Computers in Biology and Medicine 39 (2009) 1063 – 1072

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Classification of breast tissues using Moran's index and Geary's coefficient as texture signatures and SVM

Geraldo Braz Junior, Anselmo Cardoso de Paiva, Aristófanes Corrêa Silva*, Alexandre Cesar Muniz de Oliveira

Federal University of Maranhão - UFMA, Applied Computing Group NCA/UFMA, Av. dos Portugueses, SN, Campus do Bacanga, Bacanga, 65085-580 São Luís, MA, Brazil

ARTICLE INFO

Article history:
Received 2 June 2008
Accepted 24 August 2009

Keywords:
Mammography
Breast tissue classification
Moran's index
Geary's coefficient
Support vector machine

ABSTRACT

Female breast cancer is the major cause of cancer-related deaths in western countries. Efforts in computer vision have been made in order to help improving the diagnostic accuracy by radiologists. In this paper, we present a methodology that uses Moran's index and Geary's coefficient measures in breast tissues extracted from mammogram images. These measures are used as input features for a support vector machine classifier with the purpose of distinguishing tissues between normal and abnormal cases as well as classifying them into benign and malignant cancerous cases. The use of both proposed techniques showed to be very promising, since we obtained an 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 to discriminate tissues in mammograms as normal or abnormal. We also obtained 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 to discriminate tissues in mammograms as benign and malignant.

1. Introduction

Breast cancer is the major cause of cancer-related deaths among the female population. It is known that the best prevention method is the precocious diagnosis, what lessens the mortality and improves the treatment [1]. According to the American National Cancer Institute [2], it is estimated that every 3 min a woman is diagnosed with breast cancer and every 13 min, a woman dies from this disease. Mammography is currently the best technique for reliable detection of early, non-palpable, potentially curable breast cancer [1].

With the advances of computer technology, radiologists have an opportunity to improve their interpretation of mammograms using computer capabilities that can enhance the image quality of mammograms. Over the past two decades, many attempts have been made by computer scientists to assist the radiologists in detection and diagnosis of masses by developing computer-aided tools for mammography interpretation. Image processing and intelligent systems are two mainstreams of computer technologies that have been constantly explored in the development of computer-aided mammography systems. Generally, these systems are classified into two categories: computer-aided detection (CAD) and computer-aided diagnosis (CADx) [3]. CAD and CADx systems can aid radiologists by providing a second opinion and may be used in the first stage of examination in a near future, allowing the reduction of the variance among radiologists in the interpretation of mammograms.

Our research group has been investigating for some time the use of spatial statistics as texture descriptor. These techniques are largely used in spatial data analysis. Our research started applying spatial statistics to computerized tomography images for lung cancer nodule diagnose as benign and malignant [4–6].

Later, with the promising results found in those works, we looked for investigating the applicability of the same techniques for classifying breast tissues using mammogram images [7–9]. In Braz et al. [7] we used the same geostatistical techniques used in the present work, but with a more specific approach, because we wanted to investigate only the capability of classification of breast tissues as normal or abnormal in multiresolution images. In the present work, we focused on looking into other aspects of the breast tissue. Here, the previous work is extended including the classification of breast tissue as mass and non-mass (normal and abnormal tissues) using different image quantizations and also, the analysis and classification of normal tissue as benign and malignant. The geostatistical measures are used in the methodology as texture descriptors to obtain information that is imperceptible to the human eye, which can furnish specific information about the region under study. Based on the achieved results we believe that the proposed methodology is very promising.

* Corresponding author. Tel.: +55 98 33018234; fax: +55 98 33018841.
E-mail addresses: ge.braz@gmail.com (G. Braz Junior), paiva@deinf.ufma.br (A. Cardoso de Paiva), ari@dee.ufma.br (A. Corrêa Silva), acmo@deinf.ufma.br (A. Cesar Muniz de Oliveira).

0010-4825/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2009.08.009

Besides, we demonstrate that Geary's coefficient and Moran's index are texture descriptors with great potential to be explored.

This work is organized as follows. In Section 2 we present some related works. Section 3 presents the techniques for feature extraction, classification and validation. Section 4 presents a detailed description about the proposed methodology and evaluation. Next, in Section 5, the results are shown and we discuss about the application of the techniques under study. Finally, Section 6 presents some concluding remarks.

2. Related works

Many methodologies have been proposed to solve the problem providing assistance on the precocious cancer detection and diagnosis tools.

In [10], it was presented a methodology that uses independent component analysis (ICA) along with support vector machine (SVM) and linear discriminant analysis (LDA) to distinguish between mass or non-mass and benign or malign breast tissues in mammograms. As a result, it was found the following: LDA reaches 89.5% of accuracy when discriminating mass or non-mass and 95.2% when discriminating benign or malignant in the DDSM database and, in MIAS database, they obtained 85% when discriminating mass or non-mass and 88% when discriminating benign or malignant; SVM reaches 99.6% of accuracy when discriminating mass or non-mass and 99.5% when discriminating benign or malignant in DDSM database and in MIAS database we obtained 97% when discriminating mass or non-mass and 100% when discriminating benign or malignant.

Verma et al. [11] presents a new neural network technique for the classification of suspicious areas in digital mammograms. The proposed neural network technique was tested on the DDSM database and obtained accuracy of 94%.

A generalized dynamic fuzzy neural network (GDFNN) approach was used in Lim and Er [12] to classify breast tissues with accuracy of 84.4% using the DDSM database.

Oliver et al. [13], proposed a strategy based on the adaptation of the eigenfaces approach to the problem of detecting masses. Thus, they introduced the concept of eigenROIs, which span the ROI subspace of the original image space. The result of this transformation was a vector of weights describing the contribution of each eigenROI to represent the corresponding input image. They proposed the use of these vectors to construct the models for the training step. The work used one set of 160 regions of interest (ROIs) extracted from the MIAS database (40 of them were masses and the rest were normal tissue) and 196 ROIs containing masses and 392 with normal but suspicious regions obtained from the DDSM database. In Oliver et al. [14] the same authors extended their previous method by using the 2DPCA method [37] instead of the standard PCA technique, improving the performance of the false positive reduction.

The work of Tourassi et al. [15] is based on comparing a new ROI with all the ROIs in the database (template-based approach). The two clearest differences between them arise from the similarity measure and the database used. More specifically, the former developed a likelihood measure which depends on the gray-level and the shape of the ROIs. Both parameters were compared with the new ROI and the set of ROIs present in the database, which was only composed by ROIs depicting masses. From this comparison a likelihood measure was computed. The work consists in comparing all the ROIs of the database (including ROIs with and without masses) with the new one using mutual information based on similarity measure. Thus, the new ROI was labeled as belonging to the closest class. Note that with the methods based on the template-based strategy, the similarity measure used for classifying the ROIs has to be re-computed for each new element, as it measures the difference between the new ROI and all the ROIs in the database.

Varela et al. [16] proposed a strategy based on extracting gray-levels and morphologic features, and training a neural network (NNet) used to classify the new ROI. Results of FROC analysis for the test set indicate that the proposed algorithm can achieve a TP rate of 88% at 1.02 FPs/image.

In Lladó et al. [17] it is proposed the use of local binary patterns (LBP) for representing the textural properties of the masses, extending the basic LBP histogram descriptor into a spatially enhanced histogram which encodes both the local region appearance and the spatial structure of the masses. The work also uses a support vector machine (SVM) to separate the true masses from the ones which are actually normal parenchyma. The approach was evaluated using 1792 ROIs extracted from the DDSM database.

