Multimed Tools Appl (2017) 76:33–57
DOI 10.1007/s11042-015-3017-3

A comparison of different Gabor feature extraction approaches for mass classification in mammography

Salabat Khan[1] · **Muhammad Hussain**[2] · **Hatim Aboalsamh**[2] · **George Bebis**[3]

Received: 10 March 2015 / Revised: 16 September 2015 / Accepted: 19 October 2015 /
Published online: 26 October 2015

Abstract We investigate the performance of six different approaches for directional feature extraction for mass classification problem in digital mammograms. These techniques use a bank of Gabor filters to extract the directional textural features. Directional textural features represent structural properties of masses and normal tissues in mammograms at different orientations and frequencies. Masses and micro-calcifications are two early signs of breast cancer which is a major leading cause of death in women. For the detection of masses, segmentation of mammograms results in regions of interest (ROIs) which not only include masses but suspicious normal tissues as well (which lead to false positives during the discrimination process). The problem is to reduce the false positives by classifying ROIs as masses and normal tissues. In addition, the detected masses are required to be further classified as malignant and benign. The feature extraction approaches are evaluated over the ROIs extracted from MIAS database. Successive Enhancement Learning based weighted Support Vector Machine (SELwSVM) is used to efficiently classify the generated unbalanced datasets. The average accuracy ranges from 68 to 100 % as obtained by different methods used in our paper. Comparisons are carried out based on statistical analysis to make further recommendations.

Keywords Mass detection · Gabor filter bank · Directional features · Digital mammography · Feature transformation and reduction · SEL weighted SVM · PCA · LDA

✉ Salabat Khan
salabat.khan@nu.edu.pk

[1] Department of Computer Science, Comsats Institute of Information Technology, Attock, Pakistan

[2] Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

[3] Department of Computer Science and Engineering, University of Nevada, Reno, NV, USA

1 Introduction

Breast cancer is the second major deadliest cancer that affects women all over the world and listed at top among major health problems. The statistics provided by National Cancer Institute, Surveillance, Epidemiology, and End Results (SEER) program, indicate that the lifetime risk of developing breast cancer among American women is 12.2 % (aka: one in eight), exceeded only by the lung cancer [2, 45]. In the European Community, breast cancer represents 19 % of cancer deaths and the 24 % of all cancer cases [14, 24]. 25 % of all breast cancer deaths occur in women, diagnosed between the age of 40 to 49 years. In the United States for instance, breast cancer remains the leading cause of death for women in their forties [24]. The World Health Organization's International Agency for Research on Cancer (IARC) has estimated more than one million cases of breast cancer to be faced annually and reported that more than 400, 000 women die each year from this disease [28]. Cancer can be divided in different stages 0–4 based on the area it has spread in, using surgical procedures. Lower stage numbers indicate the early stage of a cancer which can easily be diagnosed. It is therefore essential to detect the breast cancer at early stage in order to reduce life fatalities [28]. However, detection of breast cancer at its early stages is difficult as it's usually has no symptoms at the beginning. The mortality of breast cancer has declined in women of all ages [28] and this fortunate reduction is considered to be related with the extensive awareness of the disease, self-screening process, widespread usage of mammographic screening and improvements in the treatment process.

Due to its reliability, mammography (an x-ray image examining method of the breast) is considered to be a most effective screening method for the detection of breast cancer. The mammograms are first digitized and then filtered/ analyzed with the help of powerful image analysis techniques in order to develop computer aided diagnosing (CAD) systems for effectively assisting the radiologists. A CAD is a set of automatic or semiautomatic tools developed to assist radiologists in the detection and / or evaluation of mammographic image [24]. There are three types of breast lesions; mass, calcification and architectural disorder [45]. The target of this research work is to identify an optimized feature extraction strategy to learn about the structure of each suspicious abnormalities in ROIs and then assigning a malignancy risk degree using an efficient classification method.

We cannot ignore the importance of biopsy (in the medical terms) in order to detect the masses, most accurately. It is however an expensive procedure and involves some risks e.g., patient discomfort, post biopsy side effects, chances of missing cancerous tissues based on different biopsy methods and is therefore recommended as an eventual solution for mass detection purpose. On the other hand, CAD systems are easy to use tools that are inexpensive and by analyzing the digital mammograms they can effectively assist the radiologists in their decision making process (as a second expert opinion). The idea of using CAD system for breast cancer detection is not recent. CAD systems are used earlier for this task and proved to be useful in the screening process of digital mammograms and in turn detection of early stage malignancies [24, 28, 45]. However, there exist controversial results and views against the usage of CAD systems mainly because of their high false positive and false negative rates in the breast cancer detection, which makes radiologist not really trust them [24]. False negative results occur when CAD system declares a mammogram to be normal even when breast cancer is present. The main cause of the false negatives is the density of the breast, as both dense tissues and tumors are appeared as white regions in the mammogram which makes it difficult to distinguish between them. As women get older, their breasts become fatty and false

negatives are less likely to occur. A false positive is a region in the mammogram that is benign but interpreted as suspicious by the CAD system. High false positive results occur commonly when analyzing the mammograms of the younger women because of the same reason of dense breast tissues. In this research work, our motivation is to investigate six different feature extraction mechanisms to optimize the performance of CAD systems.

For detection of masses in mammograms, we can identify three main stages that constitutes a CAD system: 1) detection and segmentation of potential abnormal areas, 2) false positive reduction, and 3) discrimination of benign and malignant masses. The detection and segmentation stage identifies potential mass regions, and detect their precise outlines. The detected ROIs by this stage include not only masses but suspicious normal tissues as well. The false positive reduction stage classifies the detected ROIs into mass and normal ROIs. The detected mass ROIs are further discriminated as benign and malignant in the final stage. Many efforts have been made so far for false positive reduction and benign-malignant classification but these are still challenging problems. In this research work, we investigate and compare robust, optimized and discriminative feature extraction mechanisms for false positive reduction and benign-malignant classification to effectively address these challenging issues.

The feature extraction technique proposed in [20] is observed to perform well when tested to identify normal tissues from true malignant masses. This task is however simple as compared to false positive reduction and benign-malignant classification tasks based on the fact that the normal tissues can very easily be discriminated from true malignant masses due to highly dissimilar patterns. In this paper, we are interested in observing the performance of this feature extraction method and its variations (based on two state of the art feature transformation strategies), and other Gabor feature extraction techniques exist in the literature, for these two complex classification problems. The variants of the method in [20] with the collaboration of feature transformation strategies can further ensure that only the most representative properties are used (by removing the redundant responses of the bank) for discrimination between the normal and abnormal tissues. All of these methods analyze the textural properties of masses using a bank of several Gabor filters (discussed later). The key idea behind the usage of several Gabor filters is to improve the performance of breast cancer recognition system by responding strongly to the features that best distinguishes between the normal and abnormal tissues, from different orientations of filters in different scales. Based on the size of mammogram region, filtered by the bank, extraction of textural patterns can be done either locally (sub-region of ROI is filtered) or globally (entire ROI is filtered). Both of these local and global textural descriptors, characterize the micro-patterns (e.g., edges, lines, spots and flat areas) in digital mammograms that are very helpful for detection of masses [24]. Local textural descriptors, preserve at the same time the spatial information of masses and other regions in the digital mammograms and thus become more attractive choice for the same mentioned task.

The filters in the Gabor bank are initialized with different scales and orientations to extract any possible patterns in the ROIs that might be helpful for discrimination of normal and abnormal tissues. Although, Gabor filters are used for the breast cancer detection earlier (see e.g., [20, 45] and references therein), this work propose more variants of the existing feature extraction strategies [20] which are observed to offer much better performance than their original form. Feature transformation algorithms (used in this paper) effectively remove the redundant or irrelevant responses of the Gabor filters and thus are extremely helpful for improving the performance of a CAD system. Manually extracted normal/ abnormal tissues are filtered with the Gabor filters to extract directional features which are eventually used for classification of digital mammograms.

The remainder of this paper is organized as follows. In the next section, we review the related research work. In Section 3, we present the methodology with brief discussion on the feature extraction strategies and classification algorithm. Subsequently, in Section 4, we present experimental results to show the effectiveness of the feature extraction techniques. Finally, Section 5 will conclude this work.

2 Related work

Mass detection problem has attracted the attention of many researchers, and many detection techniques have been proposed [20]. For a detailed review of these methods, an interested reader is referred to the review papers [10, 13, 31, 38]. In the following paragraphs, we give an overview of the most related recent mass detection methods.

Most of the existing methods differ in the types of features that have been used for mass detection and the way these features have been extracted. Different types of features such as texture, gradient, grey-level, shape [31] features have been employed for mass detection. Texture is an important characteristic that helps to discriminate and identify the objects. In addition to other identification/detection tasks, texture descriptors have been used for detecting normal and lesion regions in mammograms [29, 37, 42]. Wei et al. [43] extracted multiresolution texture features from wavelet coefficients and used them for the discrimination of masses from normal breast tissue on mammograms. They used linear discriminant analysis for classifying the ROIs as mass or non-mass. This method was tested with 168 ROIs containing biopsy-proven masses and 504 ROIs containing normal parenchyma, and resulted in Az (percentage area under ROC curve) equal to 0.89 and 0.86 for the training and test groups.

