

# Improved Ultrasound Based Computer Aided Diagnosis System for Breast Cancer Incorporating a New Feature of Mass Central Regularity Degree (CRD)

Ali Al-Yousef and Sandhya Samarasinghe

**Abstract** Ultrasound is one of the most frequently used methods for early detection of breast cancer. Currently, the accuracy of Computer Aided Diagnosis (CAD) systems based on ultrasound images is about 90 % and needs further enhancement in order to save lives of the undetected. A meaningful approach to do this is to explore new and meaningful features with effective discriminating ability and incorporate them into CAD systems. Some of the most powerful features used in cancer detection are based on the gross features of mass (e.g., shape and margin) that are subjectively evaluated. Recently, from an extensive investigation of ultrasound images, we extracted an easily quantifiable and easily measurable new geometric feature related to the mass shape in ultrasound images and called it Central Regularity Degree (CRD) as an effective discriminator of breast cancer. This feature takes into account a consistent pattern of regularity of the central region of the malignant mass. To demonstrate the effect of CRD on differentiating malignant from benign masses and the potential improvement to the diagnostic accuracy of breast cancer using ultrasound, this chapter evaluates the diagnostic accuracy of different classifiers when the CRD was added to five powerful mass features obtained from previous studies including one geometric feature: Depth-Width ratio (DW); two morphological features: shape and margin; blood flow and age. Feed forward Artificial Neural Networks (ANN) with structure optimized by SOM/Ward clustering of correlated weighted hidden neuron activation, K-Nearest Neighbour (KNN), Nearest Centroid and Linear Discriminant Analysis (LDA) were employed for classification and evaluation. Ninety nine breast sonograms—46 malignant and 53 benign—were evaluated. The results reveal that CRD is an effective feature discriminating between malignant and benign cases leading to improved accuracy of diagnosis of breast cancer. The best results were obtained by ANN where accuracy for training and testing using all features except CRD was 100 and 81.8 %, respectively, and 100 and 95.45 % using all features. Therefore, the overall improvement by adding CRD was about 14 %, a significant improvement.

---

A. Al-Yousef · S. Samarasinghe (✉)

Integrated Systems Modelling Group, Lincoln University, Christchurch, New Zealand  
e-mail: sandhya.samarasinghe@lincoln.ac.nz

**Keywords** Computer aided diagnostic system • Breast Cancer • Ultrasound • New mass feature • MLP-SOM model • Statistical Classifiers

## 1 Introduction

Breast tumors including benign and malignant solid masses start with few cells and grow faster than the normal cells and form a mass inside the breast tissues. A typical growing solid benign mass pushes and compresses the surrounding tissues to create a space which makes a well-defined boundary for the mass. Benign masses in most cases take an oval shape where the large axis of the oval is parallel to the skin line. Usually, the benign masses block the blood vessels which reduce the blood flow into the mass. Most of the diagnosed benign masses are fibroadinoma. Another example of benign tumors is cysts which are a fluid-filled mass with a clear margin. On the other hand, malignant masses grow and invade the surrounding tissues and not restricted to a limited area resulting in an irregular shape and ill-defined margin [1]. They also have an increased level of blood flow.

There are several changes in the breast anatomy during a female's life. In some cases, calcium salt gets deposited in the soft tissues that makes them harder; this is called calcification. Several factors contribute to this including age, past trauma to breast and breast inflammation. Also, calcification in a few cases is a sign of breast carcinoma. There are two types of calcification: the first is macro-calcification, where the size of calcium mass is large and probably benign. The second one is micro-calcification, where the size of calcium deposit is very small, less than 0.5 mm, and this type is probably malignant and requires further checking for micro-calcification morphology and distribution [2].

Ultrasound is one of the most frequently used methods for early detection of breast cancer. This method uses inaudible sound waves. The tumours are divided into four types according to their ability to return the sound waves (echogenicity). The first is anechoic lesions, and has no internal echo. This type of lesion is a benign cyst. The second one is hypoechoic lesions and has low level echoes throughout the lesion and is probably malignant. The third type is hyperechoic lesions with increased echogenicity compared to their surrounding fat, and are probably benign. The last type is isoechoic lesions with similar echogenicity compared to the fat echogenicity, and are probably benign [3, 4].

An ultrasound image is processed carefully to differentiate malignant from benign masses. Radiologists extract a number of mass features from an ultrasound image such as, shape, margin, orientation, echogenic pattern, posterior acoustic (shadow) features, effect on surrounding parenchyma and Calcifications (Table 1) [3, 5]. All these features are currently used to differentiate benign from malignant masses. Table 1 describes characteristics of each feature that help achieve this aim.