In Dominguez and Nandi [18] it is proposed a work using a set of images selected from the mini-MIAS and the DDSM database. The work proposed the extraction of six features from the ROIs for characterization of mass margins (contrast between the foreground region and the background region, coefficient of variation of edge strength, two measures of the fuzziness of mass margins, a measure of spiculation based on the relative gradient orientation, and a measure of spiculation based on edge-signature information). These features were used with three popular classifiers (Bayesian classifier, Fisher's linear discriminant and support vector machine) were used to predict the diagnosis of a set of 349 masses based on each of said features and some combinations of these. The Fisher discriminant analysis produced a sensibility of approximately 0.65, with a specificity in the range of 0.65–0.8. The experiment made with the Bayesian classifier or the SVM achieved a sensibility ranging from approximately 0.50 to 0.55, with a specificity in the range of approximately 0.7–0.9.

In Mangasarian et al. [19] a methodology is proposed to perform the nuclear analysis of tissues in breast nodules (fine needle aspirates) through linear programming techniques for studying the diagnosis and prognosis. The diagnosis is performing by extracting the following characteristics, using the Xcyt system: area, radius, perimeter, symmetry, number and size of concavities, fractal dimension (of the boundary), compactness, smoothness (local variation of radial segments) and texture (variance of gray levels inside the boundary). Applying the multisurface method (MSM) for classification of tissues into malignant and benign, resulting in 97% of accuracy.

There are also three works of our research group which are related to the method proposed in this article and which also show the evolution of researches on this matter.

In Braz et al. [7] it is proposed a methodology to distinguish mass and non-mass tissues on mammograms. It is based on the computation of geo-statistical measures (Moran's index and Geary's coefficient) over a multidimensional image representation through wavelet transform. The computed measures are classified through a support vector machine (SVM). The methodology reaches 98.36% of specificity, 98.13% of sensitivity and an accuracy of 98.24% when discriminating mass from non-mass elements, applying Geary's coefficient.

In Jr et al. [8], it is achieved accuracy up to 88% when discriminating breast masses into benign and malignant through extracted geo-statistical characteristics using MIAS database.

In Oliveira Martins et al. [9], Ripley's K function was used together with SVM to classify samples of benign and malignant tissues. In that article, the author introduces the idea of using Ripley's K function in concentric rings. This method obtained sensibility of 94% and accuracy of 94.25% using the DDSM database.

3. Background

In this section, the techniques used for developing the method proposed in this work are exposed. They are: spatial texture analysis, support vector machine and validation of classification methods.

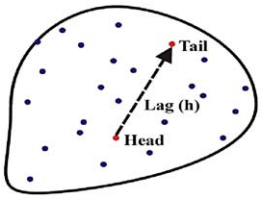

Fig. 1. Usual variables in spatial statistics calculus.

3.1. Spatial texture measures

Spatial statistics are quantitative studies about spatially continuous data in space [20]. It treats geographic attributes as random variables which depend on joined distributions and their locations. The degree to which close neighbors over a surface share a similar property is characterized by spatial autocorrelation. Herein these statistics focus on the process that occurs in space and the applied methods try to describe and analyze the environment [21].

The goal of the spatial approach is to measure the spatial association between a pair of observations separated by a certain distance. Fig. 1 exemplifies these approaches. The pair of observations, head and tail, are points separated by a distance and direction vector, lag(h). We typically assume that spatial autocorrelation does not depend on the location of the pair of observations, but only on the distance between both observations, and possibly on the direction of their relative displacement.

Head is a reference point (origin pixel) and tail is a target point (target pixel). Note that one head point could be analyzed through many tail points that the experiment might need. In our simple example, the head point is analyzed only with a certain tail point situated at a certain distance and direction represented by the lag(h) vector. This way, we can measure the existing autocorrelation between the head point and the set of tail points. The spatial statistics quantify the strength of these associations through similarity analysis of spatially distributed features. In our work we intend to analyze several combinations of pairs of pixels (several origin pixels in relation to several target pixels) given some distances and some azimuths.

In this situation, observations separated in space by a certain spatial distance lag(h) have similar values (correlationship). The objective of statistics is to measure the degree of spatial association among the observations of one or more variables. The spatial autocorrelation can be negative or positive. It will be positive when the fact observed in some place is also observed in its neighbors separated by a certain distance. The measurement can also assume null values, situation by which we can prove the inexistence of spatial correlation.

In the statistical context, texture can be described in terms of two main components associated to pixels: variability and spatial autocorrelation. The advantage of the usage of spatial statistics techniques is that both aspects can be measured together, as we will discuss in the next sections. These measurements describe the texture obtained from a certain image through the degree of spatial association present in the geographically referenced elements of the image described in the next section.

3.1.1. Moran's index and Geary's coefficient

Moran's index and Geary's coefficient summarize the strength of associations between responses as a function of distance, and possibly direction [22]. We typically assume that spatial autocorrelation does not depend on where a pair of observations is located, but only on the distance between both observations, and possibly on the direction of their relative displacement.

Moran's index is applied to zones or points which have continuous variables associated with their intensities. The statistic is used to compare the value of the variable x_i of one location with the value at all other locations x_j. It is formally defined by

$$I_h = \frac{N\sum_i\sum_j w_{ij}(x_i-\bar{X})(x_j-\bar{X})}{(\sum_i\sum_j w_{ij})\sum_i(x_i-\bar{X})^2} \quad (1)$$

where h is the lag (vector), x_i (head) is the variable value at a particular location i, x_j (tail) is the variable value at another location ($i \neq j$), N is the number of cases in the analysis that depend on the specific choice of a reference point, x_i, $\bar{X}$ is the mean of the pixel value of the referenced points and w_{ij} is a weight applied to the comparison between location i and location j. More properly speaking, w_{ij} is a distance-based weight which is the inverse distance between locations i and j, i.e. $w_{ij} = 1/d_{ij}$.

In our work N is the total number of pixels in the breast tissue image. In other words, for a certain lag(h) and azimuth there is a quantity of pixels analyzed in the image. This quantity is the N number. $\bar{X}$ represents the mean of gray level pixels in a certain lag(h) and azimuth in relation to a reference pixel. lag(h) indicates the coverage radius of the analysis. The x_i and x_j variables represent the pair of pixels to be analyzed, which depends on specific lag(h) and azimuth. That is, all of the pixels in the breast image will be analyzed within a distance(lag) and direction (azimuth). The shortest distance (lag) between i and j is 1, and the longest possible distance (lag) is the size of the breast tissue currently under analysis. In our work, however, the maximum distance (lag) was limited to 10. The same idea is applied to Geary's coefficient.

Moran's index has values that typically range from approximately $+1$, representing complete positive spatial autocorrelation, to approximately -1, representing complete negative spatial autocorrelation [23].

The Geary's coefficient assumes that the interaction is not the cross product of the deviations from the mean, but the deviation in intensities of each observation location with another one [24]. Its formal definition is

$$C(h) = \frac{(N-1)\sum_i\sum_j w_{ij}(x_i-x_j)^2}{2(\sum_i\sum_j w_{ij})\sum_i(x_i-\bar{X})^2} \quad (2)$$

where the values of x_i, x_j, N, $\bar{X}$ and w_{ij} are analogous to those of Moran's index.

The values of C typically vary between 0 and 2 [23]. The theoretical value of C as 1, indicates that values of one zone are spatially unrelated to the values of any other zone. Values less than 1 (between 0 and 1) indicate positive spatial autocorrelation while values greater than 1 indicate negative autocorrelation.

This coefficient does not provide the same information about spatial autocorrelation given by Moran's index. It emphasizes the differences in values between pairs of compared observations rather than the co-variation between the pairs. So, Moran's index gives a more global indicator, whereas the Geary's coefficient is more sensitive to differences in small neighborhoods.