If texture is described accurately, then texture descriptors can perform better than other descriptors [24]. Lladó et al. [24] used spatially enhanced LBP (Local Binary Pattern) descriptor, which is basically a texture descriptor, to represent textural properties of masses and to reduce false positives; this method achieved an overall accuracy of $Az=0.94\pm0.02$ (percentage area under ROC curve) on 512 ROIs (256 normal and 256 masses) extracted from mammograms from DDSM database. LBP based method outperforms other CAD methods for mass detection. But LBP descriptor builds statistics on local micro-patterns (dark/bright spots, edges, and flat areas etc.) and is not robust against noise. The scheme proposed by Sampaio et al. [36] used geo-statistic functions for extracting texture features, SVM for classification and obtained the accuracy of $Az=0.87$.

Gabor wavelets are among different methods which have been used for texture description in various image processing and analysis approaches [17, 40]. Gabor filters decompose an image into multiple scales and orientations and make the analysis of texture patterns easy. Mammograms contain a lot of texture, and as such Gabor filters are suitable for texture analysis of mammograms [3, 35] as well. Different texture description techniques using Gabor wavelets differ in the way the texture features are extracted. Gabor wavelets have also been used to extract features for mass detection [23, 45]. Zheng [45] employed Gabor filters to create 20 Gabor images, which were then used to extract a set of edge histogram descriptors. He used KNN along with fuzzy c-means clustering as a classifier. The method was evaluated on 431 mammograms (159 normal cases and 272 containing masses) from DDSM database using tenfold cross validation. This method achieved true positive (TP) rate of 90 % at 1.21 false positive per image. The data set used for validation is biased toward abnormal cases which will surely favor the mass cases, and it cannot be regarded as fair evaluation. This method extracts edge histograms which are holistic descriptor, and does not represent the local textures of masses.

Lahmiri and Boukadoum [23] used Gabor filters along with discrete wavelet transform (DWT) for mass detection. They applied Gabor filter bank at different frequencies and spatial orientations on HH high frequency sub-band image obtained using DWT, and extracted statistical features (mean and standard deviation) from the Gabor images. For classification, they used SVM with polynomial kernel. The method was tested on 100 mammograms from DDSM database using tenfold cross validation. This method achieved an accuracy of 98 %. Costa et al. [7] explored the use of Gabor wavelets along with principal component analysis (PCA) for feature extraction, independent component analysis (ICA) for efficient encoding, and linear discriminant analysis (LDA) for classification. The success rate of this method with feature extraction using Gabor wavelets was 85.05 % on 5090 ROIs extracted from mammograms in DDSM database.

Geralodo et al. [22] have used Moran's index and Geary's coefficients as input features for SVM classifier and tested their approach over two cases i.e., normal vs. abnormal and benign vs. malignant regions classification. They obtained accuracy of 96.04 % and Az ROC of 0.946 with Geary's coefficient and an accuracy of 99.39 % and Az ROC of 1 with Moran's index for the classification of normal or abnormal cases. For the second case (benign vs. malignant), an accuracy of 88.31 % and Az ROC of 0.804 with Geary's coefficient and accuracy of 87.80 % and Az ROC of 0.89 with Moran's index is reported. The method is tested over 1394 ROI images collected from DDSM database using tenfold cross validation. In the research work of Ioan Buciu et al. [21], raw magnitude responses of 2D Gabor wavelets are investigated as features for proximal SVM. A total of 322 mammogram images from Mammographic Image Analysis Society (MIAS) database are used for three experimental cases i.e., discrimination between the three classes: normal, benign and malign (using one against all SVM classification), normal vs. tumor (benign and malign) and benign vs. malign using 80 % data features for training and 20 % as testing sets. The features dimension in this case is equal to the number of pixels present in the downsampled mammogram images (for a single Gabor filter), later PCA is used for dimensional reduction. The best results (in terms of accuracy) for the three experimental cases are: 75, 84.37 and 78.26 %, respectively. In order to observe the robustness of the method, ROI images corrupted with quantum noise are used for feature extraction and the method achieves comparable results (lesser decrease in recognition rate) with those of noise-free ROI images.

The aforementioned research works related to 2D Gabor wavelets are mostly concerned with using a generic (non-optimized) setting of filters present in the bank [1, 4, 21, 33, 45]. Following the same trend, we identified some main contributions of this paper as follows:

- Comparison of feature extraction methods for false positive reduction and benign-malignant classification.
- A new Gabor feature extraction method named Statistical Magnitude Gabor Response (*SMGR*) is proposed which significantly reduces the feature size for classification.
- The variants of windows based SMGR method (proposed in our earlier work [20]) are supported with two state of the art feature reduction algorithms based on which they have reduced the erroneous predictions up to a significant level and thus are very attractive for the radiologists.
- With tenfold cross validation experiments, methods are confirmed to perform robustly or weakly when trained with different ratios of normal and abnormal ROIs.
- Detailed experiments using common machine learning evaluation methodologies and measures e.g., area under the ROC value, sensitivity, specificity, accuracy, are provided for a more general performance comparisons.

3 Methods

In this section, we discuss the feature extraction strategies one by one, for mass classification in digital mammograms. We review commonly used feature reduction techniques (in Section 3.2 and 3.3), which we are going to employ for extracting different types of Gabor features. We have observed (in our experiments) that different methods have quite a different impact on the recognition rate. Some methods give poor performance and the others are extremely accurate. This section is further divided in the following subsections. First, a brief overview of the Gabor filter bank is provided. Afterwards, feature transformation algorithms are discussed that are helpful for achieving better performance results followed by feature extraction methods. In the final subsection, we reviewed the SEL based weighted support vector machine that is used for classification purpose.

3.1 Gabor filter

Texture is an important part of the visual world of animals and humans and they can successfully detect, discriminate, and segment texture using their visual systems [32]. Textural properties in an image can be used to collect different information's e.g., micro-patterns like edges, lines, spots & flat areas. Masses in an ROI do contain strong edges and local spatial patterns at different frequencies and orientations. These micro-patterns are helpful in recognition of cancerous regions in a CAD system. Gabor filters can effectively be used to detect these micro-patterns & this research work aims to validate this statement. A brief overview of the Gabor filters is given in the next paragraph.

Gabor filters are biologically motivated convolution kernels [8] that have enjoyed widely usage in a myriad of applications in the field of computer vision & image processing e.g., face recognition [44], facial expression recognition, iris recognition, optical character recognition, vehicle detection [46] etc. In order to extract local/ global spatial textural micro-pa tterns in ROIs, Gabor filters can be tune with different orientations and scales thus provide powerful statistics which could be very useful for breast cancer detection. The general function $g(x,y)$ of 2D (for image) Gabor filter family can be represented as a Gaussian kernel modulated by an oriented complex sinusoidal wave can be described [46]:

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}.e^{\left[-\frac{1}{2}\left(\frac{\tilde{x}^2}{\sigma_{\tilde{x}}^2}+\frac{\tilde{y}^2}{\sigma_{\tilde{y}}^2}\right)\right]}.e^{\left(2\pi jW\tilde{x}\right)}. \tag{1}$$

$$\tilde{x} = x.\cos\theta + y.\sin\theta \quad and \quad \tilde{y} = -x.\sin\theta + y.\cos\theta. \tag{2}$$

Where σ_x and σ_y are the scaling parameters of the filter and describe the neighborhood of a pixel where weighted summation takes place. W is the central frequency of the complex sinusoidal and $\theta \in [0,\pi)$ is the orientation of the normal to the parallel stripes of the Gabor function.

A generic strategy for constructing the Gabor filter bank is adopted from [26]. A particular bank of Gabor filters contain multiple individual Gabor filters adjusted with different parameters (scaling, orientation and central frequency). In this paper, different combination of Gabor filter bank e.g., a Gabor filter bank containing 6 filters (2 scales{S}×3 orientations{O}) referred to as GS2O3, 15 filters i.e., GS3O5, 24 filters i.e., GS4O6 and 40 filters i.e.,

GS5O8 are used with initial max frequency equal to 0.2 and initial orientation set to 0. The orientations and frequency for a bank are calculated using following equations [46]:

$$orientation(i) = \frac{(i-1)*\pi}{O} where\{i = 1, 2, \ldots\ldots, O(total\,orientations)\} \tag{3}$$

$$frequency(i) = \frac{f_{\max=0.2}}{(\sqrt{2})^{i-1}} \quad where\{i = 1, 2, \ldots\ldots, S(total\,scales)\}. \tag{4}$$

3.2 Principal component analysis

Principal component analysis (PCA also known as Karhunen Loève transform) [11, 39] is a popular feature reduction technique that linearly projects a high-dimensional feature vector (e.g., Gabor feature vector without class label, Eq. 13) to a low-dimensional space whose components are uncorrelated. The low-dimensional space (eigenspace) is spanned by the principal components which are the linear combinations of the original space. Given an unlabeled Gabor feature vector ($\Gamma_i \in \mathbb{R}^J$) representing the *ith* ROI, first, an average Gabor feature vector ψ is computed for a total of N ROIs in the training data.