There are two main types of mass features in the image; combined morphological and geometric features and texture features. Morphological features are related to the structure of the mass such as, shape, margin, blood flow and the

**Table 1** Ultrasound mass features and a brief description of each feature [19]

Feature	Description
Shape	Mass takes two main shapes, regular or irregular; the regular masses (or oval) are probably benign and irregular masses (speculated) are probably malignant
Margin	The border that separates the mass from the neighbouring normal tissue; it can be clear or well defined, which is suggestive of benign or it can be blurry or ill-defined which is probably malignant.
Orientation	The long axis of the mass can be parallel to the skin line, which is suggestive of benign
Posterior acoustic	The shadow behind the mass usually caused by malignant lesion
Echogenic pattern	This feature reflects the internal mass echogenicity (indication of density of the mass); hypochoic lesions with low echogenicity are suggestive of malignant
Surrounding tissue	The effect of the mass on the surrounding tissues; depends on the mass type; for example, solid benign or malignant mass may compact the neighbouring tissues
Calcification	The presence of calcium inside the breast tissues; microcalcification is suggestive of malignant and macrocalcification is suggestive of benign
Blood flow	The speed of blood flow inside the mass tissues; high speed is suggestive of malignant
Envelop	Most benign masses are enveloped or partially enveloped

pattern of mass [6]. Geometric features, such as, volumes, spaces and measurements, can also be used to assess morphological features of the mass to extract new mass features. On the other hand, texture features are related to the arrangement of colours and other intrinsic pattern regularities, smoothness, etc., in the ultrasound image. This type of features are harder to extract as they require image processing algorithms to analyse the image.

Several studies have used texture features of ultrasound images for building Breast Cancer-Computer Aided Diagnosis Systems (BC-CAD). For example, 28 texture descriptors related to the intensities of the pixels in the Region Of Interest (ROI) including: the Mean of the Sum of the Average of intensities, Range of the Sum of the Entropy of pixels, pixel correlation,...,etc., and two Posterior Acoustic Attenuation Descriptors (PAAD) including—the difference between the average gray level within (ROI) and within regions of  $32 \times 32$  pixels inside the (ROI), and the difference between the gray level in a region of  $32 \times 32$  pixels inside the (ROI) and the average of the average of gray levels in the two adjacent  $32 \times 32$  pixel regions to the left and right of ROI—were extracted and used for differentiating a cyst from solid mass [7]. Stepwise logistic regression was used for feature selection and the study obtained 95.4 % classification accuracy using Mean of the Sum of Average of intensities, Range of the Sum of Entropy of pixels and the second PAAD descriptor. In another study, the co-variance of pixel intensity in the Region Of Interest (ROI) and the pixel similarity were used to differentiate benign from malignant masses in ultrasound image with 95.6 % classification accuracy [8].

Another study [9] used the shape and texture features of the mass including: eccentricity (the eccentricity of the ellipse that has the same second moment as the mass region); solidity of the mass, ratio of the number pixels in the mass to that of the convex including the mass; convex hull's area minus convex rectangular area that contain the convex; Cross correlation left and cross correlation right (i.e., cross correlation value between the convex rectangular area that contain the convex and the left side and right side areas, respectively) to classify the masses into benign and malignant. These features were used as inputs to a Support Vector Machine (SVM) to obtain 95 % classification accuracy. Furthermore, texture features including auto-covariance coefficients of pixel intensities, smoothness of the mass and block difference of inverse probabilities were used with Support Vector Machines (SVM) in another study [6] that produced 94.4 % classification accuracy. Another study [10] extracted eight texture features, three geometric features and two pixel histogram features, and used them as input to a Fuzzy Support Vector Machine (FSVM), Artificial Neural Networks (ANN) and SVM to classify 87 lesions into benign and malignant. The FSVM obtained the best results where the accuracy was 94.2 % followed by the ANN with 88.51 % diagnostic accuracy.

Compared with texture features, there are few studies that used the morphological features for building BC-CADs. For example, shape, orientation, margin, lesion boundary, echo pattern, and posterior acoustic features were used as inputs for a BC-CAD with 91.7 % classification accuracy (88.89 % sensitivity and 92.5 % specificity) [11]. Also, another study [12] extracted 6 morphological features of the mass shape including: Form Factor, Roundness, Aspect Ratio, Convexity, Solidity and Extent; all of these features were based on a maximum and minimum diameters of the mass and area, convex and perimeter of the mass. These features were used with a Support Vector Machine and obtained 91.7 % classification accuracy (90.59 % sensitivity and 92.22 % specificity). Another study [13] extracted seven morphological features including the number of substantial protuberances and depressions, lobulation index, elliptic-normalized circumference, elliptic-normalized skeleton, long axis to short axis ratio, depth-to-width ratio, and size of the lesion. Multilayer Perceptron MLP neural networks was used for classification. The accuracy, sensitivity, specificity measures were  $92.95 \pm 2 \%$ ,  $93.6 \pm 3 \%$  and  $91.9 \pm 5.3 \%$ , respectively.

