Feature subset selection for classification of malignant and benign breast masses in digital mammography

Ramzi Chaieb, Karim Kalti

SURVEY

CrossMark

Feature subset selection for classification of malignant and benign breast masses in digital mammography

Ramzi Chaieb[1] · **Karim Kalti**[1]

Abstract

Computer-aided diagnosis of breast cancer is becoming increasingly a necessity given the exponential growth of performed mammograms. In particular, the breast mass diagnosis and classification arouse nowadays a great interest. Texture and shape are the most important criteria for the discrimination between benign and malignant masses. Various features have been proposed in the literature for the characterization of breast masses. The performance of each feature is related to its ability to discriminate masses from different classes. The feature space may include a large number of irrelevant ones which occupy a lot of storage space and decrease the classification accuracy. Therefore, a feature selection phase is usually needed to avoid these problems. The main objective of this paper is to select an optimal subset of features in order to improve masses classification performance. First, a study of various descriptors which are commonly used in the breast cancer field is conducted. Then, selection techniques are used in order to determine the most relevant features. A comparative study between selected features is performed in order to test their ability to discriminate between malignant and benign masses. The database used for experiments is composed of mammograms from the MiniMIAS database. Obtained results show that Gray-Level Run-Length Matrix features provide the best result.

Keywords Breast cancer · Computer-aided diagnosis (CAD) · Characterization · Selection · Classification · Evaluation

1 Introduction

Breast cancer is the leading cause of cancer deaths among the female population [1]. The only way today to reduce it is its early detection using imaging techniques [2, 3]. Mammography is one of the most effective tools for prevention and early detection of breast cancer [4–6]. It is a screening tool used to localize suspicious tissues in the breast such as microcalcifications and masses. It allows also the detection of architectural distortion and bilateral asymmetry [7]. A mass is defined as a space-occupying lesion seen in, at least, two different projections [8]. Mass density can be high, isodense, low or fat containing. Moreover, mass margin can be circumscribed, microlobulated, indistinct or spiculated. Mass shape can be round, oval, lobular or irregular [9]. In recent years, screening campaigns are being organized in several countries. These campaigns generate a huge stream of mammograms, and it is still difficult for expert radiologists to provide accurate and consistent analysis. Therefore, computer-aided diagnosis (CAD) systems are developed to help the radiologists in detecting lesions and in taking diagnosis decisions [9–14].

A generic CAD system by image analysis includes two main stages: feature extraction stage followed by a classification stage. In the literature, various numbers of techniques are studied to describe breast abnormalities in digital mammograms. A lot of research has been done on the textural and shape analysis on mammographic images [1, 9, 14–19]. In this paper, the main objective is to determine which features optimize the classification performance.

In the process of pattern recognition, the goal is to achieve the best classification rate using required features. The extraction of these features from regions of the image is one of the important phases in this process. Features are defined as the input variables which are used for classification. The quality of a feature is related to its ability to discriminate observations from different classes. The characterization task often generates a large number of features, and the obtained features space may include a large number of irrelevant ones.

✉ Ramzi Chaieb
  ramzi.chaieb@hotmail.com

1 LATIS- Laboratory of Advanced Technology and Intelligent Systems, ENISo, Sousse University, Sousse, Tunisia

This will induce greater computational cost, occupy a lot of storage space and decrease the classification performance. Thus, a feature selection phase is needed to avoid these problems. In this study, we propose an automated computer scheme in order to select an optimal subset of features for masses classification in digital mammography. The obtained results can be used in other applications such as segmentation and content-based image retrieval.

The remaining part of this paper is organized as follows. The next section gives an overview of the proposed methodology. Sections 3 to 5 describe the process of selecting features. Sections 6 and 7 present the combination of three classifiers and the different measures used to evaluate the classification performance. The experimental results are evaluated and discussed in Sect. 8. Finally, concluding remarks are given in the last section.

2 Methodology

Our approach in this study is composed of two main stages: characterization and classification. Each of these stages is explained in detail by the flowchart given in Fig. 1.

2.1 Feature extraction

Various features have been proposed in the literature for the characterization of masses. These features are organized into families according to their nature [16]. The majority of studies focus on one family and analyze its performance. In this work, we propose to study the performance of a set of feature families. Then, we make a comparison between these families in order to select the best feature set. Finally, the most discriminant features are selected from the obtained feature set. The process is described in Fig. 2. $\{FS_1, FS_2, ..., FS_i\}$ denotes the set of feature families. For each feature family, we selected its most relevant features using a number
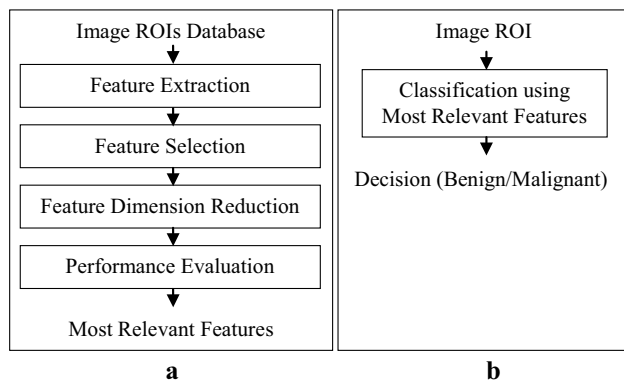
of feature selection methods $(FSM_1, FSM_2, ..., FSM_j)$. After this step, we obtained for each FS the set $\{FSM_1(FS_i)\}$, $\{FSM_2(FS_i)\}$, ..., $\{FSM_j(FS_i)\}$ where $FSM_j(FS_i)$ denotes FS_i selected features using FSM_j. Obtained results were used as input to the next step which is the selection of the best FSM that selects the most relevant features for each FS. Finally, we selected the optimal subset of features from the obtained results.

Texture and shape are the major criteria for the discrimination between the benign and malignant masses. In this study, we have followed two main kinds of description techniques. The first employs texture features extracted from ROIs. The second is based on computer-extracted shape features of masses, since morphology is one of the most important factors in breast cancer diagnosis.

2.1.1 Texture analysis

Texture analysis is performed in each ROI selected in the previous phase. The texture feature space can be divided into two subspaces: statistical and frequential features.

a) Statistical features: The statistical textural features that we have used in this study can be grouped into five sets based on what they are derived from: First-Order Statistics (FOS), Gray-Level Co-occurrence Matrices (GLCM), Gray-Level Difference Matrices (GLDM), Gray-Level Run-Length Matrices (GLRLM) and Tamura features.
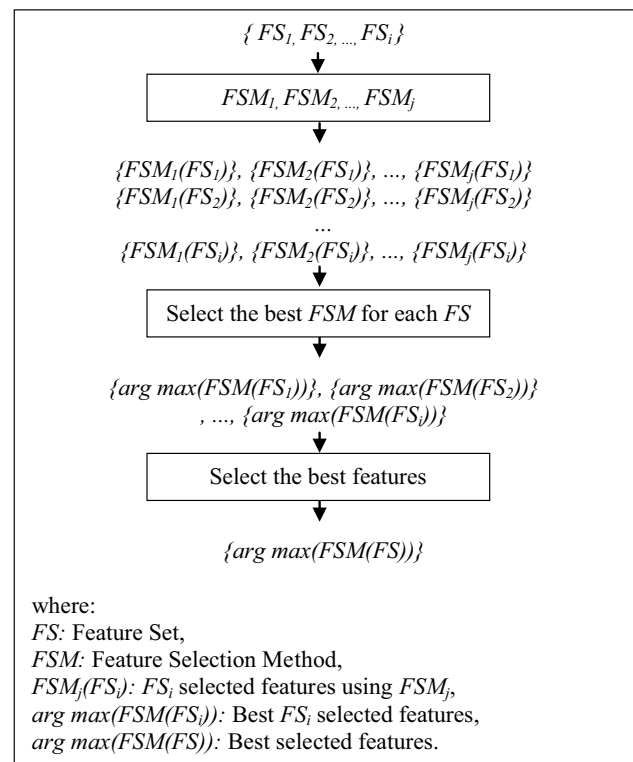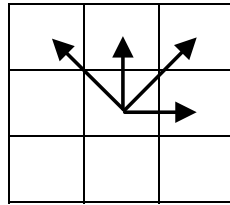

Fig. 1 Flowchart of the proposed methodology performed in this study. **a** Characterization stage and **b** classification stage

Fig. 2 Feature selection steps

Fig. 3 Four directions of adjacency for calculating the GLCM features

First-Order Statistics features: FOS provides different statistical properties of the intensity histogram of an image [20]. They depend only on individual pixel values and not on the interaction of neighboring pixels values. In this study, six first-order textural features were calculated: Mean value of gray levels, Mean square value of gray levels, Standard Deviation, Variance, Skewness and Kurtosis.

Denoting by I(x,y) the image subregion pixel matrix, the formulae used for the metrics of the FOS features are as follows:

- Mean value of gray levels:

$$M = \frac{1}{XY} \sum_{x=1}^{X} \sum_{y=1}^{Y} I(x,y) \tag{1}$$

- Mean square value of gray levels:

$$\text{MS} = \frac{1}{XY} \sum_{x=1}^{X} \sum_{y=1}^{Y} \left[I(x,y) \right]^2 \tag{2}$$

- Standard Deviation:

$$\text{SD} = \sqrt{\frac{1}{(XY-1)} \sum_{x=1}^{X} \sum_{y=1}^{Y} \left[I(x,y) - M \right]^2} \tag{3}$$

- Variance:

$$V = \frac{1}{(XY-1)} \sum_{x=1}^{X} \sum_{y=1}^{Y} \left[I(x,y) - M \right]^2 \tag{4}$$

- Skewness:

$$S = \frac{1}{XY} \sum_{x=1}^{X} \sum_{y=1}^{Y} \left[\frac{I(x,y) - M}{SD} \right]^3 \tag{5}$$

- Kurtosis:

$$K = \left\{ \frac{1}{XY} \sum_{x=1}^{X} \sum_{y=1}^{Y} \left[\frac{I(x,y) - M}{SD} \right]^4 \right\} - 3 \tag{6}$$

Gray-Level Co-occurrence Matrix features: The second feature group is a robust statistical tool for extracting second-order texture information from images [15, 21]. The

GLCM characterizes the spatial distribution of gray levels in the selected ROI. An element at location (i,j) of the GLCM represents the joint probability density of the occurrence of gray levels i and j in a specified orientation θ and specified distance d from each other (Fig. 3). Thus, for different θ and d values, different GLCMs are generated. Figure 4 shows how a GLCM with $\theta = 0°$ and $d = 1$ is generated. The number 4 in the co-occurrence matrix indicates that there are four occurrences of a pixel with gray level 3 immediately to the right of pixel with gray level 6.

Nineteen features were derived from each GLCM. Specifically, the features studied are: Mean, Variance, Entropy, Contrast, Angular Second Moment (also called Energy), Dissimilarity, Correlation, Inverse Difference Moment (also called Homogeneity), Diagonal Moment, Sum Average, Sum Entropy, Sum Variance, Difference Entropy, Difference Mean, Difference Variance, Information Measure of Correlation 1, Information Measure of Correlation 2, Cluster Prominence and Cluster Shade.