$$\psi = \frac{1}{N}\sum_{i=1}^{N} \Gamma_i. \tag{5}$$

In order to ensure that the data samples have zero mean, the difference ($\Phi_i = \Gamma_i - \psi$) of each Gabor feature vector from the average Gabor feature vector is calculated and the covariance matrix C is estimated as follows:

$$C \approx \frac{1}{N}\sum_{i=1}^{N} \Phi_i {\Phi_i}^T = AA^T. \tag{6}$$

Here, $A = [\Phi_1 \Phi_2 .... \Phi_N]$. Since, it is computationally intractable to find a number of J eigenvectors u_i and eigenvalues for this high dimensional correlation matrix $C \in \mathbb{R}^{J \times J}$ of a typical ROI image size, the eigenvectors v_i for the matrix $A^TA \in \mathbb{R}^{N \times N}$ are calculated first, where $J \gg N$. The eigenvectors u_i corresponding to the correlation matrix can then be calculated as follows [39]:

$$u_i = \sum_{j=1}^{N} v_{ij} \Phi_j \tag{7}$$

Given a high dimensional input Gabor feature vector ($\Gamma \in \mathbb{R}^J$), the subtraction from mean is done ($\Phi = \Gamma - \psi$) and projection to low dimensional space is performed as follows:

$$\tilde{\Phi} = \sum_{i=1}^{R_k} w_i u_i, \quad \text{where } w_i = {u_i}^T \Gamma. \tag{8}$$

Here, w_i are the coefficients of projection matrix and R_k are the first few *k-ranked* eigenvectors that correspond to the k largest eigenvalues. In our experiments, we have used $k = [5, 10, 15, ...All]$ where 'All' corresponds to all the eigenvectors.

3.3 Linear discriminant analysis

Linear discriminant Analysis (LDA) [11, 15] is a supervised linear transformation based feature reduction strategy. It projects a high-dimensional feature vector (e.g., Gabor feature vector with class label, Eq. 13) to a low-dimensional space such that the ratio between intra class scatter (within class) S_W and the inter class (between class) scatter S_B is maximized. Considering the same definition of the symbols as given in Section 3.2 for PCA, these scatters can be defined as follows for the multiclass classification problem containing C class labels:

$$S_B = \sum_{i=1}^{C} N_i(\Psi_i - \Psi)(\Psi_i - \Psi)^T. \tag{9}$$

$$S_W = \sum_{i=1}^{C} \sum_{x_k \in y_i} (x_k - \Psi_i)(x_k - \Psi_i)^T. \tag{10}$$

Here, ψ_i corresponds to the average Gabor feature vector for class i, N_i is the number of training samples that belong to class i and x_k is the *kth* instance in class i. So, we try to find out the optimal projection $W_{optimal}$ such that the ratio between intra class scatter matrix of projected samples and the inter class scatter matrix of projected samples is maximized, given as follows:

$$W_{optimal} = \text{argmax}(W) \frac{|W^T S_B W|}{|W^T S_W W|}. \tag{11}$$

For the selection of most representative features in the projected space, the setting of k=[5, 10, 15,.....*All*] is used, where k represents the number of projected features used for classification purpose. In our case, the number of training samples are much lesser than the number of features and therefore the intra class scatter matrix would tend to be a singular matrix and the LDA computational will be so demanding. In order to cater this issue, PCA is used as a pre-processing step to project the original training data into low-dimensional space and then LDA projection is performed.

3.4 Feature extraction strategies

A detail description of six different feature extraction methods is given in follows.

Magnitude gabor responses transformed with PCA and LDA The feature extraction method presented in [21] produces Magnitude Gabor Responses (*MGR*) by applying Gabor filters to the entire ROI image and use magnitude values of the filtered pixels, directly as feature values without any further post processing. The dimension of features in this case is equal to the number of pixels in ROIs, multiplied with the number of Gabor filters, present in the bank. For example, using *MGR* [21], the dimension of features is about 40960 for an ROI resolution of 32×32 pixels with a bank containing 40 Gabor filters. Clearly, such a huge dimension makes the classification task challenging due to the presence of several irrelevant and redundant Gabor responses. In order to handle this shortfall of *MGR*, PCA has been used [21]. In addition to PCA, LDA can also be used to overcome the same problem. In this way, two feature extraction methods are formed; first, PCA_MGR that uses PCA, and second, LDA_MGR that uses LDA, to transform the features, generated by *MGR*, into low dimension space.

First order statistics of magnitude gabor responses The third feature extraction strategy (proposed in this paper) is based on the further processing of *MGR* features such that when a Gabor filter is applied to an ROI, the resultant magnitude Gabor responses are represented with only three statistical values (mean, standard deviation and skewness), thus called Statistical Magnitude Gabor Response (*SMGR*) based feature extraction strategy. *SMGR* reduces the dimension of extracted features, significantly, as compared to *MGR* and *WSMGR* (discussed below), and offers comparable recognition rate to that of the *MGR*. The dimension of features produced with *SMGR* is already low and therefore doesn't require any further reduction.

When a Gabor filter is applied to a pixel value, it generates a complex number having real and imaginary parts. The magnitude/ absolute value of the complex number (a + bi) is calculated as follows:

$$|a + bi| \sqrt{a^2 + b^2}. \tag{12}$$

For SMGR, a single Gabor filter is applied to all the pixels of ROI and magnitude values are calculated. Later, three statistical values (mean, standard deviation and skewness) of the magnitude values are used as features for that particular Gabor filter. The method is repeated for all the Gabor filters in the bank to generate the feature vector for the given ROI.

The methods discussed so far are global feature extraction techniques; in the following we discuss some local feature extraction techniques:

Windows based first order statistics of magnitude Gabor responses In [20], moments based magnitude values for a group of pixels in the overlapping windows are used to construct a feature vector. Instead of applying a Gabor bank on the entire ROI image, an ROI is first segmented/ partitioned into overlapping windows. In particular, each hypothesized ROI image is first divided into equal sizes of square patches/ blocks and then later by combining these patches several overlapping windows are formed (for more detail [20]). In this way, by increasing/ decreasing the size of a patch, ROI image can be partitioned in different sizes & numbers of windows. Feature extraction (for *WSMGR*) is performed by convolving ROI windows with a Gabor filter bank. So, *WSMGR* method actually generates Windows based Statistical Magnitude Gabor Responses (*WSMGR*) for the textural patterns present in the ROIs. Gabor filters are applied to the overlapping windows of ROIs and statistical representative values for the filtered pixels in these windows are used as feature values. A slight modified design strategy (for feature extraction) is already used for texture base features extraction [26, 46]. Bhangaleet. al. [4] applies Gabor filter on an entire ROI and divide the filtered ROI into non-overlapping blocks and later, for a block, mean and standard deviation of the pixels intensities (in the block) are used as feature values. For WSMGR, an ROI is first partition into overlapping windows (i.e., small regions as shown in Fig. 1) and a single Gabor filter is applied to all the pixels of the window and magnitude values are calculated. Later, three statistical values (mean, standard deviation and skewness) of the magnitude values are used as features for that particular Gabor filter. The method is repeated for all the Gabor filters in the

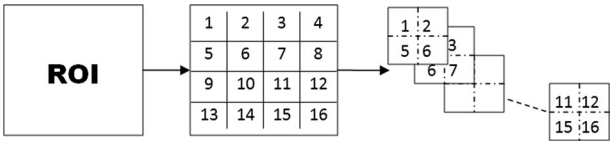

Fig. 1 Segmentation of ROI in blocks and overlapping sub-windows (*left to right*)

bank on all the windows of ROI to generate the feature vector for the given ROI. Partitioning the ROI, prior to filtering, makes the filtering process highly parallelizable e.g., in the presence of multi-core CPUs and GPUs, multiple windows can be filtered, in parallel [20].

The raw responses of Gabor filters bank (in the form of complex values) can also be used as features for classification (as was the case for *MGR*) but usually some post processing is performed to acquire most representative features e.g., (Gabor energy features, threshold Gabor features and moments based Gabor features) [17, 46]. For *WSMGR*, magnitude responses of each Gabor filter in the bank are collected from all windows and represented by three moments: the mean $\mu_{i,j}$, the standard deviation $\sigma_{i,j}$ and the skewness $k_{i,j}$ (where i corresponds to the ith filter in the bank and j to the jth window) [20].

The moments correspond to the statistical properties of a group of pixels in a window and positioning of pixels is essentially discarded which compensates for any errors that might occur during extraction/ segmentation of ROIs into overlapping windows. Suppose, we are using a Gabor bank of 40 filters (i.e., GS5O8); applying this filter on nine windows [20] of a single ROI, yields a feature vector of size 1080 + (1) class. A row feature vector in this form is shown below:

$$\left[\mu_{1,1}, \sigma_{1,1}, k_{1,1}, \mu_{2,1}, \sigma_{2,1}, k_{2,1}, \ldots, \mu_{40,1}, \sigma_{40,1}, k_{40,1}, \mu_{1,2}, \sigma_{1,2}, k_{1,2}, \ldots., \mu_{40,9}, \sigma_{40,9}, k_{40,9}, class\right]. \quad (13)$$

WSMGR significantly reduces the dimension of extracted features (without feature reduction strategy) as compared to *MGR*, as shown in Table 1. The dimension of features using SMGR is lower than WSMGR, however, the recognition performance of WSMGR is better than SMGR as discussed in Section 4. We further collaborate the *WSMGR* with two feature transformation strategies (PCA and LDA) in order to observe any performance gain. In this way, two more feature extraction methods are formed; first, PCA_WSMGR that uses PCA, and second, LDA_WSMGR that uses LDA, to transform the features, generated by *WSMGR*, into low dimension space.