Although texture features are good in differentiating benign from malignant, their values depend on the settings of the ultrasound machine and require extensive processing of ultrasound images. On the other hand, the morphological features are system independent and therefore, they are easy to incorporate into CAD systems and more commonly used in diagnosis systems [14].

However, the accuracy of current CAD systems based on the morphological features of ultrasound images is around 90 % and needs further enhancement to save the lives of the undetected. A meaningful approach to do this is to explore new and meaningful features with enhanced discriminating power and incorporate them into CAD systems.

Mass shape is one of the most useful features in breast cancer CAD systems and has been used as an input for classifiers in several studies [11, 15–17]. This feature

takes two main forms; regular or irregular shape. Irregular mass is indicative of malignancy; however, there are several types of benign mass that also take an irregular shape. Greater clarity on this issue may provide better ways to improve breast cancer diagnostic accuracy.

The aim of this chapter is to enhance the diagnostic accuracy of ultrasound based CAD systems based on morphological features by a thorough investigation of the shape of the mass in ultrasound images in order to extract a new mass feature with greater discriminating power and employ this feature in a CAD system along with few carefully selected morphological features. In doing so, we expand our work reported in [19].

## 2 Materials and Methods

This study evaluates 99 cases; 46 malignant and 53 benign. All cases were obtained from The Digital Database for Breast Ultrasound Images (DDBUI) [18]. The Second Affiliated Hospital of Harbin Medical University collected all images from 2002 to 2007. Each case in the database contains 1–6 images and a text file (Table 2) that lists important information of the patient and the lesions, such as, age, family history, shape, margin, size, blood flow, echo, micro-calcification number and shape. All these features were obtained by five experts.

**Table 2** The contents of the text file in the database

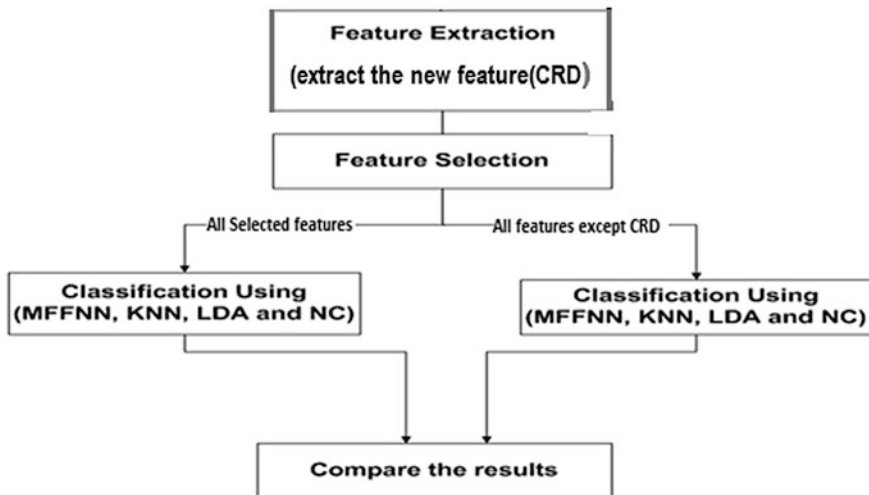
Assessment: Malignant
Ultrasound Doctor’s Depiction:
Assessment Sex: Female
Age: 32
Family History Criterion: none
Shape: Not Regular
Border: Blur
Echo: Unequable
Envelope: None
Side Echo: None
Micro-calcification Shape: None,
Micro-calcification Number: 0–0
Reduction: Has
Lymph Transformation: None
Blood: I or III Level
Mass Maximal Diameter: More than 2
Mammogram Total Number: 0
Radiologist Assessment: None
Biopsy Result: None
Biopsy Description:
Operation Result: Benign
Operation Description: breast fibroadenoma

## 2.1 Methods

The methodology is primarily aimed at extracting a new feature and assessing its effectiveness following four main steps shown in Fig. 1. The first step is feature extraction where all features indicative of breast cancer is extracted. Importantly, this step also contains the extraction of the proposed new feature related to the mass shape and called Central Regularity Degree (CRD) (see Sect. 2.1.1). All features are normalised and then, the features that are strongly related to breast cancer are selected for classification using hierarchical clustering and self organizing maps (SOM). Next, the selected features are used as inputs to four classifier systems based on neural networks and statistical methods (MFFNN—multi-layer feed forward neural network; KNN—K-nearest neighbour; LDA—Linear Discriminant Analysis; NC—Nearest Centroid). Here, two sets of inputs are used: a set without the new feature and a set including the new feature. Finally, individual classifiers are evaluated using test datasets and the results of different CAD systems are compared to assess the effectiveness of the new feature and the classifiers.