Denoting by: Ng the number of gray levels in the image, $p(i,j)$ the normalized co-occurrence matrix, $p_x(i)$ and $p_y(j)$ the row and column marginal probabilities, respectively, obtained by summing the columns or rows of $p(i,j)$:

$$p_x(i) = \sum_{j=1}^{N_g} p(i,j) \tag{7}$$

$$p_y(j) = \sum_{i=1}^{N_g} p(i,j) \tag{8}$$

$$p_{x+y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j); \quad k = i+j = 2, 3, \ldots, 2N_g \tag{9}$$

$$p_{x-y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j); \quad k = |i-j| = 0, 1, \ldots, N_g \tag{10}$$
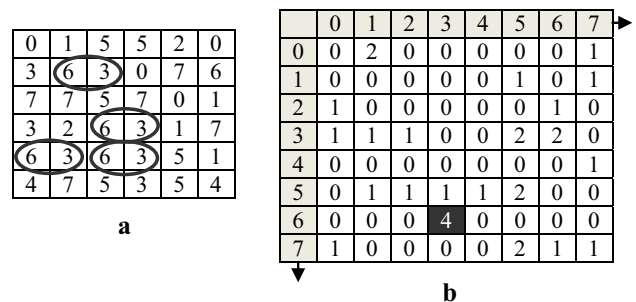

Fig. 4 Construction principle of co-occurrence matrix. **a** Initial image, **b** co-occurrence matrix ($\theta = 0°$ and $d = 1$)

The formulae used to calculate the GLCM features are as follows:

- Mean:

$$M = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} ip(i,j) \tag{11}$$

- Variance:

$$V = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \left[(i-M)^2 p(i,j) \right] \tag{12}$$

- Entropy:

$$\text{Ent} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \left[p(i,j) \log(p(i,j)) \right] \tag{13}$$

- Contrast:

$$\text{Cont} = \sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j); |i-j| = n \right\} \tag{14}$$

- Angular Second Moment (also called Energy):

$$\text{ASM} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \{p(i,j)\}^2 \tag{15}$$

- Dissimilarity:

$$\text{Diss} = \sum_{i=1}^{N_g-1} \sum_{j=1}^{N_g-1} (|i-j| p(i,j)) \tag{16}$$

- Correlation:

$$\text{Corr} = \frac{\left[\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (ij)p(i,j) \right] - \mu_x \mu_y}{\sigma_x \sigma_y} \tag{17}$$

where μ_x, μ_y, σ_x and σ_y are the Mean values and Standard Deviations of p_x and p_y, respectively.

$$\mu_x = \sum_{i=1}^{N_g} \left[i \sum_{j=1}^{N_g} p(i,j) \right] \tag{18}$$

$$\mu_y = \sum_{j=1}^{N_g} \left[j \sum_{i=1}^{N_g} p(i,j) \right] \tag{19}$$

$$\sigma_x = \sum_{i=1}^{N_g} \left[(i-\mu_x)^2 j \sum_{j=1}^{N_g} p(i,j) \right] \tag{20}$$

$$\sigma_y = \sum_{j=1}^{N_g} \left[(j-\mu_y)^2 \sum_{i=1}^{N_g} p(i,j) \right] \tag{21}$$

- Inverse Difference Moment (also called Homogeneity):

$$\text{IDM} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \left[\frac{1}{1+(i-j)^2} p(i,j) \right] \tag{22}$$

- Diagonal Moment:

$$\text{DM} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (\frac{1}{2}(|i-j|)p(i,j))^{1/2} \tag{23}$$

- Sum Average:

$$\text{SA} = \sum_{i=2}^{2N_g} \left[ip_{x+y}(i) \right] \tag{24}$$

- Sum Entropy:

$$\text{SE} = - \sum_{i=2}^{2N_g} \left[p_{x+y}(i) \log \left[p_{x+y}(i) \right] \right] \tag{25}$$

- Sum Variance:

$$\text{SV} = \sum_{i=2}^{2N_g} \left[(i-SA)^2 p_{x+y}(i) \right] \tag{26}$$

- Difference Entropy:

$$\text{DE} = - \sum_{i=1}^{N_g} \left[p_{x-y}(i) \log \left[p_{x-y}(i) \right] \right] \tag{27}$$

- Difference Mean:

$$\text{DMean} = \sum_{i=1}^{N_g} \left[ip_{x-y}(i) \right] \tag{28}$$

- Difference Variance:

$$\text{DV} = \sum_{i=1}^{N_g} \left[(i-\text{DMean})^2 p_{x-y}(i) \right] \tag{29}$$

- Information Measure of Correlation 1:

$$\text{IMC1} = \frac{HXY - HXY1}{max\{HX, HY\}} \tag{30}$$

where HX and HY are the entropy of p_x and p_y, respectively.

$$HX = -\sum_{i=1}^{N_g} p_x(i) \log \left(p_x(i)\right) \tag{31}$$

$$HY = -\sum_{j=1}^{N_g} p_y(j) \log \left(p_y(j)\right) \tag{32}$$

and

$$HXY = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \left[p(i,j)\log(p(i,j))\right] \tag{33}$$

$$HXY1 = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \left[p_x(i,j)\log\left(p_x(i)p_y(j)\right)\right] \tag{34}$$

- Information Measure of Correlation 2:

$$\text{IMC2} = \left[1 - \exp[-2(HXY2 - HXY)]\right]^{1/2} \tag{35}$$

where

$$HXY2 = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \left[p_x(i)p_y(j)\log\left(p_x(i)p_y(j)\right)\right] \tag{36}$$

- Cluster Prominence:

$$\text{CP} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i + j - 2M)^4 p(i,j) \tag{37}$$

- Cluster Shade:

$$\text{CS} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i + j - 2M)^3 p(i,j) \tag{38}$$

Gray-Level Difference Matrix features: The GLDM features are extracted from the gray-level difference matrices vector of an image [22]. The GLDM vector is the histogram of the absolute difference of pixel pairs separated by a given displacement vector $\delta = (\Delta x, \Delta y)$, where $I_\delta(x,y) = |I(x + \Delta x) - I(y + \Delta y)|$ and Δx and Δy are integers. An element of GLDM vector $p_\delta(i)$ can be computed by counting the number of times that each value of $I_\delta(x,y)$ occurs. In practice, the displacement vector $\delta = (\Delta x, \Delta y)$ is usually selected to have a phase of value as 0°, 45°, 90° or 135° to obtain the oriented texture features. GLDM method is based on the generalization of a density function of gray-level difference. In this study, five GLDM features were calculated: Mean, Contrast, Angular Second Moment, Entropy and Inverse Difference Moment.

Denoting by $f(i/\delta)$ the probability density associated with possible values of I_δ, $f(i/\delta) = P(I_\delta(x,y) = i)$ and M the number of gray-level differences, the formulae used for the metrics of the GLDM features are as follows:

- Mean:

$$\text{Mean} = \sum_{i=1}^{M} i f(i|\delta) \tag{39}$$

- Contrast:

$$\text{Cont} = \sum_{i=1}^{M} i^2 f(i|\delta) \tag{40}$$

- Angular Second Moment:

$$\text{ASM} = \sum_{i=1}^{M} \left[f(i|\delta)\right]^2 \tag{41}$$

- Entropy:

$$\text{Ent} = \sum_{i=1}^{M} -f(i|\delta) \ln \left(f(i|\delta)\right) \tag{42}$$

- Inverse Difference Moment:

$$\text{IDM} = \sum_{i=1}^{M} \frac{f(i|\delta)}{i^2 + 1} \tag{43}$$

Gray-Level Run-Length Matrix features: GLRLM provides information related to the spatial distribution of gray-level runs (i.e., pixel structures of same pixel value) within the image [22]. Each gray-level run can be characterized by its gray level, length and direction. Textural features extracted from GLRLM evaluate the distribution of small (short runs) or large (long runs) organized structures within ROIs. For each of the four directions θ (0°, 45°, 90° and 135°), we can generate a GLRLM. Figure 5 shows an example of constructing a GLRLM with $\theta = 135$°.

Denoting by Ng the number of gray levels, N_r the maximum run length and $p(i,j)$ the (i,j)th element of the run-length matrix for a specific angle θ and a specific distance d (i.e., $p_{\theta,d}(i,j)$), each element of the run-length matrix represents the number of pixels of run length j and gray level
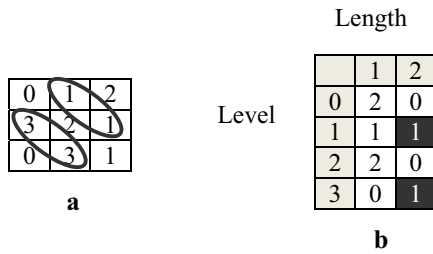
Fig. 5 Construction principle of GLRLM. **a** Initial image, **b** GLRLM ($\theta = 135°$)

i. Gray-Level Run-Number Vector and Run-Length Run-Number Vector are defined as follows:

- Gray-Level Run-Number Vector:

$$p_g(i) = \sum_{j=1}^{N_r} p(i,j) \tag{44}$$

This vector represents the sum distribution of the number of runs with gray level i.

- Run-Length Run-Number Vector:

$$p_r(j) = \sum_{i=1}^{N_g} p(i,j) \tag{45}$$

This vector represents the sum distribution of the number of runs with run length j.

- The total number of runs:

$$N_{\text{runs}} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j) \tag{46}$$

From each ROI, eleven GLRLM features are generated: Short Runs Emphasis (*SRE*), Long Runs Emphasis (*LRE*), Gray-Level Non-uniformity (*GLN*), Run-Length Non-uniformity (*RLN*), Run Percentage (*RP*), Low Gray-Level Run Emphasis (*LGRE*), High Gray-Level Run Emphasis (*HGRE*), Short Run Low Gray-Level Emphasis (*SRLGE*), Short Run High Gray-Level Emphasis (*SRHGE*), Long Run Low Gray-Level Emphasis (*LRLGE*) and Long Run High Gray-Level Emphasis (*LRHGE*).

The formulae used to calculate GLRLM features are as follows:

- Short Runs Emphasis:

$$\text{SRE} = \frac{1}{N_{runs}} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{p(i,j)}{j^2} = \frac{1}{N_{runs}} \sum_{j=1}^{N_r} \frac{p_r(j)}{j^2} \tag{47}$$

- Long Runs Emphasis:

$$\text{LRE} = \frac{1}{N_{\text{runs}}} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j).j^2 = \frac{1}{N_{\text{runs}}} \sum_{j=1}^{N_r} p_r(j).j^2 \tag{48}$$

- Gray-Level Non-uniformity:

$$\text{GLN} = \frac{1}{N_{\text{runs}}} \sum_{i=1}^{N_g} \left[\sum_{j=1}^{N_r} p(i,j) \right]^2 = \frac{1}{N_{\text{runs}}} \sum_{i=1}^{N_g} \left[p_g(i) \right]^2 \tag{49}$$

- Run-Length Non-uniformity:

$$\text{RLN} = \frac{1}{N_{\text{runs}}} \sum_{j=1}^{N_r} \left[\sum_{i=1}^{N_g} p(i,j) \right]^2 = \frac{1}{N_{\text{runs}}} \sum_{j=1}^{N_r} \left[p_r(j) \right]^2 \tag{50}$$

- Run Percentage:

$$\text{RP} = \frac{N_{\text{runs}}}{N_{\text{pixels}}} \tag{51}$$

where N_{pixels} is the total number of pixels in the image.

- Low Gray-Level Run Emphasis:

$$\text{LGRE} = \frac{1}{N_{\text{runs}}} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{p(i,j)}{i^2} = \frac{1}{N_{\text{runs}}} \sum_{i=1}^{N_g} \frac{p_g(i)}{i^2} \tag{52}$$

- High Gray-Level Run Emphasis:

$$\text{HGRE} = \frac{1}{N_{\text{runs}}} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j).i^2 = \frac{1}{N_{\text{runs}}} \sum_{i=1}^{N_g} p_r(j).i^2 \tag{53}$$

- Short Run Low Gray-Level Emphasis:

$$\text{SRLGE} = \frac{1}{N_{\text{runs}}} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{p(i,j)}{i^2.j^2} \tag{54}$$

- Short Run High Gray-Level Emphasis:

$$\text{SRHGE} = \frac{1}{N_{\text{runs}}} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{p(i,j).i^2}{j^2} \tag{55}$$

- Long Run Low Gray-Level Emphasis:

$$\text{LRLGE} = \frac{1}{N_{\text{runs}}} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{p(i,j).j^2}{i^2} \tag{56}$$

- Long Run High Gray-Level Emphasis:

$$\text{LRHGE} = \frac{1}{N_{\text{runs}}} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p(i,j).i^2.j^2 \tag{57}$$

Tamura features: Tamura et al. [23, 24] defined six texture features (coarseness, contrast, direction, linearity, regularity and roughness). The first three descriptors are based on concepts that correspond to human visual perception. They are effective and frequently used to characterize textures.

b) Frequential features: The second texture feature subspace is based on transformations. Texture is represented in a frequency domain other than the spatial domain of the image. Two different structural methods are considered: Gabor transform and two-dimensional wavelet transform.