When computing experimental directional indexes of spatial autocorrelation, it is defined the direction vector, azimuth, which corresponds to an angle in x plane. Other parameters used for spatial autocorrelation index calculations—such as lag space, lag(h) tolerance, angular tolerance and maximum bandwidth—are exemplified in Fig. 2.

The tolerance associated with the lag(h) is called lag tolerance. In practice, for a specified direction, the indexes may be computed for a number of lags. The tolerance associated to a direction is referred to as the angle tolerance. These components together determine a cone which can be constrained by the bandwidth factor. The bandwidth controls the maximum width across the cone and allows it to focus more on the specified direction.

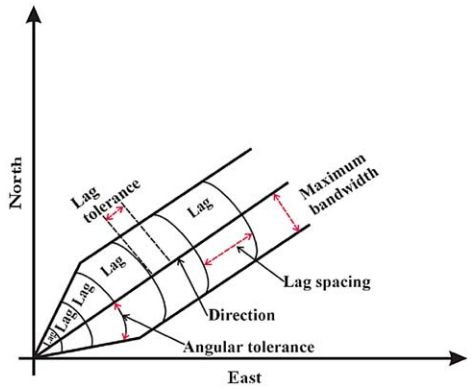

Fig. 2. Parameters used for indexes of spatial autocorrelation calculations.

3.2. Support vector machine

Support vector machine (SVM), introduced by V. Vapnik in 1995, is a method to estimate the data classification function [25]. The basic idea of SVM is to construct a hyperplane as the decision surface in such a way that the margin of separation between positive and negative examples is maximized.

The SVM term comes from the fact that the points in the training set which are closer to the decision surface are called support vectors. SVM achieves this by the structural risk minimization principle that is based on the fact that the error rate of a learning machine on the test data is bounded by the sum of the training-error rate and a term that depends on the Vapnik–Chervonenkis (VC) dimension [26].

Statistical data classification could be mapped into a problem of one, two or multiple classes. In two-class classification, the data from two classes are available. The dataset is supposedly composed of equally balanced class samples. An unbalanced dataset could lead to poor results [27]. A common problem with this approach is that the decision boundary created by two-class SVM could make a large misclassification rate if class samples are not easily separable.

The process starts with a training set of points $x_i \in \Re^n$, $i=1,2,\ldots,l$ where each point x_i belongs to one class identified by the label $y_i \in \{-1,1\}$. The goal of maximum margin classification is to separate the two classes by a hyperplane such that the distance to the support vectors is maximized.

The construction can be thought as follows: each point x in the input space is mapped to a point $z=\Phi(x)$ of a higher dimensional space, called the feature space, where the data are linearly separated by a hyperplane. The nature of data determines how the method proceeds. There are data that are linearly separable, nonlinearly separable and with impossible separation. The key property in this construction is that we can write our decision function using a kernel function $K(x,y)$ which is given by the function $\Phi(x)$ that maps the input space into the feature space. Such decision surface has the equation

$$f(x)=\sum_{i=1}^{l}\alpha_i y_i K(x,x_i)+b \tag{3}$$

where $K(x,x_i)=\Phi(x)\cdot\Phi(x_i)$, and the coefficients α_i and the variable b are the solutions of a convex quadratic programming problem [28], namely

$$\begin{aligned}&\min_{w,b,\xi} \quad \frac{1}{2}w^T\cdot w + C\sum_{i=1}^{l}\xi_i\\ &\text{s.t.} \quad y_i[w^T\cdot\phi(x_i)+b]\geq 1-\xi_i, \quad \xi_i\geq 0.\end{aligned} \tag{4}$$

where $C>0$ is a parameter to be chosen by the user, which corresponds to the strength of the penalization errors and the ξ_i's are slack variables that penalize training errors.

3.3. Validation of the classification methods

In order to evaluate the classifier with respect to its discriminatory ability, we have analyzed its sensitivity (*Se*), specificity (*Sp*) and accuracy (*Ac*). The sensitivity is defined by

$$Se=\frac{Tp}{Tp+Fn} \tag{5}$$

where *Tp* is the number of tissues correctly classified as abnormal and *Fn* is the number of tissues wrongly classified as abnormal. Sensitivity measures the performance of the method when recognizing abnormal tissues.

The specificity is defined by

$$Sp=\frac{Tn}{Tn+Fp} \tag{6}$$

where *Tn* is the number of normal tissues correctly classified and *Fp* is the number of abnormal tissues wrongly classified as normal. Specificity measures the performance of the method while recognizing normal tissues.

The accuracy is expressed by the overall rate of correctly and wrongly classified tissues:

$$Ac=\frac{Tp+Tn}{Tp+Tn+Fp+Fn} \tag{7}$$

Also, we calculate Matthews correlation coefficient [29] typically used in machine learning as a measure of the quality of binary (two-class) classifications expressed by

$$Mcc=\frac{(Tp*Tn)-(Fp*Fn)}{\sqrt{(Tp+Fp)*(Tp+Fn)*(Tn+Fp)*(Tn+Fn)}} \tag{8}$$

where *Mcc* more closely of +1 value represents a perfect prediction, 0 an average random prediction and −1 an inverse prediction.

In order to evaluate the ability of the classifier to differentiate benign from malignant nodules, the area (Az) under the ROC (receiver operation characteristic) [30] curve was used. In other words, the ROC curve describes the ability of the classifiers to correctly differentiate the set of lung nodule candidates into two class, based on the true-positive fraction (sensitivity) and false-positive fraction (1-specificity).

4. Proposed method

This section presents the method proposed in this study case which intends to classify breast tissues of mammogram images initially into normal and abnormal cases. Then, the correctly classified abnormal cases are classified again as benign or malignant. This methodology is based on six steps: image acquisition, enhancement breast tissue (histogram equalization), sample quantization, spatial texture analysis, classification (two-class SVM) and validation (sensitivity, specificity, accuracy, Matthews correlation coefficient and ROC curve). Fig. 3 shows the steps of the proposed method.

4.1. Image acquisition

For the development and evaluation of the proposed methodology, we used a publicly available database of digitized screen-film mammograms: the digital database for screening mammography (DDSM) [31]. It contains 2620 cases acquired from Massachusetts

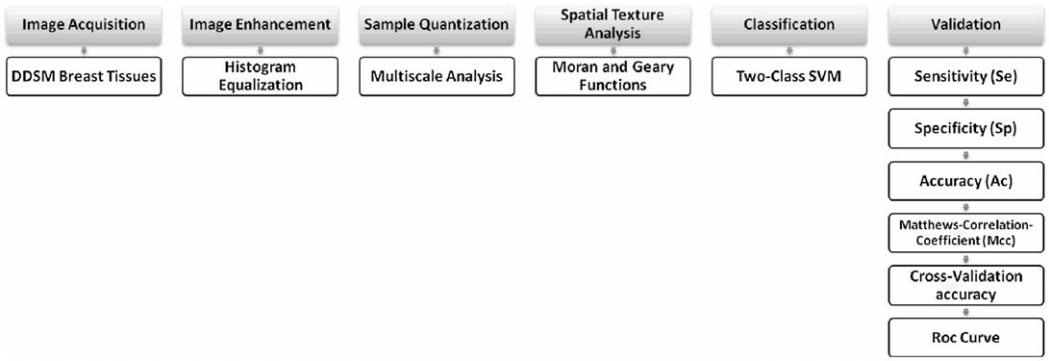

Fig. 3. Fluxogram for the proposed method.

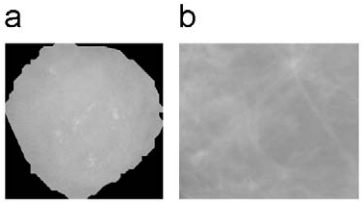

Fig. 4. Regions extracted from the DDSM mammogram base: (a) abnormal tissue and (b) normal tissue.