Table 1 Feature dimension for MGR, SMGR and WSMGR for different experimental configurations, without applying feature transformation strategies

Res.	Block WSMGR	Gabor bank	Dataset	No. of features		
				MGR	WSMGR	SMGR
32	8	GS2O3	D1	6144	162	18
		GS3O5	D2	15360	405	45
		GS4O6	D3	24576	648	72
		GS5O8	D4	40960	1080	120
64	16	GS2O3	D5	24576	162	18
		GS3O5	D6	61440	405	45
		GS4O6	D7	98304	648	72
		GS5O8	D8	163840	1080	120
128	32	GS2O3	D9	98304	162	18
		GS3O5	D10	245760	405	45
		GS4O6	D11	393216	648	72
		GS5O8	D12	655360	1080	120
			Average:	152320	573.75	63.75

3.5 Classification

In this paper, we have investigated Successive Enchantment Learning based weighted Support Vector Machine (SELwSVM) for the classification of tumors, present in the ROIs. In our case, we are dealing with a binary classification problem where the target is to build a classification model that can accurately label the unseen data to either belong to the 'mass present' or 'mass absent' classes. SVM classifiers [41] are the most advanced ones, generally, designed to solve binary classification problems; thus perfectly suite our requirements. The only difference between SELwSVM [12, 27] and normal SVM lies in the selection of training samples to be use during the training phase based on a weighting scheme assign to the class labels. SELwSVM makes use of a subset of entire training data to build the classifier and assign unequal weights to the class labels (e.g., based on their frequencies). Whereas on the other hand, normal SVM exploits complete training data to learn the classification model with equal weights assigned to each class label. Keeping in view this difference, we first discuss the weighting scheme (for the class labels) use in our work along with the successive enhancement learning strategy followed by a brief review of SVM classifier.

SELwSVM [12, 27] is recommended to be use when dealing with highly skewed datasets for the classification purpose. In general, when extracting ROIs from different locations of mammograms, most of these ROIs are labeled as "mass absent" and only a few of these belong to the "mass present" class; thus results in a highly unbalanced dataset. This property of mass classification dataset makes SELwSVM an ideal approach to be investigated for the classification purpose. Moreover, misclassification of "mass present" cases is more dangerous and has severe effects towards the causalities. Hence, accuracy of "mass present" class is more important and misclassification of this class should be given high penalty as compared to the "mass absent" class. It can be achieved by assigning higher weights to the "mass present" class and in turns assigning higher penalty for the misclassification of samples belonging to the same class. We adopt weighting scheme as used in [27] that assign the ratio of penalties for different classes to the inverse ratio of the training class sizes and same weights for the samples belonging to same class. The weight of each class is given as:

$$\begin{cases} \dfrac{W_1}{L2} = \dfrac{W2}{L1}, W_1 + W_2 = 1. \end{cases} \tag{14}$$

Here, W_1, W_2 and L_1, L_2 denotes the weight and instance numbers in majority and minority classes, respectively. A potential concern when dealing with highly skewed dataset is whether the randomly selected training samples are well representative of the majority class [12]. To address this issue, we use successive enchantment learning strategy where the basic idea is to select iteratively the most representative "MC absent" examples from all the available training images while keeping the total number of training examples small [12]. This method of learning resembles the bootstrap technique [27] and shown to improve the generalization performance of SVM [12, 27]. The pseudo code of SEL is given as follows; for more understanding, readers are kindly referred to [12] where it is discussed, how choosing "difficult training samples" from the majority class actually improves the recognition rate?

Input: Training data (Gabor textural features for the ROIs with the labels)
Output: *Classification model*
Select randomly an initial set of training examples '*Z*' from the available training data
Classification model = Train the SVM classifier with '*Z*'
REPEAT
  Apply the *Classification model* to all the mammogram regions (except those already present in '*Z*')
  Record the "mass absent" locations that have been misclassified as "mass present"
  Collect '*N*' new input examples (randomly) from the misclassified "mass absent" locations
  Update the set '*Z*' by replacing '*N*' "mass absent" examples that have been classified correctly by
  weighted SVM with the newly collected "mass absent" examples
  Classification model = Re-train the weighted SVM classifier with the updated set '*Z*'
UNTIL (convergence is not achieved i.e., accuracy doesn't improve in three consecutive iterations)

Algorithm 1. Successive enhancement learning algorithm

Considering the learning scheme of SVM, the aim is to find an optimal hyper-plane that can separate the data belonging to different classes with large margins in high dimensional space [5]. The margin is defined as the sum of distances to the decision boundary (hyper-plane) from the nearest points (support vectors) of the two classes. SVM formulation is based on statistical learning theory and has attractive generalization capabilities in linear as well as non-linear decision problems [6, 41]. SVM uses structural risk minimization as opposed to empirical risk minimization [41] by reducing the probability of misclassifying an unseen pattern drawn randomly from a fixed but unknown distribution.

Let $D=\{(x_i,y_i)\}_{i=1}^{N}\subset\mathbb{R}^J\times\{+1,-1\}$ be a training set where x_i is the ith training instance containing J features, y_i is the class label of x_i having two values $\{+1$ or $-1\}$. Finding an optimal hyper-plane based on large margin framework implies solving a constrained optimization problem using quadratic programming and can be stated as:

$$f(x)=\sum_{i=1}^{N}\alpha_i y_i k(x_i,x)+b \tag{15}$$

Where $\alpha_i>0$ are the Langrange multipliers, $k(x_i,x)$ is the kernel function and sign of $f(x)$ gives the membership class of x. For linearly separable problems or linear SVM, kernel function is simply the dot product of the two given points in the input space. However, for non-linear SVMs, the original input space is mapped to the higher dimensional space through a non-linear mapping function (possibly making the data linearly separable), using different suitable kernels (for computational efficiency) defined as a dot product in the new space and satisfies the Mercer's condition [41]. In this new formulation, the misclassification penalty or error is controlled with a user defined parameter C (regularization parameter, controlling tradeoff between error of SVM and margin maximization), and is tied with the kernel. There are several kernels available to be used e.g., linear, polynomial, sigmoid, radial basis function (RBF) etc. In our experiments, RBF kernel is used as given by:

$$k(x_i, x) = \exp\left(-\gamma \|x_i - x\|^2\right), \gamma > 0. \quad (16)$$

The γ is the width of the kernel function. There are two parameters now tied with the RBF kernel, γ and C. Tuning these parameters in an attempt to find a better hypothesis is called model selection procedure. For model selection, we have first performed a loose grid search (coarse search for computational efficiency) to find the better region in the parameter space. Later, the finer grid search is conducted in the region found by loose grid search. This model selection procedure is recommended in the work of Chih-Wei Hsu et al. [18]. The selected parameters are feed into the kernel and SVM is finally applied to our data sets. Detailed discussion on the statistical formulation and computational aspects of SVM can be found in the work of Vapnik [41].

4 Results & discussion

In this section, the experimental results for all the feature extraction strategies (discussed in Section 3) are presented and discussed in a fair amount of detail. We conducted the experiments for two problems: false positive reduction i.e., to classify ROIs into normal and mass (benign + malign) and, the classification of mass ROIs into benign and malignant. First, overview of the database used for the validation of the methods is given. Then, a fair amount of discussion is carried out for the empirical evaluation of the methods for the two diagnosis problems. The extracted ROIs are in different sizes, for processing them with Gabor filter bank, it is necessary to resize them into the same resolution; we tested three different resolutions: 128×128, 64×64 and 32×32. For extracting features (based on *WSMGR*), each ROI can be partitioned into blocks of different sizes for defining overlapping windows. We tested three block sizes: 32×32, 16×16 and 8×8. Afterwards, we perform statistical comparison of the methods (in terms of recognition rate and area under the ROC value) using a non-parametric Friedman test with Holm post-hoc test [9, 16] in order to see whether or not the differences in performance of different methods are actually statistically significant.

4.1 Database & evaluation methodology

The mammogram images used in our experiments are taken from Mammographic Image Analysis Society (MIAS) [25] database; this database consist of more than 2000 cases and is commonly used as a benchmark for testing new proposals dealing with processing and analysis of mammograms for breast cancer detection. Each case in this database is annotated by expert radiologists; the complete information is provided as an overlay file. The locations of masses in mammograms specified by experts are encoded as code-chains. We randomly selected 109 cases from the database. Using code chains, we extracted 20 ROIs which contain true masses; the sizes of these ROIs vary depending on the sizes of the mass regions. In addition, we extracted 54 ROIs containing normal but suspicious tissues and 35 benign ROIs. Some sample ROIs are shown in Fig. 2.