Gabor filters: Gabor filter is widely adopted to extract texture features from the images [25, 26] and has been shown to be very efficient. Basically, Gabor filters are a group of wavelets, with each wavelet capturing energy at a specific frequency and a specific direction. After applying Gabor filters on the ROI with different orientations at different scales, we obtain an array of magnitudes:

$$E(m, n) = \sum_x \sum_y |G_{mn}(x, y)| \tag{58}$$

where m, n and G represent, respectively, the scale, orientation and filtered image.

These magnitudes represent the energy content at different scales and orientations of the image. Texture features are found by calculating the mean and variation of the Gabor filtered image. The following mean μ_{mn} and Standard Deviation σ_{mn} of the magnitude of the transformed coefficients are used to characterize the texture of the region.

$$\mu_{mn} = \frac{E(m, n)}{P \times Q} \tag{59}$$

$$\sigma_{mn} = \frac{\sqrt{\sum_x \sum_y \left(|G_{mn}(x, y) - \mu_{mn}|\right)^2}}{P \times Q} \tag{60}$$

Five scales and six orientations are used in common implementation, and the features vector f is created using μ_{mn} and σ_{mn} as the feature components.

$$f = \left(\mu_{00}, \sigma_{00}, \mu_{01}, \sigma_{01}, \ldots, \mu_{30}, \sigma_{30}\right) \tag{61}$$

Wavelets: Wavelets are very efficient for the characterization of texture in images (especially mammographic images)

which may have different types of texture and require sufficient detail characterization. For this reason, it was chosen to be used in this study. The discrete wavelet transform (DWT) of an image is a transform based on the tree structure (successive low-pass and high-pass filters) as shown in Fig. 6. DWT decomposes a signal into scales with different frequency resolutions. The resulting decomposition is shown in Table 1. The image is divided into four bands: LL (left top), LH (right top), HL (left bottom) and HH (right bottom). High frequencies provide global information, while low frequencies provide detailed information hidden in the original image.

Energy measures at different resolution levels are used to characterize the texture in the frequency domain.

To extract wavelet features, the type of wavelet mother function chosen is the Haar wavelet, and its mother function $\psi(t)$ can be described as follows (Fig. 7):

$$\psi(t) = \begin{cases} 1 & 0 \le t \le \frac{1}{2} \\ -1 & \frac{1}{2} \le t \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{62}$$

2.1.2 Shape analysis

The shape features are also called the morphological or geometric features. These kinds of features are based on the shapes of ROIs. Seven Hu's invariant moments were adopted for shape features extraction due to their reportedly excellent performance on shape analysis [16, 27]. These moments are computed based on the information provided by both the shape boundary and its interior region. Its values are invariant with respect to translation, scale and rotation of the shape. Given a function $f(x,y)$, the (p,q)th moment is defined by:

Table 1 Wavelet transform representation of an image (two levels)

LL^2	LH^2	LH^1
HL^2	HH^2	
HL^1		HH^1

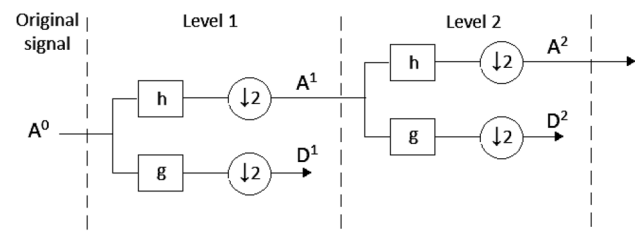1,2: decomposition level; H, high-frequency band; L, low-frequency band

Fig. 6 Two-level wavelet decomposition tree (h: low-pass decomposition filter, g: high-pass decomposition filter, ↓2 down-sampling operation, A1, A2: approximated coefficient, D1, D2: detailed coefficient)
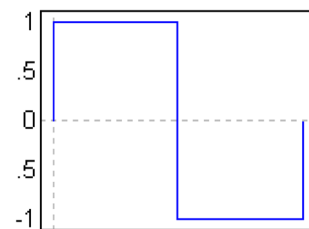

Fig. 7 Haar wavelet

$$M_{pq} = \int\limits_{-\infty}^{+\infty} \int\limits_{-\infty}^{+\infty} x^p y^q f(x,y)dxdy; \quad p,q = 0,1,2,\ldots \tag{63}$$

For implementation in digital form, Eq. 63 becomes:

$$M_{pq} = \sum_X \sum_Y x^p y^q f(x,y) \tag{64}$$

The central moments can then be defined in their discrete representation using the following formula:

$$\mu_{pq} = \sum_X \sum_Y (x - \bar{x})^p (y - \bar{y})^q \tag{65}$$

where

$$\bar{x} = \frac{M_{10}}{M_{00}} \tag{66}$$

$$\bar{y} = \frac{M_{01}}{M_{00}} \tag{67}$$

$\bar{x}$ and $\bar{y}$ are the coordinates of the image gravity center. The moments are further normalized for the effects of change of scale using the following formula:

$$\eta_{p,q} = \frac{\mu_{p,q}}{\mu_{0,0}^{\left(1 + \frac{p+q}{2}\right)}} \tag{68}$$

From the normalized central moments, a set of seven values Φ_i, $1 \leq i \leq 7$, set out by Hu, may be computed by the following formulas:

$$\Phi_1 = \eta_{20} + \eta_{02} \tag{69}$$

$$\Phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \tag{70}$$

$$\Phi_3 = (\eta_{30} - 3\eta_{12})^2 + (\eta_{03} - 3\eta_{21})^2 \tag{71}$$

$$\Phi_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{03} + \eta_{21})^2 \tag{72}$$

$$\Phi_5 = (3\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right]$$
$$+ (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \times \left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \tag{73}$$

$$\Phi_6 = (\eta_{20} - \eta_{02})\left[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right]$$
$$+ 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \tag{74}$$

$$\Phi_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})\left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right]$$
$$+ (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03}) \times \left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \tag{75}$$

Furthermore, fourteen significant descriptors are also developed in this section [28]:

- Area, defined as the number of pixels in the region.
- Perimeter, calculated as the distance around the boundary of the region.
- Compactness, calculated as:

$$C = \frac{P^2}{4 \times \pi \times A} \tag{76}$$

where P and A represent the Perimeter and Area, respectively.

- Major Axis Length, defined as the length (in pixels) of the major axis of the ellipse that has the same normalized second central moments as the region.
- Minor Axis Length, defined as the length (in pixels) of the minor axis of the ellipse that has the same normalized second central moments as the region.
- Eccentricity, defined as the scalar that specifies the eccentricity of the ellipse that has the same second moments as the region. The eccentricity is the ratio of the distance between the foci of the ellipse and its Major Axis Length. The value is between 0 and 1. 0 and 1 are degenerate cases: An ellipse whose eccentricity is 0 is actually a circle, while an ellipse whose eccentricity is 1 is a line segment.
- Orientation is defined as the angle (in degrees ranging from -90 to 90 degrees) between the x-axis and the major axis of the ellipse that has the same second moments as the region.
- Equivalent diameter specifies the diameter of a circle with the same area as the region. It is computed as:

$$\text{Eq}_{\text{diam}} = \sqrt{\frac{4 \times A}{\pi}} \tag{77}$$

- Convex polygon area specifies the proportion of the pixels in the convex hull (the smallest convex polygon that can contain the region).
- Solidity specifies the proportion of the pixels in the convex hull that are also in the region. It is computed as the Area/ConvexArea where ConvexArea is a scalar that specifies the number of pixels in the convex hull.
- Euler Number specifies the number of objects in the region minus the number of holes in those objects.
- Extent specifies the ratio of pixels in the region to pixels in the total bounding box (the smallest rectangle containing the region). It is computed as the Area divided by the area of the bounding box.
- Mean Intensity, calculated as the mean of all the intensity values in the region.
- Aspect ratio, the aspect ratio of a region describes the proportional relationship between its width and its height. It is computed as:

$$\text{Aspect}_{\text{ratio}} = \frac{x_{\max} - x_{\min} + 1}{y_{\max} - y_{\min} + 1} \tag{78}$$

where $(x_{\min}, y_{\min})$ and $(x_{\max}, y_{\max})$ are the coordinates of the upper left corner and lower right corner of the bounding box.

3 Feature selection

At the stage of feature analysis, many features are generated for each ROI. The feature space is very large and complex due to the wide diversity of the normal tissues and the variety of the abnormalities. Only some of them are significant. With a large number of features, the computational cost will increase. Irrelevant and redundant features may affect the training process and consequently minimize the classification accuracy. The main goal of feature selection is to reduce the dimensionality by eliminating irrelevant features and selecting the best discriminative ones. Many search methods are proposed for feature selection. These methods could be categorized into sequential or randomized feature selection. Sequential methods are simple and fast but they could not backtrack, which means that they are candidate to fall into local minima. The problem of local minima is solved in randomized feature selection methods, but with randomized methods it is difficult to choose proper parameters.

To avoid problems of local minima and choosing proper parameters, we have opted to use five feature selection methods. Then, using an evaluation criterion, we retain the method that gives the best results among them. The feature selection methods that we have used are: tabu search (TS), genetic algorithm (GA), ReliefF algorithm (RA), sequential forward selection (SFS) and sequential backward selection (SBS). We used all extracted features as input for selection methods individually.

3.1 Tabu search

The tabu search is a meta-heuristic approach that can be used to solve combinatorial optimization problem. TS is conceptually simple and elegant. It has recently received widespread attention [25]. It is a form of local neighborhood search. It differs from the local search techniques in the sense that tabu search allows moving to a new solution which makes the objective function worse in the hope that it will not trap in local optimal solutions. Each solution $S \in \Omega$ has an associated neighborhood $N(S) \subseteq \Omega$ where Ω is the set of feasible solutions. Each solution $S' \in N(S)$ is reached from S by an operation called a *move* to S'. Tabu search uses a short-term memory, called tabu list, to record and guide the process of the search. To avoid cycling, solutions that were recently explored are declared forbidden or tabu for a number of

iterations. In this study, the size of the tabu list is set to be equal to 2. This size is reasonable to ensure diversity. Some additional precautions can be taken to avoid missing good solutions. These strategies are known as aspiration criteria. Aspiration level is a commonly used aspiration criterion. It is employed to override a solution's tabu state. It allows a tabu move when it results in a solution with an objective value better than that of the current best-known solution. In addition to the tabu list, we can also use a long-term memories and other prior information about the solutions to improve the intensification and/or diversification of the search.

Algorithm 1: Tabu Search Algorithm [25]

Ω : Set of feasible solutions

S : Current Solution

S^* : Best admissible solution

Cost : Objective function

$N(S)$: Neighborhood of solution S

V^* : Sample of neighborhood solutions

T : Tabu list

AL : Aspiration Level

<u>Input:</u>

initial feasible solution $S \in \Omega$

<u>Output:</u>

Most relevant features

<u>Procedure:</u>

<u>Begin</u>

Start with an initial feasible solution $S \in \Omega$

Initialize tabu list and aspiration level

<u>For</u> fixed number of iterations <u>Do</u>

Generate neighbor solutions $V^* \subset N(S)$

Find best $S^* \in V^*$

<u>If</u> move S to S^* is not in T <u>Then</u>

    Accept move and update best solution

    Update tabu list and aspiration level

    Increment iteration number

<u>Else</u>

    <u>If</u> $Cost(S^*) < AL$ <u>Then</u>

        Accept move and update best solution

        Update tabu list and aspiration level

        Increment iteration number

    <u>End If</u>

<u>End If</u>

<u>End For</u>

<u>End</u>

3.2 Genetic algorithm

Genetic algorithm is based on randomness in its search procedure to escape falling in local minima. First, a population of solutions based on the chromosomes is created. Then, the solutions are evolved by applying genetic operators such as mutation and crossover to find best solution based on the predefined fitness function (Fig. 8).