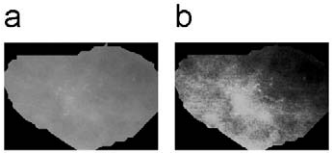

Fig. 5. (a) Original abnormal tissue and (b) abnormal tissue after equalization.

General Hospital, Wake Forest University, and Washington University in St. Louis School of Medicine. The data are composed of studies of patients from different ethnic and racial backgrounds. The DDSM contains descriptions of mammogram lesions in terms of the American College of Radiology breast imaging lexicon called the breast imaging reporting and data system (BI-RADS). Mammograms in the DDSM database were digitized by different scanners depending on the institutional source of data.

In order to validate our methodology, we used a sub-sample of 1394 breast tissues from the mammogram images, having been used 584 samples of abnormal tissues(mass), among which 273 were malignant and 311 were benign. These samples were identified by specialists according to information from DDSM, and their edges delimited by the same criterion. We used the information of edge delimitation in a text file (called overlay) to extract the mass tissues that will compose our sample set. This text file also provides the characteristics of the lesion. We still used 810 breast tissues classified as normal (non-mass). The extraction of these tissues was performed manually, that is, we choose regions that were not identified as abnormal. Fig. 4 exemplifies the normal and abnormal regions extracted from the mammograms which were used in our sample to validate the methodology.

4.2. Enhancement breast tissue

In order to improve the image quality for the recognition step, we performed a global histogram equalization [32] with all obtained tissues (mass or abnormal and non-mass or normal). This method increases the contrast between objects of the image by better distributing gray level concentrations. An abnormal tissue is exemplified in Fig. 5(a) and (b) demonstrates the same abnormal tissue after the equalization step.

4.3. Sample quantization

With the equalized image, we made six versions of the sample through quantization of the tonality range ($2^8, 2^7, 2^6, 2^5, 2^4, 2^3$). Our goal is to extract texture information for different tonality resolutions, making possible to codify information that could be otherwise omitted. I.e. the calculations for Geary's coefficient and Moran's index in each new ROI which is quantized in a specific gray tone.

4.4. Spatial texture analysis

Each of them, after the quantization step, was submitted to a spatial texture analysis using Moran's index and Geary's coefficient. This analysis is done for each pixel of the sample. In order to improve the capacity of describing texture patterns, we conduced the directional experimental indexes of spatial autocorrelation, defining a direction vector, an azimuth which corresponds to an angle in plane x, assuming a lag increasing rate, tolerances for lag and azimuth to better capture the neighborhood in a matrix arrangement of pixels. These restrictions are exemplified in Fig. 2.

In practice, for a specified direction, the indexes may be computed for a number of lags. These components together determine a cone which can be constrained by the bandwidth factor. The bandwidth controls the maximum width across the cone and allows it to focus more on the specified direction.

We calculated Moran's index and Geary's coefficient taking each pixel at a time as the reference (head) point. Therefore, each pixel is, in a certain moment, a reference to be examined for a specific combination of lag distance, azimuth value and lag/angular tolerance. The value obtained by the spatial function for this combination of parameters will be an aggregation of the analysis of each pixel as head point. The analysis guides, for each generated sample, a feature vector containing the characteristic information of the pattern in this area.

The features extracted from the breast tissues (abnormal and normal tissue), considered as texture signatures were obtained for a set of four directions, corresponding to azimuth values equal to 0°, 45°, 90° and 135°. We adopted $\pm 22.5^\circ$ as the tolerance for angle measurements. The initial lag separation distance (h) was 1. After, this lag was incremented in 1 unity. The maximum number of chosen lags was 10. The tolerance lag adopted was ± 0.45. We make this analysis for six levels of quantization, as previously described, so that the number of characteristics extracted for each sample is 10

lags × 4 directions × 6 tonality resolutions (quantization), totalizing 240 features for each sample and function.

4.5. Classification

Each feature vector was classified by two-class SVM. These two versions needs a kernel function responsible for classification of a new data point x of Eq (3). An important family of kernel functions is the radial basis function, more commonly used for pattern recognition problems, which has been used in this paper, and is defined by

$$K(x,y) = e^{-\gamma\|x-y\|^2} \quad (9)$$

where $\gamma > 0$ is also user-defined. This parameter is used to normalize the units between the feature spaces of the SVM.

The proposed methodology needs a two-step classification. The first step discriminates regions into normal and abnormal breast tissues. With all abnormal regions classified, the second step tries to discriminate them into benign and malignant breast tissues.

For both steps, the set of feature vectors is split into two parts: training and test, through a random choice. In first step, the training and test sets are formed by normal and abnormal feature vectors. In the second step, the training and test sets are formed by abnormal tissues classified later, here treated as benign and malignant samples.

For the step 1, 30% of the set was used to train an SVM. The remaining samples (70%) were used as a test. We choose this configuration trying to avoid a small number of samples in next step since samples that will be used in step 2 come from the resulting classification of step 1. We also perform many empirical tests to analyze that different percentage configurations did not represent statistical improvement.

In step 2, to recognize abnormal tissues as benign and malignant, we train an SVM classifier with 80% of the samples classified as abnormal groups in last step. The other half of samples was used for test. To avoid random results, we perform this step 50 times and do some statistical analysis like standard deviation, mean and median.

We use the library for support vector machines (LIBSVM) [33] as an implementation of SVM. For each training set used, we estimate the parameters used by SVM in the classification. The C parameter of Eq. (3) and the parameter γ of Eq. (9) were estimated using the grid function provided by LIBSVM tool.

For the first step, parameters C and γ used were 32 768 and 0.007813 using Geary's texture measures and 0.03125 and 0.0078125 using Moran's texture measures. For the second step parameters C and γ used were 128 and 0.5 using Geary's texture measures and 32 768 and 0.001953 using Moran's texture measures.

4.6. Validation

Finally, with all results obtained by the classification step we perform the validation of the performance achieved by the proposed method using the rates of sensitivity (*Se*), specificity (*Sp*), accuracy (*Ac*), Matthews correlation coefficient (*Mcc*) and the accuracy of the cross-validation (Cross) with 10-fold for each prediction to verify the correctness of the created model. Also we analyze the performance of the proposed methodology using the area (Az) under the ROC.

5. Results

This section intends to investigate the efficiency of using Moran's index, Geary's coefficient, and SVM for classification of breast tissues in digital mammograms.

In order to evaluate the discrimination power of the measures extracted from spatial texture function, we plot the measures extracted from the sample breast tissues enumerated in Fig. 6.

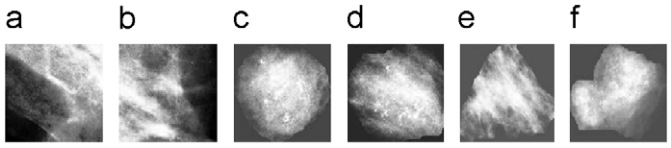

Fig. 6. ROI obtained from DDSM after equalization: (a and b) normal tissues and (c–f) abnormal tissues. Also (c and d) are benign and (e and f) are malignant.