### 2.1.1 Feature Extraction

As mentioned, each case in the dataset has been described using a text file (Table 2) and 1–6 images. The text file contains 12 features and some of these features are described using linguistic variables such as shape: regular or irregular; envelop:



**Fig. 1** The components of the proposed ultrasound CAD system (CRD—Central Regularity Degree (proposed new feature); MFFNN—multi-layer feed forward neural network; KNN—K-nearest neighbor; LDA—Linear Discriminant Analysis; NC—Nearest Centroid)

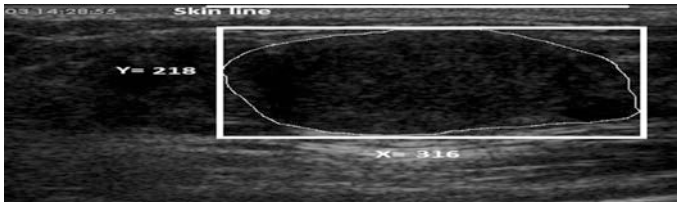
enveloped, not filled or none. To use these features in classifications, we converted the linguistic variables subjectively into numeric values as shown in Table 3.

Also we used the case images to extract two mass geometric features; Width to Depth ratio (WD) and the proposed new feature Central Regularity Degree (CRD). To compute the WD feature we applied the following steps:

- 1 For each case edit the ultrasound image of the mass using an image editor.
- 2 Draw the smallest rectangle that contains the mass as shown in Fig. 2. The X-axis of the rectangle parallels the skin line and represents the width of the mass, whereas the Y-axis represents the depth of the mass.

**Table 3** Description of the mass features and the numeric value assigned to each description

Feature	Description	Coding
Shape	Regular	1
	Irregular	2
Border	Clear	1
	Blur	2
Echo	Equable	1
	Unequable	2
Envelope	None	1
	Not filled	2
	Yes	3
Sides Echo (sides shadow)	None	1
	Yes	2
Micro-calcification shape	None	1
	Needle	2
	Cluster	3
	Large	4
Micro-calcification number	0	0
	1	1
	1-2	1.5
	2-3	2.5
	3-4	3.5
	4-5	4
Reduction (back shadow)	None	1
	Yes	2
Lymph Transformation	None	1
	Yes	2
Blood	Level 1	1
	Level 2 or 3	2
Mass Maximal Diameter	<1	1
	>1 and <2	2
	>2	3



**Fig. 2** The smallest rectangle containing the mass. *Thin line* represents the mass boundary and the *thick line* represents the boundary of the smallest rectangle that contain the mass. The *X* represents the width of the mass along horizontal axis and *Y* represents the depth of the mass along vertical axis

- 3 Compute the length of mass along *X* (*W*) and *Y* axes (*D*) (in pixels).
- 4 Compute  $WD = W/D = X/Y$ .

***A new feature—mass Central Regularity Degree (CRD).***

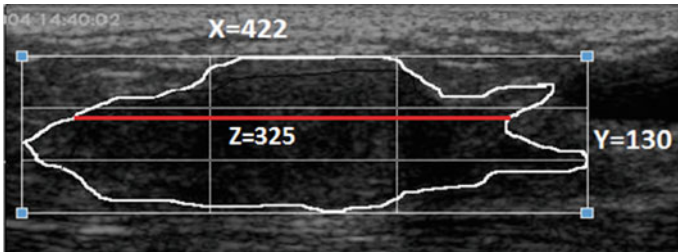
As mentioned previously, mass shape is either regular or irregular. The irregular shape of mass is an indication of malignancy but there are several benign masses that take an irregular shape too. Similarly, regular mass is suggestive of a benign case but there are few malignant masses that take a regular shape. Thus there is a need for further investigation into the mass shape to differentiate irregular benign mass from irregular malignant mass. In an in-depth investigation and closer examination of mass shape to discover potential new features, we were able to define a new geometric feature related to the mass shape called Central Regularity Degree (CRD). The CRD reflects the degree of regularity of the middle part of the mass. This was inspired by the fact that malignant masses typically have irregular shape and CRD is designed to specifically capture this irregularity in the more central part of the mass. As illustrated in Fig. 3, the mass boundary in this image was defined previously by experts as the white outline [18] and CRD involves defining the smallest width of the mass (red line).

