in mammograms: a survey. Pattern Recogn 36(12):2967–2991. doi:10.1016/s0031-3203(03)00192-4
8. Cheng HD, Shi XJ, Min R, Hu LM, Cai XP, Du HN (2006) Approaches for automated detection and classification of masses in mammograms. Pattern Recogn 39(4):646–668. doi:10.1016/j.patcog.2005.07.006
9. Christoyianni I, Dermatas E, Kokkinakis G (2000) Fast detection of masses in computer-aided mammography. IEEE Signal Proc Mag 17(1):54–64
10. Huo ZM, Giger ML, Vyborny C, Wolverton DE, Schmidt RA, Doi K (1998) Automated computerized classification of malignant and benign masses on digitized mammograms. Acad Radiol 5(3):155–168
11. Constantinidis AS, Fairhurst MC, Rahman AFR (2001) A new multi-expert decision combination algorithm and its application to the detection of circumscribed masses in digital mammograms. Pattern Recogn 34(8):1527–1537
12. Belkasim SO, Shridhar M, Ahmadi M (1991) Pattern-recognition with moment invariants—a comparative-study and new results. Pattern Recogn 24(12):1117–1138
13. Yu SY, Guan L (2000) A CAD system for the automatic detection of clustered microcalcifications in digitized mammogram films. IEEE Trans Med Imaging 19(2):115–126
14. Dhawan AP, Chitre Y, KaiserBonasso C, Moskowitz M (1996) Analysis of mammographic microcalcifications using gray-level image structure features. IEEE Trans Med Imaging 15(3):246–259
15. Wang D, Shi L, Ann Heng P (2009) Automatic detection of breast cancers in mammograms using structured support vector machines. Neurocomputing 72(13–15):3296–3302. doi:10.1016/j.neucom.2009.02.015
16. Dua S, Singh H, Thompson HW (2009) Associative classification of mammograms using weighted rules. Expert Syst Appl 36(5):9250–9259. doi:10.1016/j.eswa.2008.12.050
17. Sahiner B, Chan HP, Petrick N, Helvie MA, Hadjiiski LM (2001) Improvement of mammographic mass characterization

using spiculation measures and morphological features. Med Phys 28(7):1455–1465. doi:10.1118/1.1381548
18. Sahiner B, Chan HP, Petrick N, Helvie MA, Goodsitt MM (1998) Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis. Med Phys 25(4):516–526
19. Sahiner B, Chan HP, Petrick N, Wei DT, Helvie MA, Adler DD, Goodsitt MM (1996) Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. IEEE Trans Med Imaging 15(5):598–610
20. Haralick RM, Shanmuga K, Dinstein I (1973) Textural features for image classification. IEEE T Syst Man Cyb 3(6):610–621
21. Kim JK, Park HW (1999) Statistical textural features for detection of microcalcifications in digitized mammograms. IEEE Trans Med Imaging 18(3):231–238
22. Buciu I, Gacsadi A (2010) Directional features for automatic tumor classification of mammogram images. Biomed Signal Process Control 6(4):370–378
23. Ferreira CBR, DbL Borges (2003) Analysis of mammogram classification using a wavelet transform decomposition. Pattern Recogn Lett 24(7):973–982. doi:10.1016/s0167-8655(02)00221-0
24. Rashed EA, Ismail IA, Zaki SI (2007) Multiresolution mammogram analysis in multilevel decomposition. Pattern Recogn Lett 28(2):286–292. doi:10.1016/j.patrec.2006.07.010
25. Dhawan AP, Chitre Y, Kaiser-Bonasso C (1996) Analysis of mammographic microcalcifications using gray-level image structure features. IEEE Trans Med Imaging 15(3):246–259
26. Soltanian-Zadeh H, Rafiee-Rad F, Pourabdollah-Nejad DS (2004) Comparison of multiwavelet, wavelet, Haralick, and shape features for microcalcification classification in mammograms. Pattern Recogn 37(10):1973–1986. doi:10.1016/j.patcog.2003.03.001