The evaluation of the methods is performed using tenfold cross validation and area under the ROC curve (Az value) analysis. In particular, a data set is randomly partitioned into ten non-overlapping and mutually exclusive subsets. For the experiment of fold i, subset i is selected as testing set and the remaining nine subsets are used to train the classifier. Using

Fig. 2 (*top row*) Normal but suspicious ROIs (*middle row*) Benign mass ROIs (*bottom row*) Malignant mass ROIs

tenfold cross validation experiments, the performance of methods can be confirmed against any kind of selection biased of the samples for training and testing phases. It also helps in determining the robustness of the methods when tested over different ratios of normal and abnormal ROIs used as training and testing sets (due to random selection, ratios will be different). The SVM classifier gives a membership value of each class when an unknown pattern is presented to it. The ROC (receiver operator characteristics) curve can be obtained by varying the threshold on this membership value. The area under ROC curve (Az) is used as a performance measure. The other commonly used evaluation measures are accuracy or recognition rate (RR) = $(TP+TN)/(TP+FP+TN+FN)$, sensitivity (Sn) = $TP/(TP+FN)$, specificity (Sp) = $TN/(TN+FP)$, where $TN$is the number of true negatives, TP is that of true positives, FP is that of false positives and FN denotes the number of false negatives.

4.2 False positive reduction

In this section, the classification of diagnosis case (suspicious normal vs. masses) is investigated based on the proposed method. Mass ROIs contain two types of ROIs: 1) benign, and 2) malignant. It becomes difficult to discriminate between the normal and mass ROIs mainly because of the reason that the benign ROIs are structurally closer to both the normal and mass ROIs. One major point in favor of WSMGR is its low-dimensional feature space as compared to the huge dimensional space generated under the feature extraction strategy of MGR [21]. SMGR on the other hand results in the smallest feature size, see e.g., Table 1. It may please be noted that the block size is only relevant to WSMGR strategy as given in Table 1. In Tables 2 and 3, experimental results are given for the diagnosis case (normal vs. masses) for all the feature extraction strategies (discussed in Section 3) based on performance measures: accuracy and Az, respectively. It can easily be observed that the dimension of average number of features generated using WSMGR is substantially smaller than the dimension of features produced with MGR as given in Table 1. This reduction in feature space not only makes the WSMGR method computationally less demanding but it also improves the recognition

Table 2 Performance of feature extraction strategies over different resolutions of ROIs using tenfold cross validation based on accuracy (normal vs. masses)

Dataset	Accuracy					
	SMGR	WSMGR	PCA_MGR	PCA_WSMGR	LDA_MGR	LDA_WSMGR
D1	67.22±09.31	74.31±9.38	69.58±15.60	77.78±09.82	70.56±12.09	96.25±11.86
D2	67.08±05.71	77.08±16.63	80.42±06.46	80.69±14.78	72.08±17.14	98.75±3.95
D3	68.33±08.38	79.58±11.88	79.58±12.18	84.44±14.40	68.33±10.24	97.50±7.91
D4	69.72±09.39	80.42±14.68	82.78±10.70	82.92±10.44	74.31±14.82	98.75±3.95
D5	70.42±13.78	87.64±10.22	71.94±16.61	86.53±12.48	70.56±10.89	98.75±3.95
D6	70.83±10.02	81.53±9.07	76.94±22.92	82.92±11.98	71.53±10.78	97.50±7.91
D7	68.33±05.96	85.28±9.58	74.58±11.69	85.42±09.63	72.36±15.01	100.00±0.00
D8	67.08±05.71	80.69±19.28	76.94±08.59	84.17±15.54	72.08±14.98	97.50±7.91
D9	68.47±11.13	79.03±8.77	74.58±11.69	83.89±14.54	81.67±14.70	95.00±15.81
D10	65.83±05.12	77.08±11.17	76.81±10.94	84.03±11.66	79.17±10.35	98.75±3.95
D11	65.83±05.12	81.81±16.74	75.56±11.53	86.81±11.61	77.78±22.65	98.75±3.95
D12	67.08±05.71	81.67±12.10	75.56±09.91	81.81±15.66	74.17±15.16	98.75±3.95
Average:	68.02±1.65	80.51±3.60	76.27±3.59	83.45±2.52	73.72±3.94	98.02±1.35

rate of the SELwSVM, as discussed in follows. For convenience and ease of reference, dataset names are assigned to different experimental configurations e.g., D1 to D12.

The performance of LDA_WSMGR, in terms of all the reported average performance measures, is better than all of the other feature extraction strategies, as can be observed from Figs. 3 and 4, and Tables 2 and 3. In fact, WSMGR (without using any feature transformation

Table 3 Performance of feature extraction strategies over different resolutions of ROIs using tenfold cross validation based on Az. (normal vs. masses)

Dataset	Az.					
	SMGR	WSMGR	PCA_MGR	PCA_WSMGR	LDA_MGR	LDA_WSMGR
D1	0.083±0.180	0.350±0.201	0.260±0.336	0.439±0.259	0.237±0.349	0.927±0.232
D2	0.033±0.105	0.420±0.388	0.489±0.292	0.499±0.384	0.473±0.236	0.980±0.063
D3	0.083±0.180	0.538±0.247	0.562±0.257	0.679±0.298	0.140±0.252	0.933±0.211
D4	0.280±0.124	0.577±0.280	0.663±0.214	0.633±0.249	0.439±0.333	0.980±0.063
D5	0.270±0.271	0.708±0.242	0.403±0.297	0.710±0.251	0.220±0.253	0.967±0.105
D6	0.264±0.277	0.582±0.203	0.563±0.353	0.624±0.255	0.377±0.269	0.953±0.148
D7	0.067±0.141	0.687±0.199	0.394±0.303	0.704±0.204	0.477±0.313	1.000±0.000
D8	0.033±0.105	0.527±0.454	0.476±0.179	0.587±0.378	0.388±0.314	0.953±0.148
D9	0.170±0.228	0.573±0.197	0.338±0.279	0.655±0.295	0.602±0.328	0.927±0.232
D10	0.000±0.000	0.598±0.196	0.459±0.230	0.652±0.247	0.563±0.177	0.980±0.063
D11	0.000±0.000	0.641±0.334	0.464±0.286	0.729±0.251	0.635±0.348	0.967±0.105
D12	0.033±0.105	0.532±0.276	0.413±0.198	0.539±0.349	0.370±0.356	0.967±0.105
Average:	0.11±0.11	0.56±0.10	0.46±0.11	0.62±0.09	0.41±0.15	0.96±0.02

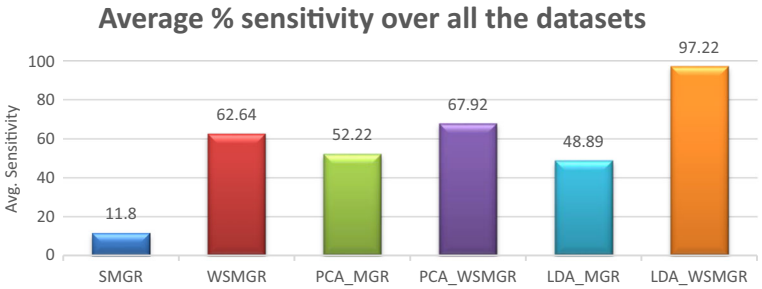

Fig. 3 Performance of feature extraction strategies over all the datasets (D1-D12) using tenfold cross validation based on sensitivity (normal vs. masses)

strategy) is better than PCA_MGR and LDA_MGR, keeping in view the average accuracy and average Az. values across all the datasets. SMGR perform very poorly with (**11.80±12.01**) average sensitivity and (**0.11±0.11**) average Az. value and therefore cannot be considered as a robust and reliable method for false positive reduction problem. The best case results are highlighted for each algorithm in Tables 2 and 3. We can see that LDA_WSMGR obtained accuracy of (100 %) with (**Az.=1**) for D7 which corresponds to the experimental configuration of 64×64 ROI resolution, 16×16 WSMGR block size and the bank with 24 filters denoted as GS4O6. This shows the importance of considering the distinguishing power of features based on the statistics of available class labels associated with the data samples (as used for LDA). The average performance of LDA is observed to be consistent (based on all performance measures) and always give best results.