To find CRD, we drew on ultrasound images (Fig. 2), a small rectangle that contains the complete mass using image editor software. The rectangle lines *X* and *Y* represent the mass width and the mass depth, respectively. Then we divided the rectangle horizontally into three equal parts; upper, middle and lower (Fig. 3). Next, for the middle part of the mass we found the length (*Z*) of the shortest line that is parallel to the horizontal axis (*X*). Finally, we found the ratio of *Z* to the rectangle width (*X*) (Eq. 1). The output value represents the Central Regularity Degree of the middle part of the mass.

$$CRD = Z \div X \tag{1}$$

After extracting and coding the mass features, there were a total of 14 features (age, 11 features listed in Table 3 and the two geometric features (WD, CRD)) as inputs for the next step. But still there is a need for data normalization to speed up





**Fig. 3** New mass feature Central Regularity Degree (CRD).  $X$  is the rectangle width parallel to skin line,  $Y$  is the rectangle depth and  $Z$  is the shortest line in the middle part of the mass [19]

the classification process and to remove the effect of large spread of some features. For this, Min-Max normalization method was applied (Eq. 2).

$$X' = \frac{X - \min_A}{\max_A - \min_A} (\max_n - \min_n) + \min_n \tag{2}$$

where  $X'$ ,  $A$ ,  $n$  are; normalised values of  $X$ , old range and new range, respectively.

### 2.1.2 Feature Selection

In this step, we used a wrapper method to select a subset of ultrasound features that are strongly related to breast cancer. Now each case in the ultrasound dataset is represented as a vector of case features  $S = \{S_1, \dots, S_n\}$ . The set of features that is strongly related to breast cancer is selected by using Hierarchical clustering (Johnson 1967) [19] and Self-Organizing Map (SOM) as follows:

- 1 Build a state space starting from empty set in the root and add features one by one until the set of all features is reached.
- 2 Use sequential search starting from the root to find the node that separates benign clusters from malignant clusters by applying the following steps:
  - a. Apply Hierarchical clustering.
  - b. Find the best cut off point that differentiates benign clusters from malignant clusters.
  - c. Compute and save the accuracy and the node index.
  - d. Repeat a-c until all nodes in the state space are visited
- 3 Select the node with the highest accuracy.
- 4 Validate the results using Self Organizing Map (SOM) combined with WARD clustering.

### 2.1.3 Classification

To evaluate the effect of the new ultrasound feature CRD on classification accuracy we applied four supervised classifiers: Multilayer Feed Forward Neural Network (MFFNN), Nearest Centroid (NC), K nearest neighbour (KNN) and Linear Discriminant Analysis (LDA) [20]. We divided the dataset randomly into two groups: training and testing, where, 77 cases were used for training and 22 for testing. Each classifier was trained and tested four times using different randomly extracted subsets. Each time the results were stored for the final evaluation.

#### Multilayer Feed Forward Neural Network (MFFNN)

Here, we used a Multilayer Feed Forward Neural Network with an input layer, one hidden layer and one neuron in the output layer. The network was trained using Scaled Conjugate Gradient Back Propagation Algorithm (SCGBP) and Logistic function as the neuron activation function.

Optimizing the number of hidden neurons in the network is still a challenge. Insufficient number of hidden neurons results in two problems: the first is under fitting that results from selecting a small number of hidden neurons giving rise to bias in generalisation. The second problem is over fitting that results from selecting a large number of hidden neurons giving rise to variance in generalisation. To overcome these problems, this study used a Self Organizing Map (SOM) based approach to optimize the number of hidden neurons in the network as described by Samarasinghe [21]. The algorithm starts with training a feed forward neural network with a relatively large number of neurons in the hidden layer. Then it reduces the number of hidden neurons by removing the redundant neurons that form correlated associations with other neurons in terms of the weighted hidden neuron activation that feeds the output neuron.

A summary of the MFFNN and its optimisation is given below. The net or weighted input  $u_i$  to neuron  $i$  and the output  $y_i$  of each neuron in the hidden layer is given by the following equations:

$$u_i = \sum_{m=1}^r x_m w_{im}; \quad y_i = \frac{1}{1 + e^{-u_i}} \quad (3)$$

where  $x$  is an input vector,  $r$  is the number of inputs including bias inputs and  $w_i$  is the weight vector between input vector  $x$  and neuron  $i$ . From Eq. 3, the weighted input to the output neuron  $v$  and the final output of the neural network are:

$$v = b_0 + \sum_{i=1}^n y_i b_i; \quad z = \frac{1}{1 + e^{-v}} \quad (4)$$

where,  $b_0$  is the bias input to the output neuron,  $b$  is the weight vector between hidden neurons and the output neuron.

Form Samarasinghe [21], the weighted activation of redundant hidden neurons across the input vectors correlate. The weighted activation for neuron  $j$  is as shown in Eq. 5 where  $y_j$  is hidden neuron output and  $b_j$  is its weight to the output neuron.

$$N_j = b_j y_j \quad (5)$$

Now, we have a hidden neuron weighted activation matrix  $N$  where each row in the matrix represents one neuron  $j$ . To reduce the complexity of the neural network, we removed the redundant neurons. For this, we applied Self Organising Map (SOM) with correlation distance measure where the activation vector of each neuron was an input vector. The SOM brings similar input vectors closer together by distributing input vectors according to similarity. Similar SOM neurons was clustered using WARD clustering. The number of different clusters on the SOM indicates the optimum number of hidden neurons.