Algorithm 2: Basic Genetic Algorithm [26]

Input:

Crossover probability P_{co}

Mutation probability P_{mut}

Population size: L-chromosomes- or classifier- by N-bits

The used criteria function *Fit( )*

Fitness threshold θ

Output:

The set of highest fitness chromosomes (best classifier)

Procedure:

Begin

Do

Determine fitness of each chromosome, *Fit(i), i=1,...,L*

Rank the chromosomes

Do

Select two chromosomes with highest score

If *(Rand[0,1] < P_{co})* Then

    Crossover the pair at a randomly chosen bit

Else

    Change each bit with probability P_{mut}

    Remove the parent chromosomes

Until *N* offspring have been created

Until Any chromosome's score *Fit( ) exceeds θ*

Return Highest fitness chromosome (best classifier)

End

The initial population of GA is created using the following formula:

$$P = round((L - 1) \times rand(DF, 200 \times DF)) + 1 \qquad (79)$$

where L and DF represent, respectively, the number of input features and the desired number of selected features. (In this work, $DF = L/2$.)

In GA, we use a fitness function based on the principle of max-relevance and min-redundancy (mRMR) [27]. The idea of mRMR is to select the set S with m features $\{x_i\}$ that satisfies the maximization problem:

$$\max \Phi_i(D, R); \ \Phi(D, R) = D - R \qquad (80)$$
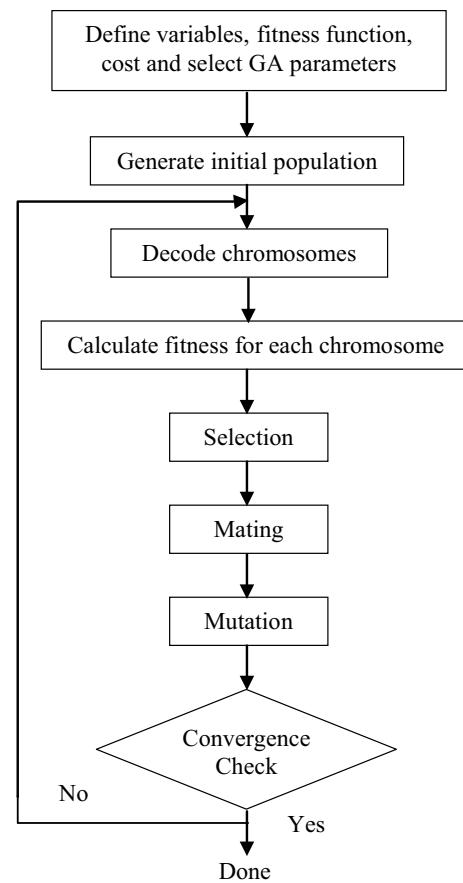
Fig. 8 Flowchart of the GA used in this study

where D and R represent the max-relevance and min-redundancy, respectively, and are computed by the following formula:

$$D = \frac{1}{|S|} \sum_{x_i} I(x_i, y) \qquad (81)$$

$$R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j) \qquad (82)$$

where $I(x_i, y)$ and $I(x_i, x_j)$ represent the mutual information, which is the quantity that measures the mutual dependence of the two random variables and is computed using the following formula:

$$I(x, y) = H(x) + H(y) - H(x, y) \qquad (83)$$

where $H(.)$ is the entropy.

3.3 ReliefF algorithm

The key idea of the ReliefF algorithm is to estimate the quality of features according to how well their values distinguish

between instances that are near to each other [29, 32]. ReliefF assigns a grade of relevance to each feature, and those features valued over a user given threshold are selected.

input list. SBS instead begins with all features and repeatedly removes a feature whose removal yields the maximal performance improvement.

Algorithm 3: ReliefF Algorithm [31]

Input:

For each training instance a vector of attribute values and the class value

Output:

The vector W of estimations of the qualities of attributes

Procedure:

Begin

set all weights $W[A]:=0.0$

For $i := 1$ to m Do Begin

      randomly select an instance R_i

      find k nearest hits H_j

For each class C≠class(R_i) Do

      from class C find k nearest misses $M_j(C)$

For $A:=1$ to a Do

$$W[A] := W[A] - \sum_{j=1}^{k} \frac{\text{diff}(A, R_i, H_j)}{m\,k} + \sum_{C \neq class(R_i)} \frac{\left[\frac{P(C)}{1-P(class(R_i))} \sum_{j=1}^{k} \text{diff}(A, R_i, M_j(C)) \right]}{m\,k}$$

End

3.4 Sequential forward selection and sequential backward selection

SFS and SBS are classified as sequential search. The main idea of these algorithms is to add or remove features to the vector of selected features, and the iteration continues until all features are checked. Adding or removing a feature is based on one of many criteria functions that could be used. Misclassification, distance, correlation and entropy are examples of criteria functions. The main disadvantage of these algorithms is that they have a tendency to become trapped in local minima. In SFS, we start with empty list of selected feature, and successively, we add one useful feature to the list until no useful feature remains in the extracted

---

Algorithm 4: SFS Algorithm

Input:

$F=\{f_1, f_2, ..., f_N\}$

M : size of the final set

Output:

$E=\{f_{s1}, f_{s2}, ..., f_{sM}\}$

Procedure:

Begin

$E = \emptyset$

For $i = 1$ to M Do

For $j = 1$ to $|F|$ Do

      Evaluate $f_j \cup E$

End For

f_{max}=best f_j

$E=E \cup f_{max}$, $F=F \setminus f_{max}$

End For

Return E

End

Algorithm 5: SBS Algorithm

Input:

$F=\{f_1, f_2, ..., f_N\}$

M : size of the final set

Output:

$E=\{f_{s1}, f_{s2}, ..., f_{sM}\}$

Procedure:

Begin

$E = F$

For $i = 1$ to N-M Do

For $j = 1$ to $|E|$ Do

    Evaluate $E \setminus f_j$

End For

f_{min}= worst f_j

E=$E \setminus f_{min}$

End For

Return E

End

---

4 Feature dimension reduction

After selecting the most relevant features, the next step is to reduce the dimensionality of the feature set in order to minimize the computational complexity. The dimension reduction is carried out using the principal component analysis (PCA) method [30]. PCA is a powerful tool for analyzing the dependencies between the variables and compressing the data by reducing the number of dimensions, without losing too much information.

5 Feature performance

This stage has two main purposes: choosing the input parameters (distance, direction, scale, etc.), giving the best results and comparing types of features. Features are evaluated according to their discriminatory power. Five measures derived from Rodrigues approach [24] can be used as criteria for this purpose and described as follows:

5.1 The class classifier (CC) measure

The first measure is based on the detachment property. For a given class, the CC states whether or not a given class is detached from the other classes. To determine this measure, it is necessary to identify the regions occupied by elements of different classes. Every element of the class corresponds to a feature vector, and it is represented by a point in a multidimensional space. Therefore, the elements of a given class form a cloud of points included in a minimum bounding sphere (MBS). To determine the MBS of a class, we must identify its most central element (*mce*) and its radius. *mce* is the element closest to the center of the class's MBS. It is the element s_i that minimizes Eq. 84.

$$\sum_{s_j \in s_c} d(s_i, s_j) \tag{84}$$

where S_C is the set of images of the class C in the test dataset S and d is a distance function for every element $s_i \in S_C$.

Then, the radius of the class C can be found as follows:

$$\max_{s_j \in s_c}(d(mce_c, s_j)) \tag{85}$$

Once the mce_c and the radius of the given class C are identified, it is required to evaluate how much the other classes invade the MBS of each class C. This can be performed by measuring the distance from the mce_c to every element of the other classes. We consider that the elements whose distance from mce_c is smaller or equal to the radius of the class C invade the MBS of class C. For a given class, the CC measure assumes 1 (fully detached) if the class is not invaded by other classes' elements, or 0 otherwise.

5.2 The class variance (CV) measure

The second measure is based on the condensation property. It states how condensed is each class. It corresponds to the distance variance of the elements to the center of the class. Knowing the mce_c and the radius of class C, we can calculate the CV measure. The average distance of the elements of the class C to the mce_c is calculated as:

$$\bar{d} = \frac{\sum_{s_j \in s_c} d(s_i, s_j)}{n-1}; \quad s_j \neq n \tag{86}$$

where n is the cardinality of class C and the variance of the class is given by:

$$S^2 = \frac{\sum_{s_j \in s_c}(d(mce_c, s_j) - \bar{d}}{n-1}; \quad s_j \in mce_c \tag{87}$$

From these two measures, we proposed three other measures.

5.3 The total class classifier (TCC) measure

It corresponds to the sum of CCs.

$$TCC = \sum_{i=1}^{t} CC_{C_i} \qquad (88)$$

where t is the number of classes in ROIs database.

5.4 The weighted average of class classifiers (WACC) measure

It corresponds to the weighted average of CCs.

$$WACC = \sum_{i=1}^{t} CC_{C_i} PC_i \qquad (89)$$

where PC_i is the number of elements in each class.

5.5 The weighted average of class variances (WACV) measure

It corresponds to the weighted average of CVs.

$$WACV = \sum_{i=1}^{t} CV_{C_i} PC_i \qquad (90)$$

6 Classification

The main goal of this stage is to test the ability of most relevant features to discriminate between benign and malignant masses. Once the features relating to masses have been extracted and selected, they can then be used as inputs to the classifier to classify the regions of interest into benign and malignant. In the literature, various classifiers are described to distinguish between normal and abnormal masses. Each classifier has its advantages and disadvantages. In this work, we used three classifiers: multilayer perceptron (MLP), support vector machines (SVMs) and K-nearest neighbors (K-NNs) which have performed well in mass classification [16]. Then, we make a combination of these classifiers in order to exploit the advantages of each one of them and to improve the accuracy and efficiency of the classification system.

6.1 Multilayer perceptron (MLP)

MLP is regarded as one prominent tool for training a system as a classifier [16]. It uses simple connection of the artificial neurons to imitate the neurons in human. It also has a huge capacity of parallel computing and powerful remembrance ability. MLP is organized in successive layers: an input layer, one or more hidden layers and output layer.

6.2 Support vector machines (SVMs)

SVM is an example of supervised learning methods used for classification [28]. It is a binary classifier that for each given input data it predicts which of two possible classes comprise the input. It is based on the idea to look for the hyperplane that maximizes the margin between two classes.

6.3 K-nearest neighbors (K-NNs)

The K-NN classifier is well explored in the literature and has been proved to have good classification performance on a wide range of datasets [26]. The K-nearest neighbors of an unknown sample are selected from the training set in order to predict the class label as the most frequent one occurring in the K-neighbors. K is a parameter which can be adjusted, and it is usually an integer. It specifies the number of nearest neighbors. In this work, K is equal to 3 and Euclidean distance is used as a distance metric.

In order to accumulate the classifiers advantages and to exploit the complementarity between the three classifiers, we applied six fusion methods: Majority Vote (MV) which is based on voting algorithm and five other methods: Maximum (Max), Minimum (Min), Median (Med), Product (Prod) and Sum (S) which are based on posterior probability. Then, the best accuracy obtained after this combination is used as final result of classification.

MV technique is based on the principle of voting. The final decision is made by selecting the class with the greatest number of votes. MV is based on a majority rule expressed as follows:

$$E(x) = \begin{cases} C_i \ if \ \sum_i e(i) = \max_{c_i \in \{1,\dots,M\}} \sum_j e(j) \geq \alpha K \\ else \ reject \end{cases} \qquad (91)$$

where K is the number of classifiers (in our case $K = 3$); M is the number of classes (in our case $M = 2$); $e_j(x) = C_i$ ($i \in \{1, \dots, M\}$) indicates that the classifier j assigned class C_i to x.

To calculate the five other methods (Max, Min, Med, Prod and Sum), we make the following assumptions:

- The decision of the kth classifier is defined as $d_{k,j} \in \{0,1\}$, $k = 1, \dots, K; j = 1, \dots, M$. If kth classifier chooses class ω_j, then $d_{k,j} = 1$, and 0, otherwise.
- X is the feature vector (derived from the input pattern) presented to the kth classifier.
- The outputs of the individual classifiers are $P_k(\omega_j | X)$, i.e., the posterior probability belonging to class j given the feature vector X.