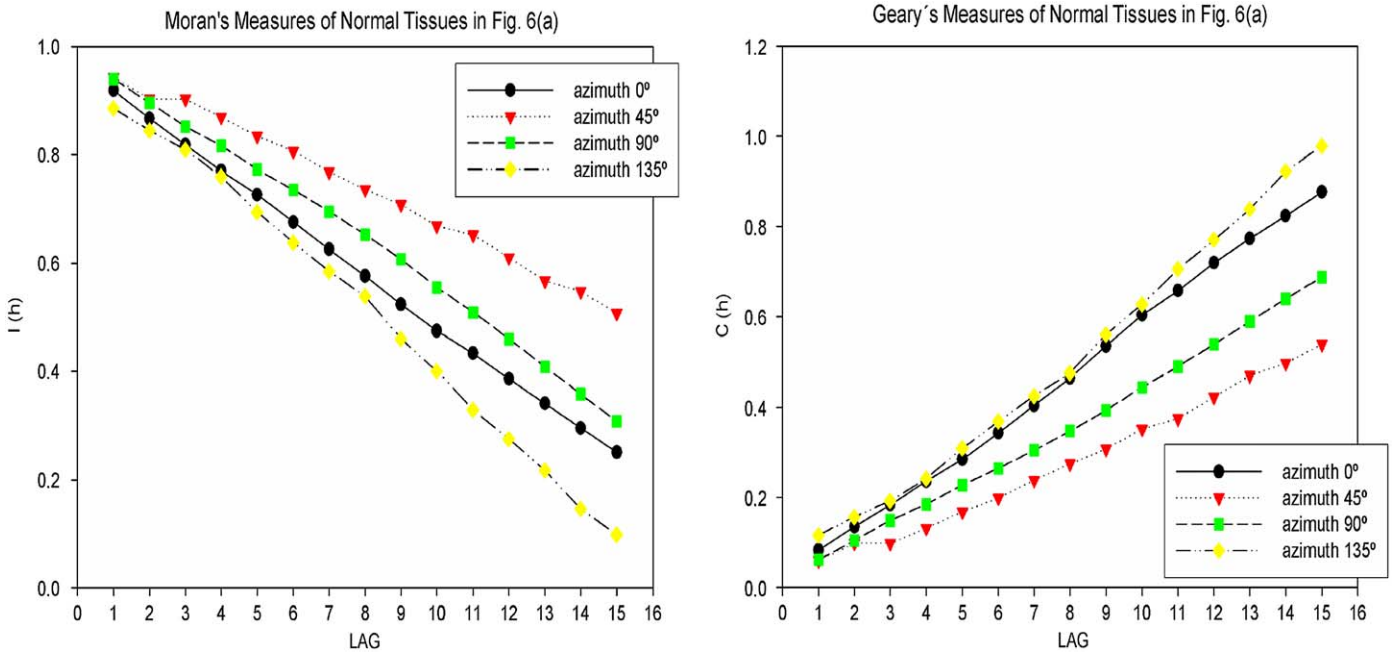

Fig. 7. Geary's coefficient and Moran's index applied to normal tissue.

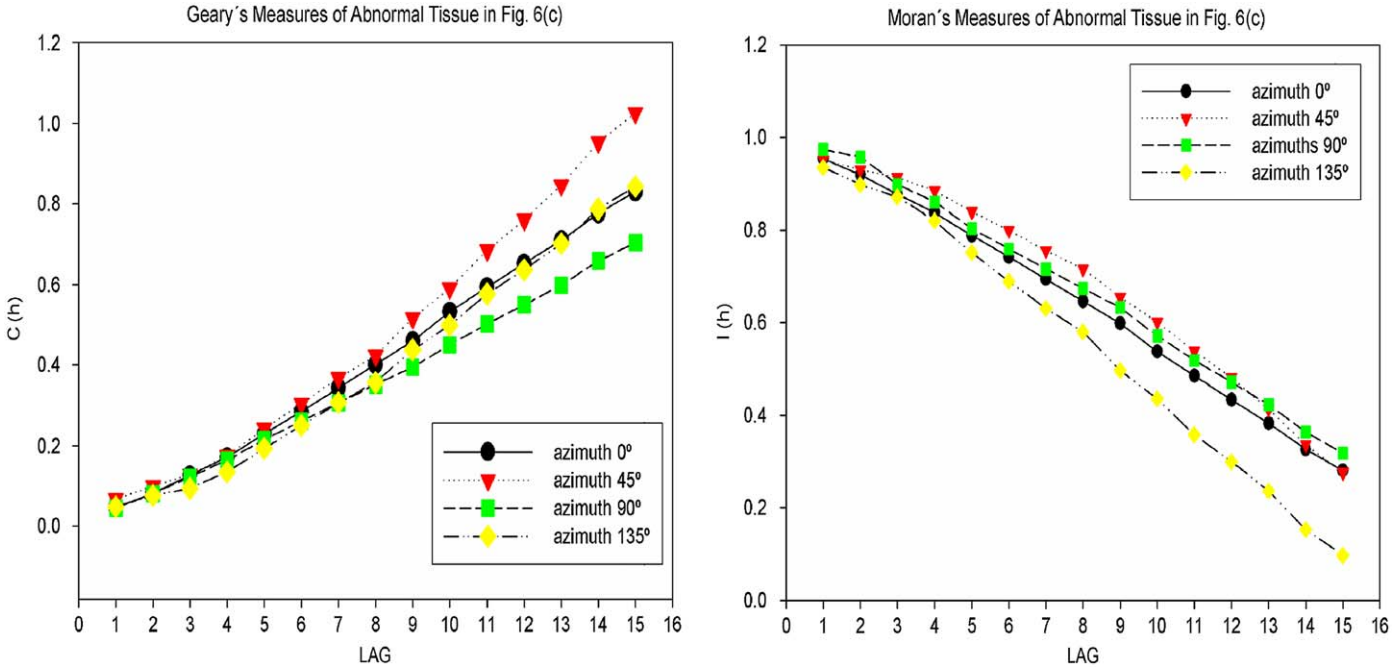

Fig. 8. Geary's coefficient and Moran's index applied to abnormal tissue.

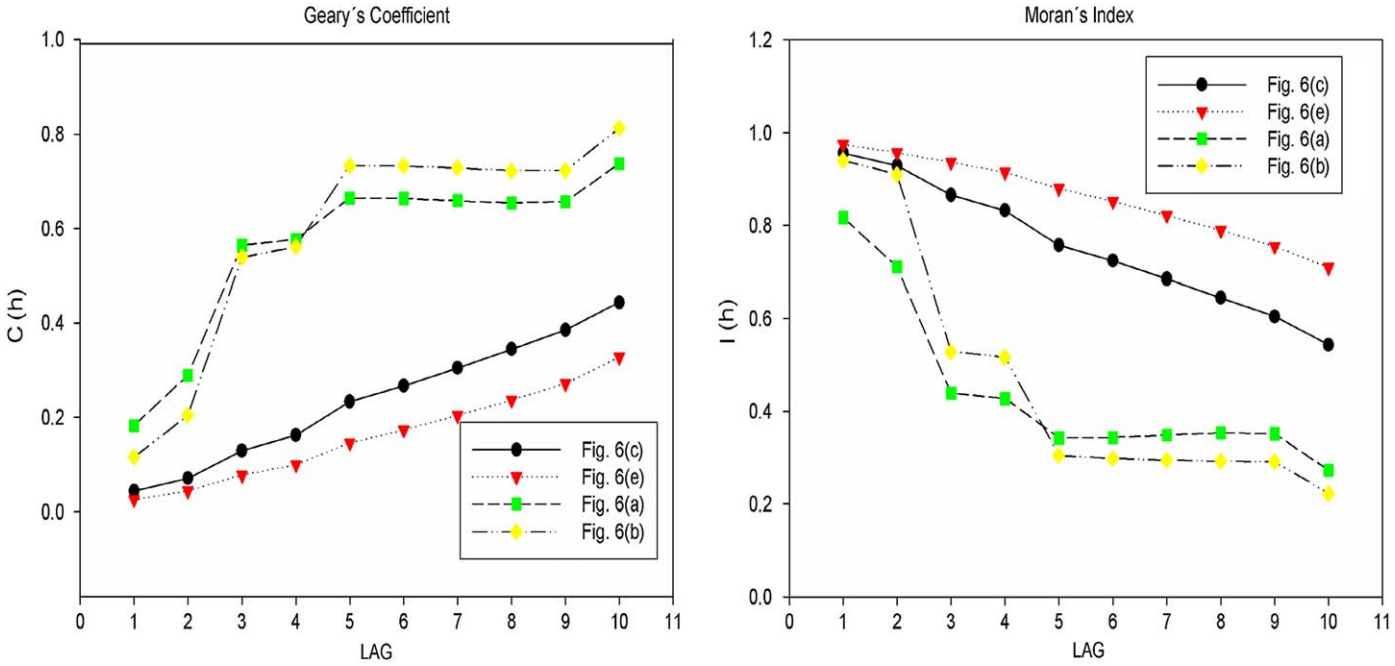

Fig. 9. Geary's coefficient and Moran's index applied to two samples of normal and abnormal tissues.