#### K-Nearest Neighbour (KNN)

In KNN, data vectors or the experimental samples are represented as marked points in the space where each point belongs to a known class (benign or malignant, for example). For a new instance, the classifier represents the instance in the same space and calculates the distance between it and the experimental samples. The label of the new instance depends on the labels of the  $K$  closest points to the new instance. The instance is labelled with the class label that has the largest number of points within  $K$  closest points [22]. To select the value of  $K$  we applied the following steps:

- 1 Set  $F = 0$ .
- 2 For  $I = 2$  to  $N$ , where  $N$  is a large number:
  - a. Set the value of  $K = I$
  - b. Classify the samples using KNN.
  - c. Compute the overall **accuracy**.
  - d. If **accuracy**  $\geq F$  then set  $F = \mathbf{accuracy}$  and  $\text{Best} = I$ .
- 3 Set  $K = \text{Best}$

#### Nearest Centroid (NC)

In this classifier, the classification is done by calculating the mean (centroid) of both classes, malignant and benign, from the available data. For the new object  $x$ , the algorithm calculates the distance between the new object and the means of the classes; the object is then labelled with the label of the closest class centroid [23].

#### Linear Discriminant Analysis (LDA)

This classifier uses a covariance matrix to build a hyperplane between the benign and malignant classes by maximizing between to within variance ratio for the classes (Eq. 6) [24] such that

$$P(i|x) > P(j|x) \quad \forall i \neq j \quad (6)$$

The probability that  $x$  belongs to class  $i$  is not easy to compute; therefore, the simplest mathematical formula of LDA is:

$$f_i(x_k) = \mu_i C^{-1} x_k^T - \frac{1}{2} \mu_i C^{-1} \mu_i^T + \ln(p_i) \quad (7)$$

where  $\mu_i$  is the mean vector of class  $i$ ,  $C^{-1}$  is the inverse of covariance matrix of the dataset,  $p_i$  is the probability of class  $i$  and  $T$  is transpose.

The  $x_k$  belongs to class  $i$  if and only if:

$$f_i(x_k) \geq f_j(x_k) \quad \forall i \neq j \quad (8)$$

### 2.1.4 Evaluation

The aim of this step is to evaluate the performance of the two CAD systems (with and without the new feature CRD) using sensitivity, specificity, False Positive (FP) rate, False Negative (FN) rate and overall accuracy [25].

$$\text{Sensitivity} = \frac{\text{Number of malignant samples correctly classified}}{\text{Total number of malignant samples}} \quad (9)$$

$$\text{Specificity} = \frac{\text{Number of benign samples correctly classified}}{\text{Total number of benign samples}} \quad (10)$$

$$\text{Accuracy} = \frac{\text{Number of samples correctly classified}}{\text{Total number of samples}} \quad (11)$$

$$\text{FP} = \frac{\text{Number of benign samples falsely classified}}{\text{Total number of samples in malignant class}} \quad (12)$$

$$\text{FN} = \frac{\text{Number of malignant samples falsely classified}}{\text{Total number of samples in benign class}} \quad (13)$$

## 3 Results and Discussion

### 3.1 Feature Selection

Frequencies of specific ultrasound features in both malignant and benign cases are shown in Table 4. A feature would be considered effective or good if it clearly separated benign from malignant cases; for example, the margin was considered an

**Table 4** Frequency of ultrasound features in the 99 cases (46 malignant (M) and 53 benign (B)) and their effectiveness as a feature that discriminates between malignant and benign cases [19]

Feature	M	B	Effectiveness
Age (mean)	46.4	38	Good
<i>Shape</i>			
Regular	2	29	Good
Irregular	44	24	
<i>Margin</i>			
Clear	8	38	Good
Blur	38	15	
<i>Echo</i>			
Equable	4	19	Not good
Not equable	42	34	
<i>Envelope</i>			
Enveloped	8	9	
Partially	5	9	Not good
No	33	35	
<i>Microcalcification</i>			
Big	2	0	Good
Cluster	2	1	
Needle	21	6	
None	21	46	
<i>Blood level</i>			
Level 1	14	41	Good
>1	32	12	
<i>WD ratio</i>			
> = 1.34	17	36	Good
<1.34	29	17	
<i>CRD</i>			
> = 0.7	10	39	Good
<0.7	36	14	
<i>Diameter</i>			
<= 1	2	8	Not good
>1 and <=2	24	26	
>2	20	19	