The final ensemble decision is the class j that receives the largest support $\mu_j(x)$. Specifically

$$h_{\text{final}}(x) = \arg\max_j \mu_j(x) \tag{92}$$

where $\mu_j(x)$ are computed as follows:

- Maximum rule:

$$\mu_j(x) = \max_{k=1,\ldots,K} \left\{ d_{k,j}(x) \right\} \tag{93}$$

- Minimum rule:

$$\mu_j(x) = \min_{k=1,\ldots,K} \left\{ d_{k,j}(x) \right\} \tag{94}$$

- Median rule:

$$\mu_j(x) = \underset{k=1,\ldots,K}{med} \left\{ d_{k,j}(x) \right\} \tag{95}$$

- Sum rule:

$$\mu_j(x) = \sum_{k=1}^{K} d_{k,j}(x) \tag{96}$$

- Product rule:

$$\mu_j(x) = \prod_{k=1}^{K} d_{k,j}(x) \tag{97}$$

7 Classification performance evaluation

A number of different measures are commonly used to evaluate the classification performance. These measures are derived from the confusion matrix which describes actual and predicted classes as shown in Table 2.

- Rate of Positive Predictions:

$$RPP = \frac{TP + FP}{TP + FN + FP + TN} \tag{98}$$

- Rate of Negative Predictions:

$$RNP = \frac{TN + FN}{TP + FN + FP + TN} \tag{99}$$

Table 2 Confusion matrix

Actual	Predicted	
	Positive	Negative
Positive	TP (true positive)	FP (false positive)
Negative	FN (false negative)	TN (true negative)

TP predicts abnormal as abnormal; FP predicts abnormal as normal; TN predicts normal as normal, and FN predicts normal as abnormal

- True Positive Rate (Sensitivity):

$$TPR = \frac{TP}{TP + FN} \tag{100}$$

- False Negative Rate:

$$FNR = \frac{FN}{TP + FN} \tag{101}$$

- False Positive Rate:

$$FPR = \frac{FP}{TN + FP} \tag{102}$$

- True Negative Rate (Specificity):

$$TNR = \frac{TN}{TN + FP} \tag{103}$$

- Positive Predictive Value:

$$PPV = \frac{TP}{TP + FP} \tag{104}$$

- Negative Predictive Value:

$$NPV = \frac{TN}{TN + FN} \tag{105}$$

- Accuracy:

$$AC = \frac{TP + TN}{TP + FP + TN + FN} \tag{106}$$

- Mathews Correlation Coefficient:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{107}$$

- F-measure:

$$F = \frac{2 \times PPV \times TPR}{PPV + TPR} \tag{108}$$

- G-measure:

$$G = \sqrt{TNR \times TPR} \tag{109}$$

A receiver operating characteristic (ROC) curve is also used for this stage. It is a plotting of true positive as a function of false positive. Higher ROC, approaching the perfection at the upper left hand corner, would indicate greater discrimination capacity (Fig. 9).

8 Experimental results

All experiments were implemented in MATLAB. The source of the mammograms used in this experiment is the MiniM-IAS database [31] which is publicly available for research.
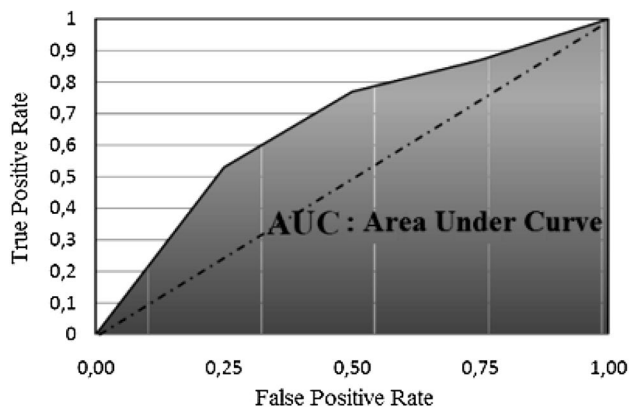
Fig. 9 ROC curve
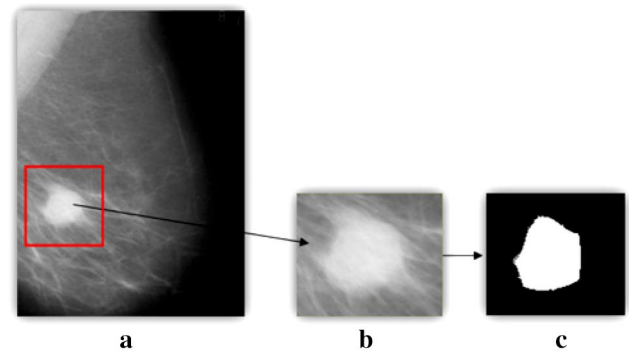
Fig. 12 Resulting binary image. **a** Selection of ROI specified by an expert radiologist, **b** ROI containing only one mass, **c** mass shape in binary format
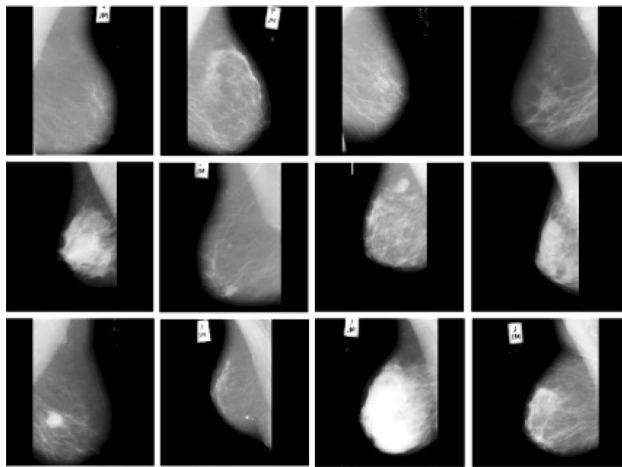
Fig. 10 Some samples from MiniMIAS database (normal (first line), benign (second line) and cancer (third line))

Fig. 13 By multiplying the enhanced image (first raw) with the mask (second raw), we obtain the segmented mass shown in the third raw

Fig. 11 Two examples of masses. **a** Benign and **b** malignant

It contains left and right breast images for 161 patients. It is composed of 322 mammograms which includes 208 normal breasts, 63 ROIs with benign lesions and 51 ROIs with malignant (cancerous) lesions, and 80% set of abnormal images are used for training and 20% used for testing. Every X-ray film is 1024 × 1024 pixels in size, and each pixel is represented with an 8-bit word. The database includes a readme file, which details expert radiologist's markings for each mammogram (MiniMIAS database reference number, character of background tissue, class of abnormality present, severity of abnormality, x,y image coordinates of center of abnormality and approximate radius (in pixels) of a circle enclosing the abnormality).

Figure 10 shows some cases representing three types of mammograms: normal (first line), benign (second line) and cancer (third line), downloaded from MiniMIAS database.

The first step in our experiments is to perform ROI selection. This involves separating the suspected areas they may contain abnormalities from the image. The ROIs are usually very small and limited to areas being determined as suspicious regions of masses. The suspicious area is an area that is brighter than its surroundings, has a regular shape with varying size and has fuzzy boundaries. Figure 11 shows two examples of masses (benign and malignant).

The suspicious regions (benign or malignant) are marked by experienced radiologists who specified the center location and an approximate radius of a circle enclosing each

Table 3 Features used in this study

	Features	
Statistical texture features	First-Order Statistics (FOS)	Six features: Mean value of gray levels, Mean square value of gray levels, Standard Deviation, Variance, Skewness and Kurtosis. Total number: 6 features
	Gray-Level Co-occurrence Matrices (GLCM)	Nineteen features: Mean, Variance, Entropy, Contrast, Angular Second Moment (also called Energy), Dissimilarity, Correlation, Inverse Difference Moment (also called Homogeneity), Diagonal Moment, Sum Average, Sum Entropy, Sum Variance, Difference Entropy, Difference Mean, Difference Variance, Information Measure of Correlation 1, Information Measure of Correlation 2, Cluster Prominence and Cluster Shade. Eight values were obtained for each feature corresponding to the four directions ($\theta = 0°$, 45°, 90° and 135°) and to the two distances ($d = 1$ and 2). Total number: $19 \times 4 \times 2 = 152$ features
	Gray-Level Difference Matrices (GLDM)	Five features: Mean, Contrast, Angular Second Moment, Entropy and Inverse Difference Moment. Twenty values were computed for each feature corresponding to the four displacement vectors ($\delta = (0, d)$, $(-d, d)$, $(d, 0)$ and $(-d, -d)$) and to the five distances ($d = 1, 2, 3, 4$ and 5). Total number: $5 \times 4 \times 5 = 100$ features
	Gray-Level Run-Length Matrices (GLRLM)	Eleven features: Short Runs Emphasis (SRE), Long Runs Emphasis (LRE), Gray-Level Non-uniformity (GLN), Run-Length Non-uniformity (RLN), Run Percentage (RP), Low Gray-Level Run Emphasis (LGRE), High Gray-Level Run Emphasis (HGRE), Short Run Low Gray-Level Emphasis (SRLGE), Short Run High Gray-Level Emphasis (SRHGE), Long Run Low Gray-Level Emphasis (LRLGE) and Long Run High Gray-Level Emphasis (LRHGE). Two values were computed for each feature corresponding to the angles of 0° and 90° (horizontal and vertical directions). Total number: $11 \times 2 = 22$ features
	Tamura features	Three features: Coarseness, Contrast and Direction. Total number: 3 features
Frequential texture features	Gabor transform	Five scales (1/2, 1/3, 1/4, 1/5 and 1/8) and six orientations (0, $\pi/6$, $\pi/3$, $\pi/2$, $2\pi/3$ and $5\pi/6$) are used. The features vector f is created using μ_{mn} and σ_{mn} as the feature components. Total number: $5 \times 6 \times 2 = 60$ features
	Two-dimensional wavelet transform	Eight wavelet coefficients and eight energy measures at different resolution levels. Total number: $8 + 8 = 16$ features
Shape features	Hu's invariant moments	Seven features: First Moment, Second Moment, Third Moment, Fourth Moment, Fifth Moment, Sixth Moment and Seventh Moment. Total number: 7 features
	Other features	Fourteen features: Area, Perimeter, Compactness, Aspect ratio, Major Axis Length, Minor Axis Length, Eccentricity, Orientation, Convex polygon area, Euler Number, Equivalent diameter, Solidity, Extent and Mean Intensity. Total number: 14 features

Table 4 Examples of FOS features

Features	ROI of Mdb001.pgm (Benign)	ROI of Mdb028.pgm (Malignant)
Mean	133.36	184.72
Mean square	238.02	255
Standard Deviation	$0 + 132.46i$	$0 + 184.02i$
Variance	-17547	-33865
Skewness	$0 - 0.096059i$	$0 - 0.0015662i$
Kurtosis	0.14268	0.00065438

abnormality. We used this information and a graphic editing tool to locate and cut the smallest square that contains the marked suspicious region. Each ROI contains only one mass. After running this step, we obtain manually for each mammogram an approximate binary mask (zero for black and one for white). Figure 12c shows one of the outputs acquired from this operation by using Fig. 12a as input. In Fig. 12c, the white region indicates the mass and the black region indicates the background. The resulting binary image will be useful in shape description.

By multiplying the mask image with the origin image, we obtain the segmented mass as shown in Fig. 13.