4.3 Discrimination of benign and malignant

This section summarizes the results for another difficult classification problem i.e., the discrimination between benign and malignant masses. The discrimination task is relatively hard in this case due to highly identical patterns and similar structures of the two classes (benign and malignant) present in the selected digital mammograms. Figures 5 and 6 shows the comparison of the methods based on sensitivity and specificity. For the experimental case (benign vs. malignant), best average percentage accuracy of (**100.00±0.00**) and best average Az. value of (**1.000±0.000**) corresponds to a number of different configurations of input

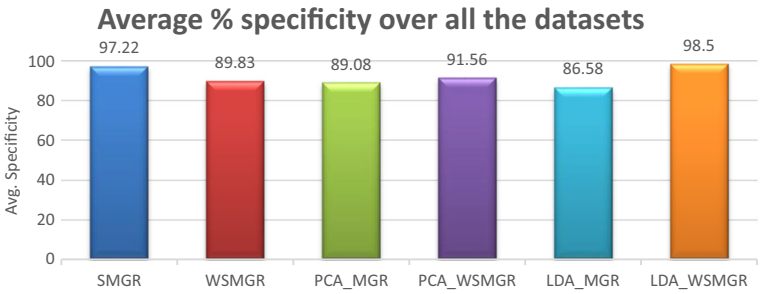

Fig. 4 Performance of feature extraction strategies over all the datasets (D1-D12) using tenfold cross validation based on specificity (normal vs. masses)

Table 4 Performance of feature extraction strategies over different resolutions of ROIs using tenfold cross validation based on accuracy (benign vs. malign)

Dataset	Accuracy					
	SMGR	WSMGR	PCA_MGR	PCA_WSMGR	LDA_MGR	LDA_WSMGR
D1	72.67±12.35	69.00±21.03	77.33±24.33	74.33±15.72	79.67±14.18	98.00±6.32
D2	70.67±16.61	69.67±19.47	77.00±17.32	76.67±14.14	78.00±17.44	96.00±12.65
D3	74.33±19.94	74.33±15.72	74.67±11.67	77.67±24.40	74.33±18.33	98.00±6.32
D4	72.33±19.63	74.00±16.39	75.00±14.25	78.00±12.39	82.00±18.41	100.00±0.00
D5	72.33±24.60	78.00±11.24	76.00±17.55	87.67±16.11	74.00±14.38	94.00±18.97
D6	73.00±18.56	73.33±22.28	79.33±20.42	80.67±21.13	75.00±14.25	96.00±12.65
D7	78.67±17.79	72.67±12.35	81.67±14.68	80.33±16.51	83.67±15.75	96.00±12.65
D8	71.00±21.03	71.67±22.18	80.00±10.66	84.33±16.71	71.67±18.48	100.00±0.00
D9	75.00±14.25	72.00±15.96	77.67±18.13	74.33±24.29	71.33±14.50	98.00±6.32
D10	70.67±9.91	71.67±16.94	76.67±15.07	72.67±13.41	69.00±25.05	100.00±0.00
D11	77.00±10.71	72.00±13.90	76.67±17.98	74.00±18.91	69.67±12.91	96.00±12.65
D12	77.00±15.43	73.67±18.75	75.00±16.27	81.00±15.56	72.00±13.90	96.00±12.65
Average:	73.72±2.69	72.67±2.33	77.25±2.15	78.47±4.53	75.03±4.81	97.33±1.97

resolution size, block size and Gabor bank as shown in Tables 4 and 5 for LDA_WSMGR. LDA_WSMGR is again observed to be more accurate (in terms of all performance measures) and consistent as compared to its five competitors and thus recommended to be use for mass classification problem. All the algorithms (except LDA_WSMGR) have resulted in almost the same performance with slight differences based on average accuracy and Az. values, observed

Table 5 Performance of feature extraction strategies over different resolutions of ROIs using tenfold cross validation based on Az. (benign vs. malign)

Dataset	Az.					
	SMGR	WSMGR	PCA_MGR	PCA_WSMGR	LDA_MGR	LDA_WSMGR
D1	0.442±0.261	0.446±0.355	0.538±0.373	0.371±0.391	0.546±0.305	0.967±0.105
D2	0.333±0.333	0.471±0.328	0.500±0.312	0.463±0.250	0.488±0.336	0.933±0.211
D3	0.450±0.352	0.533±0.319	0.413±0.167	0.538±0.373	0.425±0.369	0.950±0.158
D4	0.471±0.357	0.425±0.284	0.408±0.339	0.471±0.244	0.688±0.306	1.000±0.000
D5	0.467±0.341	0.567±0.206	0.492±0.310	0.758±0.325	0.388±0.208	0.900±0.316
D6	0.350±0.412	0.533±0.401	0.571±0.376	0.658±0.330	0.350±0.337	0.933±0.211
D7	0.588±0.353	0.375±0.281	0.538±0.373	0.571±0.376	0.688±0.283	0.933±0.211
D8	0.479±0.369	0.350±0.412	0.579±0.197	0.721±0.326	0.338±0.408	1.000±0.000
D9	0.404±0.325	0.300±0.350	0.533±0.373	0.508±0.348	0.250±0.354	0.967±0.105
D10	0.308±0.283	0.463±0.221	0.383±0.393	0.379±0.367	0.538±0.355	1.000±0.000
D11	0.517±0.194	0.354±0.309	0.483±0.337	0.513±0.355	0.250±0.264	0.933±0.211
D12	0.517±0.311	0.442±0.351	0.496±0.303	0.588±0.332	0.275±0.381	0.900±0.316
Average:	0.44±0.08	0.44±0.08	0.49±0.06	0.54±0.12	0.44±0.16	0.95±0.04

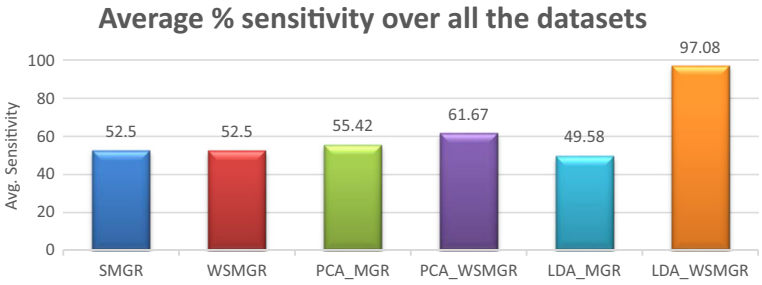

Fig. 5 Performance of feature extraction strategies over all the datasets (D1-D12) using tenfold cross validation based on sensitivity (benign vs. malign)

across all the datasets. It may please also be noted that the experimental configurations are playing an important role in achieving results with varying performance levels, e.g., the Az. value of PCA_WSMGR is 0.371±0.391 for D2 (32×32 ROI resolution, 8×8 WSMGR block size and the bank denoted as GS3O5) which get improves up to 0.758±0.325 using the same feature extraction strategy in case of D5 (64×64 ROI resolution, 16×16 WSMGR block size and the bank denoted as GS2O3).

We empirically described that the WSMGR method is better than MGR in terms of feature complexity and discrimination power for both the diagnosis cases (false positive reduction and discrimination of benign and malignant). Moreover, when WSMGR is compared with SMGR, both gives same performance for (benign vs. malign) case. However, the performance of SMGR is poor for (normal vs. masses) case and therefore this feature extraction strategy cannot be considered as robust method for mass classification problem (in general). WSMGR is a robust method for feature extraction and gives better results; it can be observed however that the performance of this method is still turned to be poor in terms of average sensitivity and Az. and therefore need further refinement. As a refinement of WSMGR, its variant LDA_WSMGR (that uses LDA transformation strategy) has significantly improved the results and achieved 100 % accurate recognition rate for both the diagnosis cases. In the next subsection, we elaborate about the statistical differences between the performance of feature extraction methods based on average accuracy and Az. values.

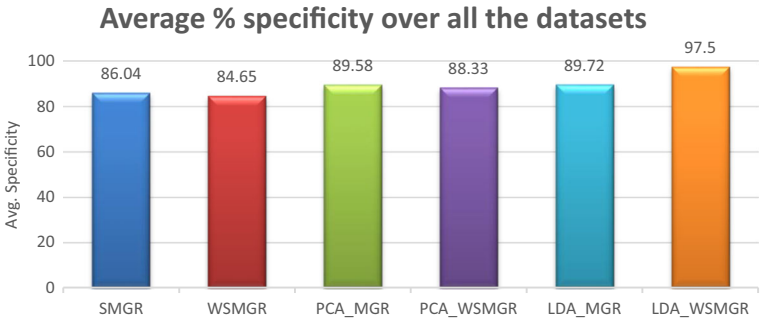

Fig. 6 Performance of feature extraction strategies over all the datasets (D1-D12) using tenfold cross validation based on specificity (benign vs. malign)

Table 6 Summary of the comparisons of the LDA_WSMGR feature extraction algorithm with the remaining algorithms according to the non-parametric Friedman test with the Holm's post-hoc test in terms of (i) Az. value and (ii) percentage accuracy

Comparison on the basis	Experimental case	Algorithm	Avg. rank	*p*	*Holm*
Accuracy	Normal vs. Masses	LDA_WSMGR (control)	1.0000	–	–
		PCA_WSMGR	2.0833	0.1560	0.05
		WSMGR	3.2916	0.0026	0.025
		PCA_MGR	4.1250	4.2843E-5	0.0166
		LDA_MGR	4.5416	3.5327E-6	0.0125
		SMGR	5.9583	8.4714E-11	0.01
	Benign vs. Malignant	LDA_WSMGR (control)	1.0000	–	–
		PCA_WSMGR	3.0000	0.0088	0.05
		PCA_MGR	3.0833	0.0063	0.025
		LDA_MGR	4.2083	2.6609E-5	0.0166
		SMGR	4.7500	9.1121E-7	0.0125
		WSMGR	4.9583	2.1872E-7	0.01
Az.	Normal vs. Masses	LDA_WSMGR (control)	1.0000	–	–
		PCA_WSMGR	2.0833	0.1560	0.05
		WSMGR	3.4166	0.0015	0.025
		PCA_MGR	3.9999	8.5682E-5	0.0166
		LDA_MGR	4.5833	2.7096E-6	0.0125
		SMGR	5.9166	1.2151E-10	0.01
	Benign vs. Malignant	LDA_WSMGR (control)	1.0000	–	–
		PCA_WSMGR	3.2916	0.0026	0.05
		PCA_MGR	3.8333	2.0750E-4	0.025
		SMGR	4.2500	2.0881E-5	0.0166
		LDA_MGR	4.2916	1.6339E-5	0.0125
		WSMGR	4.3333	1.2749E-5	0.01