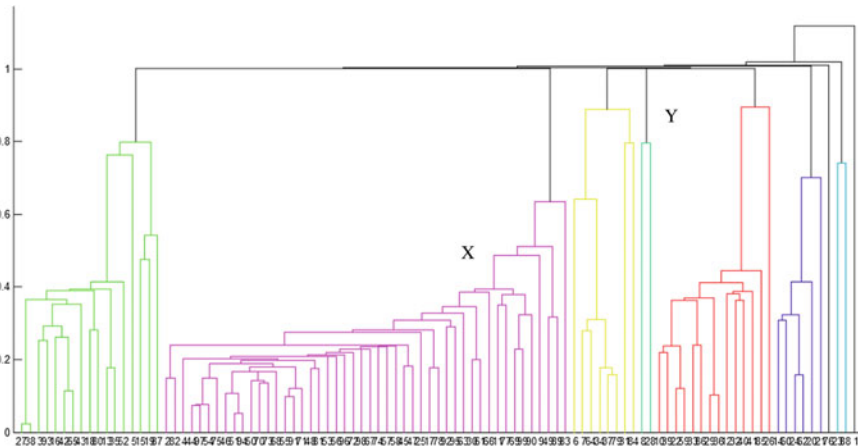
effective feature because most malignant cases (38 out of 46) were blurry and most benign cases (38 out of 53) were clear. On the other hand, the mass echo was not considered effective because most benign (34 out of 53) and malignant (42 out of 46) cases were not equable.

Table 4 shows seven effective features and these are: Age, Shape, Margin, Micro-calcification, Blood level, WD ratio and CRD. In particular, it shows that the new feature CRD is on a par with the most effective feature- margin. According to

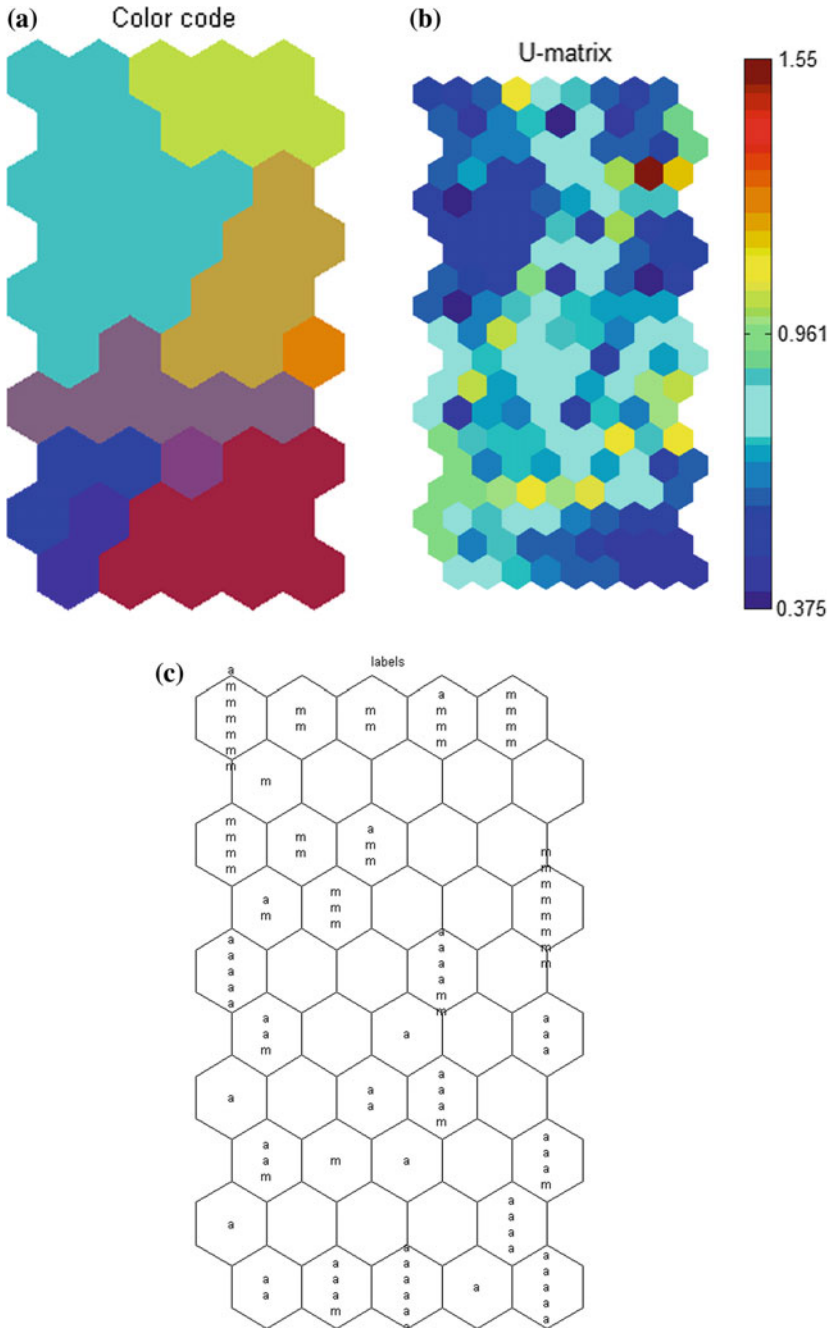
these effective features, a malignant mass has irregular shape, blurry margin, potential micro-calcification, higher speed of blood flow, lower WD ratio (below 1.34) indicative of less elongated shape and smaller CRD (below 0.7) indicative of more irregular shape.