Table 5 Examples of GLCM features ($d=1$ and $\theta=0°$)

Features	ROI of Mdb001.pgm (Benign)	ROI of Mdb028.pgm (Malignant)
Mean	134.32	185.8
Variance	4445.8	580.64
Entropy	10.604	9.2733
Contrast	195.97	6.704
Angular Second Moment	0.001256	0.002924
Dissimilarity	2.6367	1.7896
Correlation	0.97796	0.99423
Inverse Difference Moment	0.43164	0.43003
Diagonal Moment	58.412	31.654
Sum Average	266.63	369.61
Sum Entropy	8.4392	7.1642
Sum Variance	17587	2315.9
Difference Entropy	2.5552	2.4776
Difference Mean	2.6367	1.7896
Difference Variance	189.01	3.5013
Information Measure of Correlation 1	-0.57913	-0.49988
Information Measure of Correlation 2	0.99991	0.99896
Cluster Prominence	6.8848e+08	1.1895e+07
Cluster Shade	-1795600	-77908

Table 6 Examples of GLDM features

Features	ROI of Mdb001.pgm (Benign)	ROI of Mdb028.pgm (Malignant)
Mean $(0, d)$	2.0043e+07	1.5826e+06
Contrast $(0, d)$	3.4291e+09	2.7073e+08
Angular Second Moment $(0, d)$	6.1028e+12	3.8073e+10
Entropy $(0, d)$	$-4.7067e+08$	$-2.9282e+07$
Inverse Difference Moment $(0, d)$	63805	5192.5
Mean $(-d, d)$	2.0039e+07	1.5817e+06
Contrast $(-d, d)$	3.4291e+09	2.7072e+08
Angular Second Moment $(-d, d)$	6.0774e+12	3.765e+10
Entropy $(-d, d)$	$-4.6941e+08$	$-2.9062e+07$
Inverse Difference Moment $(-d, d)$	60786	4460.3
Mean $(d, 0)$	2.009e+07	1.5823e+06
Contrast $(d, 0)$	3.437e+09	2.7073e+08
Angular Second Moment $(d, 0)$	6.1285e+12	3.7919e+10
Entropy $(d, 0)$	$-4.7176e+08$	$-2.9194e+07$
Inverse Difference Moment $(d, 0)$	65286	4646.8
Mean $(-d, -d)$	2.0036e+07	1.5823e+06
Contrast $(-d, -d)$	3.429e+09	2.7073e+08
Angular Second Moment $(-d, -d)$	6.0559e+12	3.7878e+10
Entropy $(-d, -d)$	$-4.6822e+08$	$-2.9174e+07$
Inverse Difference Moment $(-d, -d)$	56340	4724.9

Textural and shape features are calculated and extracted from the cropped ROIs. The features used in this study are listed in Table 3. For each selected ROI, the total number of computed features is $6 + 152 + 100 + 22 + 3 + 60 + 16 + 7 + 14 = 380$.

Tables 4, 5, 6, 7, 8, 9, 10, 11 and 12 show the obtained texture and shape features of two examples of selected ROIs (ROI of "Mdb001.pgm" and ROI of "Mdb028.pgm").

To extract Gabor transform features, five scales (1/2, 1/3, 1/4, 1/5 and 1/8) and six orientations (0, π/6, π/3, π/2, 2π/3

Table 7 Examples of GLRLM features

Features	ROI of Mdb001.pgm (Benign)	ROI of Mdb028.pgm (Malignant)
Short Runs Emphasis	0.3059	0.43212
Long Runs Emphasis	256.94	25.689
Gray-Level Non-uniformity	1613.7	280.46
Run-Length Non-uniformity	1806.1	711.52
Run Percentage	0.11259	0.28934
Low Gray-Level Run Emphasis	0.050656	0.021487
High Gray-Level Run Emphasis	104.58	130.8
Short Run Low Gray-Level Emphasis	407.97	40.79
Short Run High Gray-Level Emphasis	546.03	771.33
Long Run Low Gray-Level Emphasis	407.97	40.79
Long Run High Gray-Level Emphasis	458640	45856

Table 8 Examples of Tamura features

Features	ROI of Mdb001.pgm (Benign)	ROI of Mdb028.pgm (Malignant)
Coarseness	54.2441	34.3909
Contrast	54.6442	19.7844
Direction	0.3352	0.8865

Table 9 Examples of Gabor features

Features	ROI of Mdb001.pgm (Benign)	ROI of Mdb028.pgm (Malignant)
μ_{00}	0.87091	0.33105
σ_{00}	2854800	3264.6
μ_{01}	1.0883	0.49597
σ_{01}	3408300	8025.7
μ_{02}	1.5626	0.86589
σ_{02}	3261200	21590

Table 10 Examples of wavelet features

Features	ROI of Mdb001.pgm (Benign)	ROI of Mdb028.pgm (Malignant)
Wavelets coefficients	58070	20668
	230.75	94.563
	751.56	67.018
	105.26	38.718
	45.152	16.632
	81.859	35.554
	35.696	18.86
	64.009	28.533
Energy measures	$-3.5587\mathrm{e}+10$	$-4.4883\mathrm{e}+09$
	-463990	-18817
	-2538200	-9103.5
	-45653	-956.58
	8742.7	759.4
	-16012	-397.88
	8568.2	622.67
	1222	398.35

Table 11 Examples of Hu's invariant moments

Features	ROI of Mdb001.pgm (Benign)	ROI of Mdb028.pgm (Malignant)
First Moment	7.1105	7.3297
Second Moment	22.412	28.899
Third Moment	25.16	37.159
Fourth Moment	27.902	34.361
Fifth Moment	54.713	70.852
Sixth Moment	40.634	48.911
Seventh Moment	54.706	74.143

Table 12 Examples of other shape features

Features	ROI of Mdb001.pgm (Benign)	ROI of Mdb028.pgm (Malignant)
Area	7	9482
Perimeter	10.243	495.39
Compactness	5.7735	119.22
Aspect ratio	2.43	103.44
Major Axis Length	0.90711	0.49717
Minor Axis Length	45	81.074
Eccentricity	9	10133
Orientation	1	-4
Convex polygon area	2.9854	109.88
Euler Number	0.77778	0.93575
Equivalent diameter	0.4375	0.72492
Solidity	1.7143	1
Extent	0.83846	0.48554
Mean Intensity	1	0.90833

and 5π/6) are used. Figures 14 and 15 show, respectively, the obtained filter banc and two examples of filtered images.

Applying levels of decomposition produced a large number of wavelet coefficients. So it was decided to use only two-level decomposition to reduce complexity. Figure 16
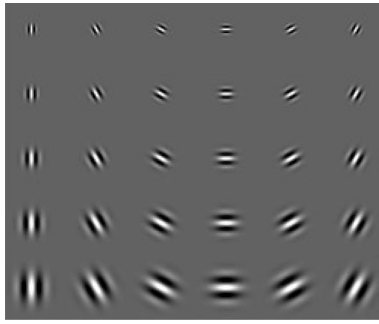
Fig. 14 Filter banc used in this study
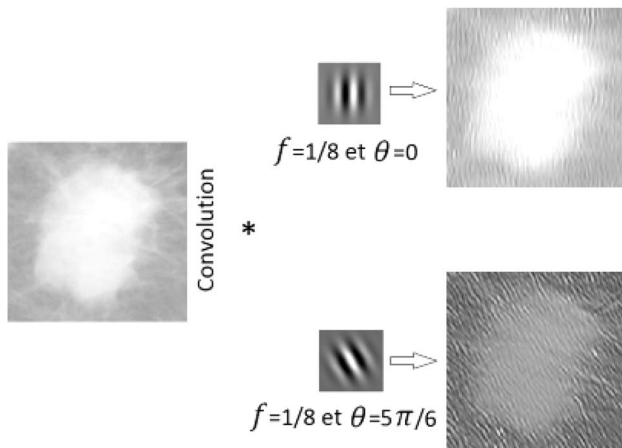
Fig. 15 Two examples of filtered images (ROI of "Mdb069.pgm")

shows an example of a two-level decomposition through the Haar wavelet transform.

With the aim of achieving optimum discrimination, a normalization procedure can be performed by assigning a
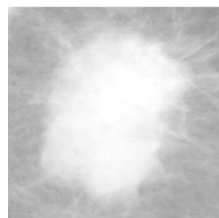
weight to all features in order to measure their similarity on the same basis. The technique used was to project each feature onto a unit sphere. Features values are normalized between 0 and 1 according to the following formula:

$$Y = (X - \text{Min})/(\text{Max} - \text{Min}) \tag{110}$$

where X is the initial feature value and Y is the feature value after normalization.

The next step is to select the most relevant features. To further explore our search space and to improve the optimal solution, we applied TS and GA ten times for feature selection, and each time the results are slightly changed from previous. Tables 13 and 14 show the selected features for each iteration. (Selected and rejected features are represented by 1 and 0, respectively.)

Then, the obtained selected features (for both TS and GA) are ordered with respect to the number of occurrences of each feature in the 10 rounds and its score. The best solution corresponds to the highest score for both TS and GA. Tables 15 and 16 show the number of occurrences of each feature in the ten iterations.

The weight of each feature is calculated by applying ReliefF algorithm to the feature sets. Table 17 shows the obtained weights of HU's invariant moments. Most relevant features correspond to the most high weights. In this case, $\{Mom_i, i = 1, \ldots, 4\}$ are the selected features from HU's invariant moments.

Table 18 shows the selected features using SFS and SBS for GLRLM features. (Selected and rejected features are represented by 1 and 0, respectively.)

Our experiment involved a binary classification, in which the classifier should recognize, whether a given ROI contains benign or malignant tissue. The goal of the experiments is to test which descriptor is best for describing mammographic images, i.e., to find out which combination of

Fig. 16 Two-level decomposition of mammogram (ROI of "Mdb069.pgm") through the Haar wavelet transform
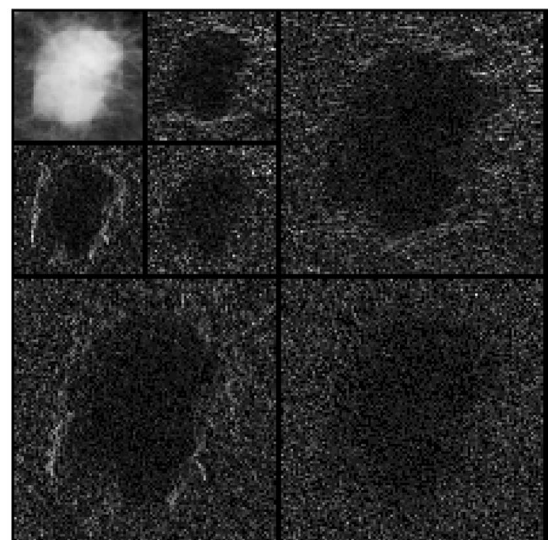
Table 13 Results of TS feature selection (FOS features)

Feature/Round (R)	Mean	Mean square	Standard Deviation	Variance	Skewness	Kurtosis	Cost
R1	1	1	0	1	0	1	0.1562
R2	1	1	0	1	0	0	0.1558
R3	1	0	0	1	0	0	0.1540
R4	1	0	1	1	0	0	0.1545
R5	1	1	0	1	0	1	0.1562
R6	1	1	0	1	0	1	0.1562
R7	1	1	1	1	0	0	0.1563
R8	1	1	1	1	0	0	0.1563
R9	1	1	1	1	0	0	0.1563
R10	1	1	1	1	0	0	0.1563

Table 14 Results of GA feature selection (GLDM features)

Feature/Round (R)	Mean $(0, d)$	Cont $(0, d)$	ASM $(0, d)$	Ent $(0, d)$	IDM $(0, d)$	Cost
R1	1	0	1	1	0	-0.4541
R2	1	0	1	1	0	-0.4539
R3	1	0	1	1	0	-0.4540
R4	1	0	1	1	0	-0.4535
R5	0	1	1	1	0	-0.4528
R6	1	0	1	1	0	-0.4544
R7	1	0	1	1	0	-0.4546
R8	0	0	1	1	0	-0.4532
R9	1	0	1	1	0	-0.4536
R10	1	0	1	1	0	-0.4540

Table 15 Selected features and the number of times it is selected by TS

Feature	Mean	Variance	Mean square	Standard Deviation	Kurtosis	Skewness
Number of occurrences	10	10	8	5	3	0

Table 16 Selected features and the number of times it is selected by GA

Feature	Mean $(0, d)$	Cont $(0, d)$	ASM $(0, d)$	Ent $(0, d)$	IDM $(0, d)$	Mean $(-d, d)$	Cont $(-d, d)$	ASM $(-d, d)$	Ent $(-d, d)$	IDM $(-d, d)$
Number of occurrences	10	10	10	10	8	1	1	0	0	0

Table 17 Weights of HU's invariant moments using ReliefF algorithm

Feature	First Moment	Second Moment	Third Moment	Fourth Moment	Fifth Moment	Sixth Moment	Seventh Moment
Weight	-0.0037	-0.0068	-0.0070	-0.0078	-0.0092	-0.0101	-0.0282