4.4 Discussion based on statistical comparison

In this study, to test whether the LDA_WSMGR based feature extraction method performs significantly better than those of the other five competitors, a non-parametric statistical test (Friedman) is conducted. The Friedman test is chosen because it does not make any assumptions about the normal distribution of the underlying data (a requirement for equivalent parametric tests) and it is a recommended and suitable test to compare a set of classification strategies over multiple performance output values, according to the guidelines presented in [9, 16]. Table 6 presents the summary of the comparisons of the LDA_WSMGR feature extraction algorithm (the algorithm with the best average rank, considered as control algorithm) with the remaining algorithms used in our experiments according to the non-parametric Friedman test with the Holm's post-hoc test [9, 16] in terms of percentage accuracy and Az. values for all the datasets as given in Tables 2, 3, 4 and 5 for the two diagnosis cases.

For each algorithm, the average rank (the lower the average rank the better the algorithm's performance), the *p-value* (when the average rank is compared to the average rank of the algorithm with the best rank i.e., control algorithm, in our case, it is LDA_WSMGR) and *Holm*

critical value obtained by Holm's post-hoc test are reported. Based on the fact that the *p-value* is lower than the critical value (at 5 % significance level), entries in a row are shown in bold when there is a significant difference between the average ranks of an algorithm and the control algorithm (LDA_WSMGR). The rows containing bold entries indicate that the control algorithm has significantly outperformed the corresponding algorithms present in these rows.

According to the statistics of Table 6, LDA_WSMGR performs statistically significantly better than all of its competitors in terms of both percentage average accuracy and Az. values in all of the experiments presented in Tables 4 and 5 for benign vs. malignant classification problem. Almost, same is the situation when considering normal vs. masses case, that LDA_WSMGR is better than all of its competitors (except PCA_WSMGR) based on percentage average accuracy as well Az. values. When observed for normal vs. masses classification problem, PCA_WSMGR seem to perform comparatively equivalent to LDA_WSMGR based on a large *p-value* for both the performance measures. Interestingly, based on average ranking for both the performance measures, WSMGR is placed at 3rd position for normal vs. masses case but for benign vs. malignant case, WSMGR performs the worst and thus placed at the last position. The difference between the performance of LDA_WSMGR as compared to SMGR is very significant based on smallest *p-values* for normal vs. masses case. With these statistics, LDA_WSMGR can be considered as attractive choice for the problems being targeted. LDA_WSMGR improve the performance of the proposed system to extreme level and also reduces the feature dimension, significantly (as compared to MGR), which is very helpful to cope with the problem known as *curse of dimensionality* and offer better *generalization* ability for a classification scheme; such as SELwSVM.

4.5 Comparison with other methods

It is rather difficult to meaningfully compare the proposed method with other methods in the literature due to many factors. For example, which mammogram database was used for

Table 7 Comparison with state-of-the-art methods based on average Acc. and Az values

Problem	Research work	Database	No. of ROIs	Avg. Acc. (%)	Avg. Az.
Normal vs. Masses	Geraldo B. J et al. [22] (2009)	DDSM	584	99.39(max)	1.00(max)
	X. Liado et al. [24] (2009)	DDSM	512	–	0.94±0.02
	Fatemeh et al. [27] (2010)	MIAS	90	85.9±0.03	–
	Daniel et al. [7] (2011)	DDSM	5090	90.07	–
	Ioan B. et al. [21] (2011)	MIAS	322	84.37	0.79
	Reyad et al. [34] (2014)	DDSM	512	98.63(max)	–
	Oliveira et al. [30] (2015)	DDSM	3404	98.88(max)	–
	Our Method	MIAS	109	100±0.00	1.00±0.00
Benign vs. Malignant	Geraldo B. J et al. [22] (2009)	DDSM	584	88.31	0.89
	Fatemeh et al. [27] (2010)	MIAS	90	87.00±0.008	–
	Daniel et al. [7] (2011)	DDSM	3240	84.22	–
	Ioan B. et al. [21] (2011)	MIAS	114	78.26	0.78
	Hussain [19] (2014)	DDSM	512	85.53±5.43	0.87±0.05
	Our Method	MIAS	109	100±0.00	1.00±0.00

evaluation? Given that the same database was employed, were the same sample of mammograms selected for evaluation? How many samples were used? Which evaluation approach (validation methodology, training and testing set formation with different percentages of ROIs) was used? Were the ratios of ROIs for different classes (e.g., normal, malignant and benign) the same? Even if other methods are implemented and evaluated on the same dataset, it might still not be a fair comparison because the tuning of parameters involved in different methods are not necessarily the same.

In any case, to give a general trend of the performance of our method (LDA_WSMGR) and compare it with state-of-the-art methods in terms of accuracy and Az., we have compiled information from various studies as shown in Table 7. The quantities that are not reported in the literature are indicated with a dash symbol. For some methods, standard deviation values are not available. For the two problems (i.e., normal vs. masses and benign vs. malignant), only the best case mean and standard deviation results are reported for all the methods being compared. For the first problem (normal vs. masses), the proposed method performs better than all the other methods. It may please be noted that some entries in the Table 7 represent the maximum (max) accuracy and Az. values as reported in their paper; the mean and standard deviation of each measure are not given.

For the second problem (benign vs. malignant), the proposed method also outperforms all reported methods. In general, the proposed method performs better than state-of-the-art techniques for the two classification problems. Note that Ioan et al. [21], Daniel et al. [7] and Hussain [45] also used Gabor filter banks for the description of masses but their descriptors are different; in the first two methods, the descriptors are global while in the third method, the descriptor is local.

5 Conclusions

In this research article, we have discussed and compared six different directional feature extraction methods for mass classification problem. These methods use a bank of Gabor filters to extract the features from textural micro-patterns (present in the ROIs) at different scales and orientations. The features extracted based on LDA_WSMGR feature extraction strategy are shown to best discriminates between the three tissue types (normal, benign and malign masses) used in the experiments and in general, improves the recognition rate of a breast cancer detection system, up to a significant level.

The comparison based on Friedman statistical test reveals that the LDA_WSMGR method is actually statistically significantly better than its competitors on 5 % significance level. All of the feature extraction methods are evaluated over ROI images extracted from MIAS database, using an application oriented fitness function based on successive enhancement learning based weighted Support Vector Machine (SELwSVM) to cater with the skewed/ unbalanced dataset problem. Two state of the art feature transformation algorithms have reduced the dimension of feature space, remarkably. With compact data space, recognition rate of cancerous tissues in the digital mammograms has improved. Model compactness indirectly implies that the feature space will be low dimensional and thus better computational efficiency and better generalization of the classification model is expected and observed.

The methods are empirically analyzed over two diagnosis cases i.e., discrimination between: (i) normal but suspicious and masses (malignant and benign) and (ii) benign and malignant masses. For the two diagnosis cases, we achieved encouraging results, reported as

(percentage mean accuracy, mean area under the ROC value over tenfold cross validation): i.e., (100 %, 1.00 for normal vs. masses) and (100 %, 1.00 for benign vs. malignant) based on LDA_WSMGR. It can be observed that LDA_WSMGR has the potential to be further explored in more complex recognition tasks related to breast cancer detection problem. LDA_WSMGR is shown to outperform other state-of-the-art methods available in the literature and thus offer promising capabilities.

There are several future avenues in order to extend the LDA_WSMGR technique. It will be interesting to investigate the performance of LDA_WSMGR method in more complex problem scenarios e.g., recognition and identification of more breast abnormalities like micro-calcification, breast structural disorder etc. The preprocessing of mammogram images for enhancing their quality is also an area that is required to be further investigated. LDA_WSMGR method is required to be tested over noisy mammogram images in order to investigate its robustness power. Other optimization strategies e.g., Genetic Algorithm, Cuckoo optimization are worth enough to be investigated for Gabor filter parameter optimization in the targeted area.

Acknowledgments This project was supported by NSTIP strategic technologies programs, grant number 08-INF325-02 in the Kingdom of Saudi Arabia.