Figs. 7 and 8 exemplify, respectively, the application of Geary's coefficient and Moran's index to a normal and an abnormal region, respectively, represented by Fig. 6(a) and (c). The curves show the computed spatial autocorrelation for all specified directions, at varying distances (lag). We can conclude from the results that all functions present isotropic characteristics in the undertaken tests, meaning that only one curve (one direction) is needed to represent the signature of each class of tissues.

Fig. 9 shows the application of Geary's coefficient and Moran's index to two samples of normal and abnormal tissues given in Fig. 6(a)–(c) and (e). We verify that abnormal tissues have typically a continuous falling for Moran's measures, and continuous growth for Geary's measures. On the other hand, for normal tissues, we observe much more dispersion in the curves. The graphic analysis

Table 1
Analysis of accuracy in the classification of breast tissues in mammograms into normal and abnormal.

Function	*Tp*	*Tn*	*Fp*	*Fn*	*Se* (%)	*Sp* (%)	*Ac* (%)	*Mcc*	Mean accuracy Cross-validation (%)
Geary	384	562	10	29	92.98	98.25	96.04	0.92	99
Moran	408	562	6	0	100	98.94	99.39	0.99	100

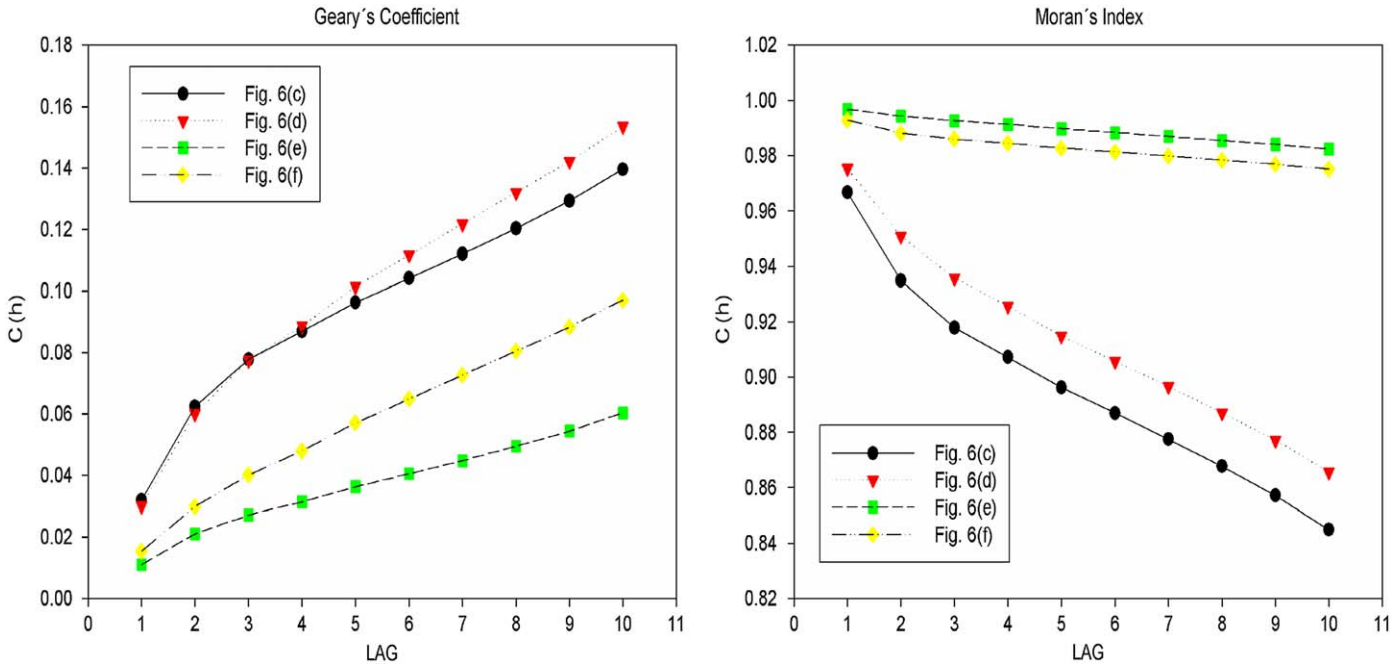

Fig. 10. Geary's coefficient and Moran's index applied to two samples of benign and malignant breast tissues.

Table 2
Analysis of accuracy in the classification of abnormal breast tissues in mammograms into benign and malignant.

Function	*Tp*	*Tn*	*Fp*	*Fn*	*Se* (%)	*Sp* (%)	*Ac* (%)	*Mcc*	Mean accuracy Cross-validation (%)
Geary	39	29	2	7	84.78	93.55	88.31	0.77	77.52
Moran	34	38	4	6	85.00	90.48	87.80	0.76	75.46

shows the presence of greater dispersion between the regions in normal rather than in abnormal cases.

Table 1 shows the results obtained for this classification step. It is important to stress that this table presents the results of the test of the model with 70% (975 out of 1394 breast tissues) of the whole sample, since the remaining 30% (415 out of 1394 breast tissues) were used for training the model. Based on the Az ROC, we have observed that the Az for Geary's coefficient and Moran's index were 0.946 and 1, respectively.

Fig. 10 shows the application of Geary's coefficient and Moran's index to two samples of benign and malignant breast tissues represented by Fig. 6(c)–(f). We verify that for benign tissues, Geary's coefficient has more accentuated growth rate than in malignant tissues and Moran's index has more decreasing rate for benign tissues than for malignant ones. The graphic analysis shows the presence of a separation between the measures extracted for benign and malignant tissues using these two functions.

Table 2 shows the results obtained while classifying abnormal tissues into benign and malignant classes. In this step, 80% of the breast tissues classified as abnormal were used (Table 1) for both Geary's coefficient (308 out of 384 tissues classified as abnormal) and Moran's index (328 out of 408 tissues classified as abnormal). The remaining 20% of the sample was used for training the model. Based on the Az ROC, we have observed that the Az for Geary's coefficient and Moran's index were 0.804 and 0.89, respectively (Fig. 11). Also, Table 3 presents a more detailed statistic analysis of the 50 runs executed for this step.

Comparing this and the various works in this area in detail is a hard task since: (1) they use databases that are different from the one used in this work; (2) they use the same database as this work (DDSM), but it is impossible to know which cases are used for training and testing; (3) they use a specific database that is not publicly available and (4) they use different evaluation metrics (Az ROC, sensibility, specificity, accuracy, etc.).

Despite of these considerations, we compared some works based on accuracy or on Az ROC, which are two common metrics among the cited works.

We also included two tables of comparisons with the works cited in Section 2. Table 4 compares our methodology with the works of classification of breast tissues into mass and non-mass. Table 5 compares our methodology with works of classification of breast tissues into benign and malignant.