27. Meselhy Eltoukhy M, Faye I, Belhaouari Samir B (2010) A comparison of wavelet and curvelet for breast cancer diagnosis in digital mammogram. Comput Biol Med 40(4):384–391. doi:10.1016/j.compbiomed.2010.02.002
28. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE computer society conference on computer vision and, pattern recognition, vol 881, pp 886–893. doi:10.1109/cvpr.2005.177
29. Heath M, Bowyer K, Kopans D, Moore R, Kegelmeyer P (2000.) The digital database for screening mammography. In: Proceedings of the 5th international workshop on digital mammography, pp 212–218
30. Ramos Pollán R, Rubio del Solar M, Franco Valiente JM, Moriche JE, Gonzalez de Posada N, Valdes Torres JA, Pires Vaz MA, Guevara López MA (2010) Exploiting e-infrastructures for medical image storage and analysis: a grid application for mammography CAD. In: Hierlemann A (ed) Proceedings of the 7th IASTED international conference on, biomedical engineering
31. de Oliveira JEE, Machado AMC, Chavez GC, Lopes APB, Deserno TM (2010) MammoSys: a content-based image retrieval system using breast density patterns. Comput Methods Programs Biomed 99(3):289–297. doi:10.1016/j.cmpb.2010.01.005
32. Oliveira JEE, Gueld MO, Araújo AA, Ott B, Deserno TM (2008) Towards a standard reference database for computer-aided mammography. In: Proceedings SPIE 6915, medical imaging 2008: computer-aided diagnosis, 69151Y, pp 1Y1–1Y9. doi:10.1117/12.770325
33. Gonzalez RC, Woods RE, Eddins SL (2004) Digital image processing. Prentice Hall, New Jersey
34. Sheshadri HS, Kandaswamy A (2007) Experimental investigation on breast tissue classification based on statistical feature extraction of mammograms. Comput Med Imaging Graph 31(1):46–48. doi:10.1016/j.compmedimag.2006.09.015
35. Kinoshita S, Azevedo-Marques P, Pereira R Jr, Rodrigues J, Rangayyan R (2007) Content-based retrieval of mammograms using visual features related to breast density patterns. J Digit Imaging 20(2):172–190. doi:10.1007/s10278-007-9004-0
36. Hu MK (1962) Visual-pattern recognition by moment invariants. Ire T Inform Theor 8(2):179–187
37. Teague MR (1980) Image-analysis via the general-theory of moments. J Opt Soc Am 70(8):920–930
38. Wei C-H, Chen SY, Liu X (2011) Mammogram retrieval on similar mass lesions. Comput Methods Programs Biomed 106(3):234–248. doi:10.1016/j.cmpb.2010.09.002
39. Galloway MM (1975) Texture analysis using gray level run lengths. Comput Graph Image Process 4(2):172–179. doi:10.1016/s0146-664x(75)80008-6
40. Daugman JG (1985) Uncertainty relation for resolution in space, spatial-frequency, and orientation optimized by two-dimensional visual cortical filters. J Opt Soc Am A 2(7):1160–1169
41. Candes EJ, Donoho DL (2004) New tight frames of curvelets and optimal representations of objects with piecewise C2 singularities. Commun Pure Appl Math 57(2):219–266
42. Candes E, Demanet L, Donoho D, Ying L (2006) Fast discrete curvelet transforms. Multiscale Model Simul 5(3):861–899. doi:10.1137/05064182X
43. Lowe D (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110. doi:10.1023/b:visi.0000029664.99615.94
44. Deniz O, Bueno G, Salido J, De la Torre F (2011) Face recognition using histograms of oriented gradients. Pattern Recogn Lett 32(12):1598–1603
45. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explor 11(1):10–18
46. Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30
47. Andreadis II, Spyrou GM, Nikita KS (2011) A comparative study of image features for classification of breast microcalcifications. Meas Sci Technol 22(11):114005–114013. doi:10.1088/0957-0233/22/11/114005