Table 18 Selected features using SFS and SBS (GLRLM features)

Feature	LRE	SRE	GLN	RLN	RP	LGRE	HGRE	SRLGLE	SRHGLE	LRLGLE	LRHGLE
Weight											
SFS	1	1	0	0	0	0	1	1	1	0	0
SBS	0	0	0	0	1	1	1	0	1	1	1

Table 19 Performance obtained for FOS features

Selection method	Benign class		Malignant class		Total		
	CC	CV	CC	CV	TCC	WACC	WACV
TS	0	0.1016	0.9853	0.1409	0.9853	0.4223	0.1184
GA	0	0.0746	0.9559	0.0898	0.9559	0.4097	0.0811
ReliefF	0	0.1109	0.9559	0.1487	0.9559	0.4097	0.1271
SFS	0	0.0181	0.0294	0	0.0294	0.0126	0.0104
SBS	0	0.1451	0.9559	0.2051	0.9559	0.4097	0.1708

Table 20 Performance obtained for GLCM features

Selection method	d	θ	Benign class		Malignant class		Total		
			CC	CV	CC	CV	TCC	WACC	WACV
TS	1	0	0	0.1209	0.8824	0.2301	0.8824	0.3782	0.1677
		45	0	0.1234	0.8529	0.2942	0.8529	0.3655	0.1966
		90	0	0.1243	0.8824	0.2472	0.8824	0.3782	0.1770
		135	0	0.1194	0.8529	0.2512	0.8529	0.3655	0.1759
	2	0	0	0.1187	0.8676	0.2625	0.8676	0.3718	0.1803
		45	0	0.1206	0.8529	0.2750	0.8529	0.3655	0.1867
		90	0	0.1294	0.8529	0.2743	0.8529	0.3655	0.1915
		135	0.0196	0.1231	0.8529	0.2708	0.8725	0.3768	0.1864
GA	1	0	0.0196	0.0836	0.9118	0.2323	0.9314	0.4020	0.1473
		45	0.0196	0.0843	0.9412	0.2217	0.9608	0.4146	0.1432
		90	0	0.0877	0.9118	0.2357	0.9118	0.3908	0.1511
		135	0.0196	0.0887	0.9412	0.2217	0.9608	0.4146	0.1457
	2	0	0.0196	0.0979	0.9265	0.2359	0.9461	0.4083	0.1571
		45	0.0196	0.0959	0.9265	0.2154	0.9461	0.4083	0.1472
		90	0	0.0951	0.9118	0.2378	0.9118	0.3908	0.1562
		135	0.0196	0.0958	0.9265	0.2163	0.9461	0.4083	0.1474
ReliefF	1	0	0.0196	0.0970	0.8971	0.3402	0.9167	0.3957	0.2012
		45	0.0196	0.1171	0.8824	0.3666	0.9020	0.3894	0.2240
		90	0.0392	0.1235	0.8824	0.3372	0.9216	0.4006	0.2151
		135	0.0196	0.1054	0.8824	0.3509	0.9020	0.3894	0.2106
	2	0	0.0196	0.0999	0.8971	0.3463	0.9167	0.3957	0.2055
		45	0.0196	0.1063	0.8824	0.3599	0.9020	0.3894	0.2149
		90	0.0196	0.1168	0.8971	0.3190	0.9167	0.3957	0.2035
		135	0.0196	0.1103	0.8824	0.3667	0.9020	0.3894	0.2202
SFS	1	0	0	0.1037	0.7794	0.1869	0.7794	0.3340	0.1394
		45	0.0784	0.0339	0.9118	0.3674	0.9902	0.4356	0.1768
		90	0	0.1138	0.8971	0.2578	0.8971	0.3845	0.1755
		135	0	0.0755	0.4706	0.0427	0.4706	0.2017	0.0614
	2	0	0.0784	0.0347	0.8971	0.3797	0.9755	0.4293	0.1826
		45	0.0392	0.0891	0.9118	0.3053	0.9510	0.4132	0.1818
		90	0.0392	0.1003	0.8529	0.3570	0.8922	0.3880	0.2103
		135	0.0196	0.1280	0.6912	0.3625	0.7108	0.3074	0.2285
SBS	1	0	0	0.1185	0.8529	0.2732	0.8529	0.3655	0.1848
		45	0.0392	0.1320	0.8382	0.3432	0.8775	0.3817	0.2225
		90	0.0196	0.1249	0.8971	0.2700	0.9167	0.3957	0.1871
		135	0.0196	0.1200	0.8676	0.2797	0.8873	0.3831	0.1884
	2	0	0	0.1336	0.8824	0.3332	0.8824	0.3782	0.2191
		45	0.0196	0.1305	0.8235	0.3059	0.8431	0.3641	0.2057
		90	0	0.1281	0.9118	0.2535	0.9118	0.3908	0.1819
		135	0	0.1313	0.8529	0.3043	0.8529	0.3655	0.2055

Table 21 Performance obtained for GLDM features

Selection method	d	Benign Class		Malignant Class		Total		
		CC	CV	CC	CV	TCC	WACC	WACV
TS	1	0	0.0579	0.9559	0.9300	0.9559	0.4097	0.4317
	2	0	0.0579	0.9559	0.9299	0.9559	0.4097	0.4316
	3	0	0.0581	0.9559	0.9295	0.9559	0.4097	0.4316
	4	0	0.0582	0.9559	0.9290	0.9559	0.4097	0.4314
	5	0	0.0585	0.9559	0.9284	0.9559	0.4097	0.4313
GA	1	0	0.0521	0.9559	0.9471	0.9559	0.4097	0.4357
	2	0	0.0521	0.9559	0.9471	0.9559	0.4097	0.4357
	3	0	0.0523	0.9559	0.9467	0.9559	0.4097	0.4356
	4	0	0.0527	0.9559	0.9451	0.9559	0.4097	0.4352
	5	0	0.0530	0.9559	0.9443	0.9559	0.4097	0.4350
ReliefF	1	0	0.0520	0.9559	0.9471	0.9559	0.4097	0.4356
	2	0	0.0512	0.9559	0.9486	0.9559	0.4097	0.4358
	3	0	0.0515	0.9559	0.9482	0.9559	0.4097	0.4358
	4	0	0.0507	0.9559	0.9491	0.9559	0.4097	0.4357
	5	0	0.0510	0.9559	0.9485	0.9559	0.4097	0.4356
SFS	1	0	0.0641	0.9559	0.9090	0.9559	0.4097	0.4262
	2	0	0.0265	0.9559	0.9896	0.9559	0.4097	0.4392
	3	0	0.0265	0.9559	0.9896	0.9559	0.4097	0.4392
	4	0	0.0265	0.9559	0.9894	0.9559	0.4097	0.4392
	5	0	0.0643	0.9559	0.9090	0.9559	0.4097	0.4263
SBS	1	0	0.0632	0.9559	0.9117	0.9559	0.4097	0.4269
	2	0	0.0635	0.9559	0.9106	0.9559	0.4097	0.4265
	3	0	0.0627	0.9559	0.9132	0.9559	0.4097	0.4272
	4	0	0.0624	0.9559	0.9142	0.9559	0.4097	0.4274
	5	0	0.0628	0.9559	0.9124	0.9559	0.4097	0.4269

Table 22 Performance obtained for GLRLM features

Selection method	Direction	Benign Class		Malignant Class		Total		
		CC	CV	CC	CV	TCC	WACC	WACV
TS	H	0.019	0.175	0.911	0.487	0.931	0.402	0.309
	V	0.039	0.1662	0.955	0.338	0.995	0.432	0.239
GA	H	0.019	0.158	0.955	0.417	0.975	0.420	0.269
	V	0.039	0.088	0.985	0.324	1.024	0.444	0.189
ReliefF	H	0.039	0.160	0.941	0.513	0.980	0.425	0.311
	V	0.039	0.127	0.926	0.341	0.965	0.419	0.219
SFS	H	0.019	0.167	0.897	0.525	0.916	0.395	0.321
	V	0	0.205	0.897	0.335	0.897	0.384	0.260
SBS	H	0.019	0.184	0.897	0.515	0.916	0.395	0.326
	V	0.039	0.138	0.970	0.341	1.009	0.438	0.225

Table 23 Performance obtained for Tamura features

Selection method	Benign class		Malignant class		Total		
	CC	CV	CC	CV	TCC	WACC	WACV
TS	0	0.1734	0.9163	0.4592	0.9323	0.3874	0.2621
GA	0	0.2204	0.9559	0.5219	0.9559	0.4097	0.3496
ReliefF	0.0392	0.0902	0.7353	0.4119	0.7745	0.3375	0.2281
SFS	0.0392	0.1929	0.4706	0.1837	0.5098	0.2241	0.1890
SBS	0.0196	0.2348	0.9412	0.4148	0.9608	0.4146	0.3119

Table 24 Performance obtained for Gabor features

Selection method	Benign class		Malignant class		Total		
	CC	CV	CC	CV	TCC	WACC	WACV
TS	0	0.1324	0.9559	0.4799	0.9559	0.4097	0.2814
GA	0	0.1131	0.9265	0.6256	0.9265	0.3971	0.3327
ReliefF	0	0.0856	0.9265	0.8033	0.9265	0.3971	0.3932
SFS	0	0.0553	0.9706	0.9254	0.9706	0.4160	0.4282
SBS	0	0.1220	0.9265	0.5457	0.9265	0.3971	0.3036

Table 25 Performance obtained for wavelet features

Selection method	Benign class		Malignant class		Total		
	CC	CV	CC	CV	TCC	WACC	WACV
TS	0.0196	0.1310	0.9412	0.5519	0.9608	0.4146	0.3114
GA	0.0392	0.0864	0.5294	0.4471	0.5686	0.2493	0.2410
ReliefF	0.0196	0.0786	0.9559	0.5854	0.9755	0.4209	0.2958
SFS	0.0784	0.0072	0	0.1865	0.0784	0.0448	0.0840
SBS	0	0.1407	0.9706	0.6530	0.9706	0.4160	0.3602

Table 26 Performance obtained for HU's invariant moments features

Selection method	Benign class		Malignant class		Total		
	CC	CV	CC	CV	TCC	WACC	WACV
TS	0	0.3836	0.8088	0.4074	0.8088	0.3466	0.3938
GA	0	0.4421	0.8382	0.4258	0.8382	0.3592	0.4351
ReliefF	0	0.3667	0.8088	0.4281	0.8088	0.3466	0.3930
SFS	0	0.3915	0.9559	0.3056	0.9559	0.4097	0.3547
SBS	0	0.3732	0.8088	0.3824	0.8088	0.3466	0.3771

Table 27 Performance obtained for other shape features

Selection method	Benign class		Malignant class		Total		
	CC	CV	CC	CV	TCC	WACC	WACV
TS	0	0.2694	0.4118	0.2440	0.4118	0.1765	0.2585
GA	0	0.2353	0.7941	0.2008	0.7941	0.3403	0.2205
ReliefF	0	0.2235	0.2941	0.2273	0.2941	0.1261	0.2251
SFS	0	0.2174	0.9412	0.1933	0.9412	0.4034	0.2071
SBS	0	0.2116	0.2353	0.1997	0.2353	0.1008	0.2065

Table 28 Performance obtained for all statistical texture features

Selection method	Benign class		Malignant class		Total		
	CC	CV	CC	CV	TCC	WACC	WACV
TS	0	0.1536	0.9559	0.6580	0.9559	0.4097	0.3698
GA	0	0.1009	0.9559	0.4942	0.9559	0.4097	0.2695
ReliefF	0	0.1281	0.9559	0.6731	0.9559	0.4097	0.3617
SFS	0	0.9118	0.1628	0.4123	0.9118	0.3908	0.2697
SBS	0	0.1437	0.9559	0.6154	0.9559	0.4097	0.3459

descriptors will give the best results. The effectiveness of the different feature sets is listed in Tables 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32 and 33.