References

1. Alam RN et al (2009) Computer-aided mass detection on digitized mammograms using a novel hybrid segmentation system. Int J Biol Biomed Eng 3(4):51–58
2. Altekruse SF, Kosary CL, Krapcho M et al (2010) SEER Cancer Statistics Review, 1975–2007. National Cancer Institute, Bethesda
3. Bhangale T, Desai UB, Sharma U (2000) An unsupervised scheme for detection of microcalcifications on mammograms. Proc. IEEE Int Conf Image Proc. Vancouver, BC, Canada, pp. 184–187
4. Bhangale T, Desai UB, Sharma U (2000) An unsupervised scheme for detection of microcalcifications on mammograms. IEEE Int Conf Image Proc 184–187
5. Boser BE, Guyon IM, Vapnik V (1992) A training algorithm for optimal margin classifiers. In Proc. of the fifth annual workshop on Computational learning theory 144–152
6. Burges C (1998) Tutorial on support vector machines for pattern recognition. Data Min Knowl Disc 2(2): 955–974
7. Costa DD, Campos LF, Barros AK (2001) Classification of breast tissue in mammograms using efficient coding. Bio-Medical Engineering, On-Line, 2011, 10:55, http://www.biomedical-engineering-online.com/content/10/1/55
8. Daugman JG (1980) Two-dimensional spectral analysis of cortical receptive field profiles. Vision Res 20: 847–856
9. Demˇsar J (2006) Statistical comparisons of classifiers over multiple data sets. Mach Learn Res 7:1–30
10. Domínguez AR, Nandi AK (2009) Towards breast cancer diagnosis based on automated segmentation of masses in mammograms. Pattern Recogn 42(6):1138–1148
11. Duda RO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. Wiley, New York
12. El-Naqa I, Yang Y, Wernick M, Galatsanos N, Nishikawa R (2002) A support vector machine approach for detection of microcalcifications. IEEE Trans Med Imaging 21(12):1552–1563
13. Elter M, Horsch A (2009) CADx of mammographic masses and clustered micro calcifications: a review. Med Phys 36(6):2052–2068
14. Esteve J, Kricker A, Ferlay J, Parkin D (1993) Facts and figures of cancer in the European Community. In: Tech. Rep., International Agency for Research on Cancer
15. Fisher RA (1936) The use of multiple measures in taxonomic problems. Ann Eugen 7:179–188
16. García S, Herrera F (2008) An extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all pairwise coparisons. Mach Learn Res 9:2677–2694

17. Grigorescu S, Petkov N, Kruizinga P (2002) Comparison of texture features based on Gabor filters. IEEE Trans Image Proc 11(10):1160–1167
18. Hsu CW, Chang CC, Lin CJ (2010) A practical guide to support vector classification. Technical report, Department of Computer Science and Information Engineering, National Taiwan University
19. Hussain M (2014) False positive reduction in mammography using multiscale spatial weber law descriptor and support vector machines. Neural Comput Appl 25(1):83–93, Springer-Verlag
20. Hussain M, Khan S, Muhammad G, Mohamed B, Bebis G (2012) Mass detection in digital mammograms using gabor filter bank. IET Image Proc 1–5
21. Ioan B, Gacsadi A (2011) Directional features for automatic tumor classification of mammogram images. Biomed Signal Process Control 6(4):370–378
22. Junior GB et al (2009) Classification of breast tissues using Moran's index and Geary's coefficient as texture signatures and SVM. Comput Biol Med 39:1063–1072
23. Lahmiri S, Boukadoum M (2011) Hybrid discrete wavelet transform and gabor filter banks processing for mammogram features extraction. Proc. NEWCAS, France. IEEE Comput Soc 53–56
24. Lladó X, Oliver A, Freixenet J, Martí R, Martí J (2009) A textural approach for mass false positive reduction in mammography. Comput Med Imaging Graph 33(6):415–422
25. Mammographic Image Analysis Society, http://www.wiau.man.ac.uk/services/MIAS/MIASweb.html
26. Manjunath BS, Ma WY (1996) Texture features for browsing and retrieval of image data. IEEE Trans Pattern Anal Mach Intell 18(8):837–842
27. Moayedi F et al (2010) Contourlet-based mammography mass classification using the SVM family. Comput Biol Med 40:373–383
28. Mohamed ME, Ibrahima F, Brahim BS (2010) Breast cancer diagnosis in digital mammogram using multiscale curvelet transform. Comput Med Imaging Graph 34(4):269–276
29. Nunes AP, Silva AC, de Paiva AC (2010) Detection of masses in mammographic images using geometry, Simpson's diversity index and SVM. Int J Signal Imaging Syst Eng 3(1):43–51
30. Oliveira FSS, Filho AOC, Silva AC, Paiva AC, Gattass M (2015) Classification of breast regions as mass and non-mass based on digital mammograms using taxonomic indexes and SVM. Comput Biol Med 57(1): 42–53
31. Oliver A, Freixenet J, Martí J et al (2010) A review of automatic mass detection and segmentation in mammographic images. Med Image Anal 14(2):87–110
32. Peter K, Nikolay P (1999) Nonlinear operator for oriented texture. IEEE Trans Image Process 8(10):1395–1407
33. Rangayyan RM, Ferrari RJ, Desautels JEL, Frère AF (2000) Directional analysis of images with Gabor wavelets. Proc. XIII Braz Symp Comput Graphics Image SIBGRAPI 170–177
34. Reyad YA, Berbar MA, Hussain M (2014) Comparison of statistical, LBP, and multi-resolution analysis features for breast mass classification. J Med Syst 38:100. doi:10.1007/s10916-014-0100-7
35. Rogova GL, Stomper PC, Ke C (1999) Microcalcification texture analysis in a hybrid system for computer aided mammography. Proc SPIE 1426–1433
36. Sampaio WB, Diniz EM, Silva AC, Paiva AC, Gattass M (2011) Detection of masses in mammogram images using CNN, geostatistic functions and SVM. Comput Biol Med 41:653–664
37. Székely N, Tóth N, Pataki B (2006) A hybrid system for detecting masses in mammographic images. IEEE Trans Instrum Meas 55(3):944–952
38. Tang J, Rangayyan RM, Xu J et al (2009) Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. IEEE Trans Inform Technol Biomed 13(2):236–251
39. Turk M, Pentland A (1991) Eigenfaces for recognition. J Cogn Neurosci 3:71–86
40. Turner MR (1986) Texture discrimination by Gabor functions. Biol Cybern 55:71–82
41. Vapnik V (1995) Statistical learning theory. Springer, New York
42. Wang Y, Gao X, Li J (2007) A feature analysis approach to mass detection in mammography based on RF-SVM", ICIP07 9–12
43. Wei D, Chan H, Helvie M, Sahiner B, Petrick N, Adler D, Goodsitt M (1995) Classification of mass and normal breast tissue on digital mammograms: multiresolution texture analysis. Med Phys 22(9):1501–1513
44. Yu S, Shiguan S, Xilin C, Wen G (2009) Hierarchical ensemble of global and local classifiers for face recognition. IEEE Trans Image Process 18(8):1885–1896
45. Yufeng Z (2010) Breast cancer detection with gabor features from digital mammograms. Algorithms 3(1): 44–62
46. Zehan S, George B, Ronald M (2006) Monocular Precrash vehicle detection: features and classifiers. IEEE Trans Image Process 15(7):2019–2034

Salabat Khan received the BS. degree in computer science from Virtual University of Pakistan in 2007, the MS and PhD degrees in computer science from FAST - National University of Computer and Emerging Sciences in 2009 and 2015, respectively. He is currently an Assistant Professor in the Department of Computer Science, Comsats Institute of Information Technology. His current research interests include data mining, pattern analysis, medical image processing, bio-informatics and all kinds of bio-inspired algorithms (mainly, evolutionary and swarm based algorithms and in particular, ant colony algorithms).

Muhammad Hussain is an Associate Professor in the Department of Computer Science, King Saud University, Saudi Arabia. He received both his M.Sc. and M. Phil., from University of the Punjab, Lahore, Pakistan, in 1990 and 1993 respectively. In 2003, He received a Ph.D. in Computer Science, specializing in Computer Graphics from Kyushu University, Fukuoka, Japan. He worked as a researcher at Japan Science and Technology Agency from April 2003 to September 2005. In September 2005, he joined King Saud University as an Assistant Professor. He worked on a number of funded projects in Kyushu University, Japan and King Saud University, Saudi Arabia. His current research interests include multiresolution techniques in Computer Graphics and Image Processing.

Hatim Aboalsamh is a professor in the Department of Computer Science, King Saud University, Riyadh, Saudi Arabia. He is former dean of College of Computer & Information Sciences. He was also the Vice President for development and Quality of King Saud University. He is an IT consultant and a fellow member of British Computer Society (BCS), a senior member of ACM, and a member of many other professional societies. He has over 70 referred publications and co-editor of two books.

George Bebis received his BS degree in Mathematics from the University of Crete in 1987, the MS degree in Computer Science from the University of Crete in 1991, and the PhD degree in Electrical and Computer Engineering from the University of Central Florida in 1996. He is currently a Professor in the Department of Computer Science and Engineering at the University of Nevada, Reno (UNR), director of UNR's Computer Vision Laboratory (CVL), and Visiting Professor at King Saud University. Prior to joining UNR, he was a Visiting Assistant Professor in the Department of Mathematics and Computer Science at the University of Missouri-St. Louis (UMSL).