We may observe that the work proposed herein achieves a result comparable to the best result published in the recent literature for normal and abnormal breast tissue classification, as depicted in Table 4. The Az ROC is greater than the majority of the observed related works and the accuracy is in the range of the best methods, what indicates that this is a promising methodology that must be better investigated to obtain more conclusive works.

6. Conclusion

This paper has presented a pair of spatial texture functions with the purpose of characterizing breast tissues using a public database of mammograms (DDSM). The results of discriminating breast tissue patterns obtained with the features extracted by Moran's index and Geary's coefficient and evaluated with a support vector machine classifier had good performance for distinguishing abnormal from normal and benign from malignant tissues.

We verified that extracting characteristics from the sample images by applying these two functions provide accuracy to classification of normal and abnormal breast tissues above 96% and 99% for Geary's and Moran's measures and Az ROC reaches 0.946 and 1, respectively. The specificity found was above 98% for each function, indicating a successful discrimination power while separating normal regions from abnormal regions. The value of sensitivity had a rate of 92.98% using Geary's coefficient (sensitivity for Moran's

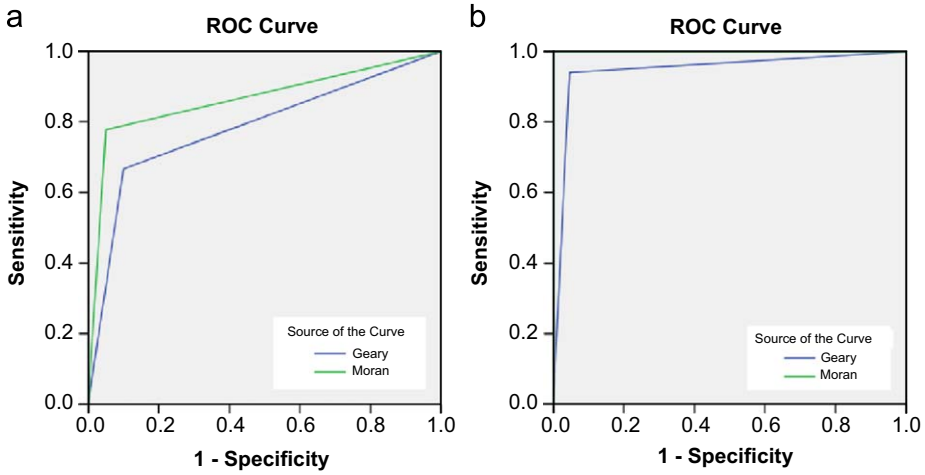

Fig. 11. Az ROC curve for proposed methodology: (a) discriminating normal and abnormal and (b) discriminating benign and malignant.

Table 3
More statistical analysis in the classification of abnormal breast tissues in mammograms into benign and malignant.

Function	*Se* (%)	*Sp* (%)	*Ac* (%)	*Mcc*	Mean accuracy Cross-validation (%)
Geary					
Min	69.23	80.56	80.77	0.61	74.84
25%-quartile	75.61	85.45	82.05	0.64	79.04
Mean	78.05	89.96	83.37	0.68	80.16
Std deviation	4.68	5.69	1.96	0.04	2.09
Median	78.38	89.61	82.66	0.66	80.59
75%-quartile	80.82	93.89	85.01	0.72	81.46
Max	86.84	100.00	88.31	0.77	83.77
Moran					
Min	67.44	78.38	79.27	0.58	75.46
25%-quartile	75.00	83.72	80.49	0.61	77.67
Mean	78.08	86.71	82.25	0.65	78.89
Std deviation	4.97	4.47	2.58	0.05	1.77
Median	77.57	86.48	81.71	0.63	78.68
75%-quartile	82.33	89.94	84.15	0.70	79.75
Max	87.18	94.87	87.80	0.75	83.13

Table 4
Comparison of some works regarding the classification of breast tissues into normal × abnormal.

Works	Database	Accuracy (%)	Az ROC
Braz et al. [7]	DDSM	98.24	–
Jr et al. [8]	MIAS	88	–
Verma et al. [11]	DDSM	94	–
Costa et al. [10]	DDSM	85	–
	MIAS	88	–
Lim and Er [12]	DDSM	84.4	–
Oliver et al. [13]	MIAS		0.83
Oliver et al. [14]	DDSM		0.83
Tourassi et al. [15]	Private		0.81
Varela et al. [16]	Private		0.90
Lladó et al. [17]	DDSM		0.94
Our method (Geary's coefficient)	DDSM	96.04	0.94
Our method (Moran's index)	DDSM	99.39	1

index was 100.00%) indicating a performance that may be increased for abnormal pattern discrimination with this function.

Classifying abnormal cases as benign and malignant, the methodology reaches accuracy of 88.31% and 87.80% for Geary's and Moran's function, respectively. The Az ROC reaches 0.804 for Gearys's coefficient and 0.89 for Moran's index indicating a very similar classifier performance. The rates obtained for sensitivity were 84.78% and 85.00% for each function indicating a high precision on correct recognition of malignant abnormal breast tissue. Also, the rates obtained by specificity demonstrate the effectiveness of the spatial texture measures to discriminate patterns in abnormal breast tissues as benign and malignant cancerous cases.

Table 5
Comparison of some works regarding the classification of abnormal breast tissues into benign × malignant.

Works	Database	Accuracy (%)	Az ROC
Oliveira Martins et al. [9]	DDSM	94.94	–
Costa et al. [10]	DDSM	99.6	–
	MIAS	97	–
Gorgel et al. [34]	Private	84.8	–
Jr et al. [8]	MIAS	88	–
Our method (Geary's coefficient)	DDSM	88.31	0.80
Our method (Moran's index)	DDSM	87.80	0.89

Based on these results, we have observed that such measures provide significant support to a more detailed clinical investigation and the results were very encouraging, in special when tissues were classified with support vector machines. Nevertheless, there is the need to perform tests with a larger database and more complex cases in order to obtain a more precise behavior pattern.

Despite the good results obtained only by analyzing the texture, further information can be obtained by analyzing the shape. Shape is a very good characteristic extracted from mammography tissues which is useful for discrimination in abnormal cases. As future work, we propose a combination of texture and shape measures for a more precise and reliable classification.

Conflict of interest statement

There is no conflict of interest.

Acknowledgments

The authors acknowledge CAPES, CNPq and FAPEMA for financial support.

References

[1] A.C.S. (AMS), Learn about breast cancer, 2009, available at ⟨http://www.cancer.org⟩.