The best performances obtained for the texture and shape features correspond to GLRLM features using GA and Hu's invariant moments using SFS, respectively. The GLRLM

Table 29 Performance obtained for all frequential texture features

Selection method	Benign class		Malignant class		Total		
	CC	CV	CC	CV	TCC	WACC	WACV
TS	0.0196	0.1475	0.9706	0.5172	0.9902	0.4272	0.3060
GA	0	0.1040	0.9118	0.5876	0.9118	0.3908	0.3113
ReliefF	0	0.1123	0.8971	0.6235	0.8971	0.3845	0.3314
SFS	0.0784	0.0072	0	0.1865	0.0784	0.0448	0.0840
SBS	0	0.1250	0.9412	0.6016	0.9412	0.4034	0.3293

Table 30 Performance obtained for all texture features

Selection method	Benign class		Malignant class		Total		
	CC	CV	CC	CV	TCC	WACC	WACV
TS	0	0.1576	0.9265	0.5499	0.9265	0.3971	0.3258
GA	0	0.1087	0.9265	0.5562	0.9265	0.3971	0.3005
ReliefF	0	0.1077	0.9118	0.6140	0.9118	0.3908	0.3247
SFS	0.0392	0.0608	0.9559	0.6260	0.9951	0.4321	0.3030
SBS	0	0.1456	0.9412	0.5996	0.9412	0.4034	0.3402

Table 31 Performance obtained for all shape features

Selection method	Benign class		Malignant class		Total		
	CC	CV	CC	CV	TCC	WACC	WACV
TS	0	0.3527	0.4118	0.3266	0.4118	0.1765	0.3415
GA	0	0.3061	0.8971	0.2588	0.8971	0.3845	0.2858
ReliefF	0	0.3502	0.9118	0.3504	0.9118	0.3908	0.3503
SFS	0	0.4097	0.8235	0.4274	0.8235	0.3529	0.4173
SBS	0	0.3501	0.7500	0.3382	0.7500	0.3214	0.3450

Table 32 Performance obtained for all features

Selection method	Benign class		Malignant class		Total		
	CC	CV	CC	CV	TCC	WACC	WACV
TS	0	0.2407	0.9412	0.5932	0.9412	0.4034	0.3918
GA	0	0.1586	0.9559	0.5255	0.9559	0.4097	0.3158
ReliefF	0	0.1119	0.9559	0.7472	0.9559	0.4097	0.3842
SFS	0	0.0570	0.9559	0.8099	0.9559	0.4097	0.3797
SBS	0	0.2030	0.9706	0.5849	0.9706	0.4160	0.3667

performance is more than performance obtained for GLRLM and Hu's invariant moments features as shown in Table 34.

Most relevant GLRLM and Hu's invariant moments features are GLNU, RLNU, LGRE, SRLGLE, LRLGLE and Second Moment.

These features are then passed through three combined classifiers which determine which category the region of interest falls into to benign or malignant. Fivefold cross-validation is performed in the classification phase. We used this approach because the dataset has a relatively small number of samples. Finally, evaluation results are performed using measures presented in the previous section. The confusion matrix for selected features is presented in Tables 35, 36 and 37.

Table 38 shows the obtained performance measures of selected features.

Table 39 depicts the classification precision and the area under ROC curve (AUC) of combined classifiers. The results show that the selected GLRLM features give best overall classification precision of 90.9%. Selected HU's invariant moments provide the worst performance with 72.7%. Also, it is obvious that the classification of selected GLRLM features and HU's invariant moments gives worse accuracy with an accuracy of 77.2%.

Table 33 Best features performance

Features	Input parameters	TCC		WACC		WACV
		Rank	Value	Rank	Value	
GLRLM (using GA)	Vertical	1	1.024	1	0.444	0.189
All Texture Features (using SFS)		2	0.995	3	0.432	0.303
GLCM (using SFS)	$d=1, \theta=45°$	3	0.990	2	0.435	0.176
All texture frequential features (using TS)		4	0.990	4	0.427	0.306
FOS (using TS)	None	5	0.985	5	0.422	0.118
Wavelets (using ReliefF)	None	6	0.975	6	0.420	0.295
All Features (using SBS)		7	0.970	7	0.416	0.366
Gabor filters (using SFS)	None	8	0.970	8	0.416	0.428
Tamura (using SBS)	None	9	0.960	9	0.414	0.311
All texture statistical features (using GA)		10	0.955	10	0.409	0.269
Hu's invariant moments (using SFS)	None	11	0.955	11	0.409	0.354
GLDM (using SFS)	$d=1$	12	0.955	12	0.409	0.426
Shape (using SFS)	None	13	0.941	13	0.403	0.207
All shape features (using ReliefF)		14	0.911	14	0.390	0.350

Table 34 Performance obtained for GLRLM and HU's invariant moments features

Selection method	Benign class		Malignant class		Total		
	CC	CV	CC	CV	TCC	WACC	WACV
TS	0.019	0.290	0.970	0.365	0.990	0.427	0.322
GA	0.039	0.234	0.955	0.342	0.995	0.432	0.281
ReliefF	0.019	0.274	0.970	0.367	0.990	0.427	0.314
SFS	0	0.406	0.941	0.404	0.941	0.403	0.405
SBS	0.019	0.275	0.955	0.368	0.975	0.420	0.315

Table 35 Confusion matrix for selected GLRLM features using GA

Actual	Predicted	
	Positive	Negative
Positive	19	0
Negative	2	1

Table 36 Confusion matrix for selected HU's invariant moments using SFS

Actual	Predicted	
	Positive	Negative
Positive	14	1
Negative	5	2

Table 37 Confusion matrix for selected GLRLM and HU's invariant moments features

Actual	Predicted	
	Positive	Negative
Positive	15	0
Negative	5	2

9 Conclusion

Digital mammography is the most common method for early breast cancer detection. Automated analysis of these images is very important, since manual analysis of these images is costly and inconsistent. In this paper, we made an analysis using nine different techniques for feature extraction. The proposed methodology is composed of two main stages that work in pipeline mode. The first stage consists of characterization and selection of the most relevant features. The second stage consists of classification of breast masses into benign and malignant. According to the provided examination, we can conclude that the best feature performance was achieved in the case of GLRLM descriptor. The obtained results can be used in other applications such as segmentation and content-based image retrieval.

Table 38 Performance measures

Criterion	GLRLM features	HU's invariant moments	GLRLM features and HU's invariant moments
Rate of Positive Predictions	0.863	0.681	0.681
Rate of Negative Predictions	0.136	0.318	0.318
True Positive Rate (Sensitivity)	0.904	0.736	0.75
False Negative Rate	0.095	0.263	0.25
False Positive Rate	0	0.333	0
True Negative Rate (Specificity)	1	0.66	1
Positive Predictive Value	1	0.933	1
Negative Predictive Value	0.33	0.28	0.28
Accuracy	0.909	0.727	0.772
Mathews Correlation Coefficient	0.549	0.297	0.462
F-measure	0.95	0.823	0.857
G-measure	0.951	0.7	0.866

Table 39 Evaluation results

Feature set	Most relevant features	AUC (%)	Precision (%)
GLRLM features	GLNU, RLNU, LGRE, SRLGLE and LRLGLE	66.66	90.9
HU's invariant moments	Second Moment	60.95	72.7
GLRLM features and HU's invariant moments	GLNU, RLNU, LGRE, SRLGLE, LRLGLE and Second Moment	64.28	77.2

References

1. Razavi AR, Gill H, Ahlfeldt H, Shahsavar N (2007) Predicting metastasis in breast cancer: comparing a decision tree with domain experts. J Med Syst 31:263–273
2. Smart CR, Hendrick RE, Rutledge JH, Smith RA (1995) Benefit of mammography screening in women ages 40 to 49 years: current evidence from randomized controlled trials. Cancer 75:1619–1626
3. Cady B, Michaelson JS (2001) The life-sparing potential of mammographic screening. Cancer 91:1699–1703
4. Bowes MP (2012) Digital mammography: process, guidelines, and potential advantages. eRadimaging. https://www.eradimaging.com/site/article.cfm
5. Tabar L, Fagerberg C, Gad A, Baldetorp L, Holmberg L, Grontoft O, Ljungquist U, Lundstrom B, Manson J, Eklund G et al (1985) Reduction in mortality from breast cancer after mass screening with mammography. Lancet 1:829–832
6. Bird RE, Wallace TW, Yankaskas BC (1992) Analysis of cancers missed at screening mammography. Radiology 184:613–617
7. American Cancer Society (2003) Cancer prevention and early detection facts and figures. American Cancer Society, Atlanta
8. American College of Radiology, ACR BI-RADS (2003) Mammography, ultrasound & magnetic resonance imaging, 4th edn. American College of Radiology, Reston
9. Bozek J, Mustra M, Delac K, Grgic M (2009) A survey of image processing algorithms in digital mammography. J Recent Adv Multimedia Signal Process Commun 231:631–657
10. Oliver A, Freixenet J, Mart J, Pérez E, Pont J, Denton ER et al (2010) A review of automatic mass detection and segmentation in mammographic images. Med Image Anal 14(2):87–110
11. Rojas Domnguez A, Nandi AK (2009) Toward breast cancer diagnosis based on automated segmentation of masses in mammograms. Pattern Recognit 42(6):1138–1148
12. Tang J, Rangayyan RM, Xu J, El Naqa I, Yang Y (2009) 'Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. Inf Technol Biomed IEEE Trans IEEE 13(2):236–251
13. Elter M, Horsch A (2009) CADx of mammographic masses and clustered microcalcifications a review. Med Phys Am Assoc Phys Med 36(6):2052–2068
14. Narváez F, Romero E (2012) Breast mass classification using orthogonal moments. In: Breast imaging. Springer, Berlin, pp 64–71
15. Haralick RM, Shanmugam K, Dinstein I (1973) Textural features for image classification. IEEE Trans Syst Man Cybernet SMC-3:610–621
16. Cheng HD, Shi XJ, Min R, Hu LM, Cai XP, Du HN (2006) Approaches for automated detection and classification of masses in mammograms. Pattern Recognit 39:646–668
17. Székely N, Tóth N, Pataki B (2006) A hybrid system for detecting masses in mammographic images. IEEE Trans Instrum Meas 55(3):944–952
18. Nunes AP, Silva AC, Paiva ACD (2010) Detection of masses in mammographic images using geometry, Simpson's Diversity Index and SVM. Int J Signal Imaging Syst Eng Indersci 3(1):40–51
19. Khuzi AM, Besar R, Zaki WW, Ahmad N (2009) Identification of masses in digital mammogram using gray level co-occurrence matrices. Biomed Imaging Interv J 5(3):e17
20. Gonzalez RC, Woods RE (2002) Digital image processing. Prentice-Hall Inc, New Jersey, pp 76–142
21. Galloway MM (1975) Texture classification using gray level run length. Comput Graph Image Process 4:172–179
22. Tamura H, Mori S, Yamawaki T (1978) Texture features corresponding to visual perception. IEEE Trans Syst Man Cybernet Smc-8(6):460–473

23. Manjunath BS, Ma WY (1996) Texture features for browsing and retrieval of large image data. IEEE Trans Pattern Anal Mach Intell (Spec Issue Digit Libr) 18(8):837–842

24. Rodrigues JF Jr, Traina AJM, Traina C Jr (2005) Enhanced visual evaluation of feature extractors for image mining. In: The 3rd ACS/IEEE international conference on computer systems and applications

25. Tahir MA, Bouridane A, Kurugollu F (2007) Simultaneous feature selection and feature weighting using Hybrid Tabu Search/K-nearest neighbor classifier. Pattern Recognit Lett 28:438–446

26. Duda RO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. Wiley, New York

27. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238

28. Vapnik V (1995) The nature of statistical learning theory. Springer, Berlin

29. Sun Y, Lou X, Bao B (2011) A novel relief feature selection algorithm based on mean-variance model. J Inf Comput Sci 8(16):3921–3929

30. Shalizi C (2010) Principal Component Analysis. Lecture Notes, 36–490, https://www.stat.cmu.edu/~cshalizi/490/10/pca/pca-handout.pdf.

31. Suckling J et al (1994) 'The mammographic image analysis society digital mammogram database. In: Exerpta Medica, international congress series 1069, pp 375–378

32. Robnik-Šikonja M, Kononenko I (2003) Theoretical and empirical analysis of ReliefF and RReliefF. Mach Learn J 53:23–69