[2] N.C.I. (NCI), Cancer stat fact sheets: cancer of the breast, 2009, available at ⟨http://seer.cancer.gov/statfacts/html/breast.html⟩.
[3] A. Bovik, J. Gibson, Handbook of Image and Video Processing, Academic Press, Inc., Orlando, FL, USA, 2000.
[4] A. Silva, M. Gattass, P. Carvalho, Analysis of spatial variability using geostatistical functions for diagnosis of lung nodule in computerized tomography images, Pattern Analysis and Applications 7 (3) (2004) 227–234.
[5] A. Silva, A. Paiva, P. Carvalho, M. Gattass, Semivariogram and SGLDM methods comparison for the diagnosis of solitary lung nodule, in: Second Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2005), Lecture Notes in Computer Science, vol. 3523, Springer, Berlin, 2005, pp. 479–486.
[6] E. Silva, A. Silva, A. de Paiva, R. Nunes, Diagnosis of lung nodule using Moran's index and Geary's coefficient in computerized tomography images, Pattern Analysis and Applications 11 (1) (2008) 89–99.
[7] Braz júnior, Geraldo, E.C., Silva, A.C., Paiva, Silva, Aristófanes Corrêa. Breast Tissues Classification Based on the Application of Geostatistical Features and Wavelet Transform. In: Information Technology Applications in Biomedicine, 2007. ITAB 2007. 6th International Special Topic Conference on, 2007, Tokyo. IEEE Proceedings, 2007, pp. 227–230.
[8] V.R.S. Jr, A.C.d. Paiva, A.C. Silva, A.C.M. Oliveira, Semivariogram applied for classification of benign and malignant tissues in mammography, in: Lecture Notes in Computer Science, Springer, Berlin, 2006, pp. 570–579.
[9] L. Oliveira Martins, E.C. Silva, A.C. Silva, A.C. Paiva, M. Gattass, Classification of breast masses in mammogram images using Ripley's K function and support vector machine, in: MLDM '07: Proceedings of the Fifth International Conference on Machine Learning and Data Mining in Pattern Recognition, Springer, Berlin, Heidelberg, pp. 784–794, doi: http://dx.doi.org/10.1007/978-3-540-73499-4_59 (Independent component analysis in breast tissues mammograms images classification using LDA and SVM).
[10] D. Costa, L. Campos, A. Barros, A. Silva, Independent component analysis in breast tissues mammograms images classification using LDA and SVM, 2007, pp. 231–234, doi: 10.1109/ITAB.2007.4407389.
[11] B. Verma, P. McLeod, A. Klevansky, A novel soft cluster neural network for the classification of suspicious areas in digital mammograms, Pattern Recognition 42 (9) (2009) 1845–1852 ISSN 0031-3203, doi: 10.1016/j.patcog.2009.02.009.
[12] W.K. Lim, M.J. Er, Classification of mammographic masses using generalized dynamic fuzzy neural networks, Medical Physics 2009 (5) (2004) 1288–1295 doi: 10.1118/1.1708643, http://link.aip.org/link/?MPH/31/1288/1.
[13] A. Oliver, J. Marti, R. Marti, A. Bosch, J. Freixenet, A new approach to the classification of mammographic masses and normal breast tissue, in: International Conference on Pattern Recognition, vol. 4, 2006, pp. 707–710, ISSN 1051-4651, doi: http://doi.ieeecomputersociety.org/10.1109/ICPR.2006.113.
[14] A. Oliver, X. Lladó, J. Martí, R. Martí, J. Freixenet, False positive reduction in breast mass detection using two-dimensional PCA, in: IbPRIA '07: Proceedings of the Third Iberian Conference on Pattern Recognition and Image Analysis, Part II, Springer, Berlin, Heidelberg, 2007, ISBN 978-3-540-72848-1, pp. 154–161, doi: http://dx.doi.org/10.1007/978-3-540-72849-8_20.
[15] G.D. Tourassi, B. Harrawood, S. Singh, J.Y. Lo, C.E. Floyd, Evaluation of information-theoretic similarity measures for content-based retrieval and detection of masses in mammograms, Medical Physics 34 (1) (2007) 140–150 doi: 10.1118/1.2401667, http://link.aip.org/link/?MPH/34/140/1.
[16] C. Varela, P.G. Tahocesb, A.J. Mendez, M. Souto, J.J. Vidal, Computerized detection of breast masses in digitized mammograms, Computers in Biology and Medicine (2) (2007) 214–226.
[17] X. Lladó, A. Oliver, J. Freixenet, R. Martí, J. Martí, A textural approach for mass false positive reduction in mammography. Computerized Medical Imaging and Graphics, Elsevier, 2009.
[18] A.R. Dominguez, A. Nandi, Toward breast cancer diagnosis based on automated segmentation of masses in mammograms, Pattern Recognition 42 (2009) 1138–1148.
[19] O.L. Mangasarian, W.N. Street, W.H. Wolberg, Breast cancer diagnosis and prognosis via linear programming, Operations Research 43 (4) (1995) 570–577 doi: 10.1287/opre.43.4.570, http://or.journal.informs.org/cgi/content/abstract/43/4/570.
[20] I. Clark, Practical Geostatistics, Applied Science Publishers, London, 1979.
[21] A.P. Krempi, Recursos de Estatística Espacial a para Análise da Acessibilidade da Cidade de Bauru, Master's Thesis, Departamento de Transportes, Escola de Engenharia de São Carlos - USP, 2004.
[22] L. Anselin, Computing environments for spatial data analysis, Journal of Geographical Systems 2 (2001) 201–220.
[23] T. Shimada, Global Moran's I and Small distance adjustment: spatial pattern of crime in Tokyo, National Research Institute of Police Science, National Police Agency, Chiba, Japan, available at http://www.icpsr.umich.edu/CRIMESTAT/files/CrimeStatChapter.4.pdf, 2002.
[24] M.R.T. Dale, P. Dixon, M.-J. Fortin, P. Legendre, D.E. Myers, M.S. Rosenberg, Conceptual and mathematical relationships among methods for spatial analysis, Ecography 25 (2002) 558–577.
[25] C.J.C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Kluwer Academic Publishers, Dordrecht, 1998.
[26] Zhuang, L. and Dai, H. Parameter Optimization of Kernel-based One-class Classifier on Imbalance Learning. Journal of Computers 1 (7) (2006).
[27] B. Schölkopf, A. Smola, Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, Cambridge, MA, 2002.
[28] S. Haykin, Redes Neurais: Princípios e Prática, second ed., Bookman, Porto Alegre, 2001.
[29] B. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, Biochimica et Biophysica Acta 405 (2) (1975) 442.
[30] J.A. Swets, ROC analysis applied to the evaluation of medical imaging techniques, Investigative Radiology 4 (1979) 109–121.
[31] M. Heath, K. Bowyer, D. Kopans, Current Status of the Digital Database for Screening Mammography Digital Mammography, Kluwer Academic Publishers, Dordrecht, 1998, pp. 457–460.
[32] R. Gonzalez, R. Woods, Digital Image Processing, Addison-Wesley, Reading, MA, 1992.
[33] C.-C. Chang, C.-J. Lin, LIBSVM—a library for support vector machines, 2009, available at ⟨http://www.csie.ntu.edu.tw/~cjlin/libsvm/⟩.
[34] P. Gorgel, A. Sertbas, N. Kilic, O.N. Ucan, O. Osman, Mammographic mass classification using wavelet based support vector machine, Journal of Electrical and Electronics Engineering 9 (1) (2009) 867–875.

Geraldo Braz Junior received the master degree in electric engineering at the Federal University of Maranhão, Brazil, in 2008. Currently he is a professor of Informatics Department at the Federal University of Maranhão (UFMA), Brazil, where he teaches operational systems and programming languages. His researches include image processing, pattern recognition and machine learning.

Anselmo Cardoso de Paiva received BSc in civil engineering from Maranhão State University, Brazil, in 1990, an MSc in civil engineering structures and a PhD in informatics from Pontiphical Catholic University of Rio de Janeiro, Brazil, in 1993 and 2002. He is currently a professor at the Informatics Department, Federal University of Maranhão, Brazil. His current interests include medical image processing, geographical information systems and scientific visualization.

Aristófanes Corrêa Silva received a PhD degree in Informatics from Pontiphical Catholic University of Rio de Janeiro, Brazil, in 2004. Currently he is a professor at the Federal University of Maranhão (UFMA), Brazil. He teaches image processing, pattern recognition and programming language. His research interests include image processing, image understanding, medical image processing, machine vision, artificial intelligence, pattern recognition and machine learning.

Alexandre Cesar Muniz de Oliveira received his PhD in applied computing by the National Institute for Space Research (2004). He is currently an adjunct professor at the same Federal University of Maranhão, where he teaches computer science with emphasis on operation research, discrete and continuous optimization, search metaheuristics, parallel algorithms and pattern recognition.