Hierarchical clustering (HC) and self-organizing maps were used for feature selection. We started with hierarchical clustering to find a set of features that separates benign cases and malignant cases into different clusters. Hierarchical Clustering (HC) found age, shape, margin, level of blood flow, DW and our new feature CRD as the features that strongly demarcate breast cancer cases. HC divided the dataset into 9 different clusters. The distribution of malignant samples was: 39 out of 46 cases were distributed over 2 different clusters (clusters X and Y) with 0.84 sensitivity (ratio of malignant cases in the 2 clusters to total malignant cases) (Fig. 4). On the other hand, 42 out of 53 benign cases were distributed over 7 clusters with 0.793 specificity (ratio of benign cases in the 7 clusters to total benign cases) (Fig. 4). Hierarchical clustering produced 81.8 % accuracy.

To confirm the above results, we used SOM to find the distribution of the 99 ultrasound samples over SOM map using the same features (Fig. 5). The dataset was distributed over different regions on the SOM map where most of the malignant cases (41 out of 46) were distributed in the upper part of SOM and most benign cases (37 out of 53) were distributed in the lower part of SOM (Fig. 5c). The SOM U-matrix clearly divided the upper part of SOM into three clusters that appear in the U-matrix as dark blue regions (Fig. 5b). To clarify the boundary of each cluster in SOM map, we used K-mean clustering (k = 9), same number of clusters in hierarchical clustering, to cluster the neurons of SOM (Fig. 5a). By analysing the 9



**Fig. 4** Hierarchical clustering of the 99 cases using the 6 selected features. Each color represents a cluster. Hierarchical clustering has spread the 99 cases over 9 different clusters. (Clusters X and Y are the two clusters containing most of the malignant cases. Other clusters contain benign cases)



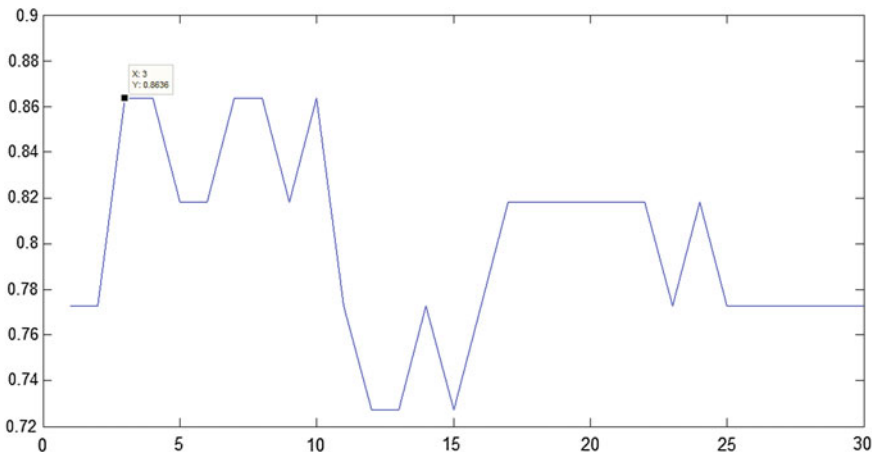
**Fig. 5** SOM clustering of ultrasound data [19]: **a** Clusters (clarified by K-means) (colour coded). **b** SOM U-matrix—the distance of a neuron to the neighboring nodes in the SOM lattice is represented by a color according to the bar that appears on the right side of the figure; the distances range from *dark blue* (small distance) to *dark red* (large distance). Large distances indicate potential cluster boundaries. **c** The distribution of the benign (a) and malignant (m) cases over the SOM lattice which indicates that most malignant cases are in the upper part of SOM whereas most benign cases are in the lower part of SOM

clusters we found 89 % of malignant cases were distributed over 3 clusters (1, 2 and 3) in the upper part of SOM and 70 % of benign cases were distributed over the other 6 clusters. Both Hierarchical clustering and SOM found the above features strongly related to breast cancer.

### 3.2 Classifications and Evaluation

The selected features and the new feature (CRD) were tested for their ability to differentiate between malignant and benign cases. This step was divided into two main stages: In the first stage, we applied the four classifiers; KNN, MFFNN, NC and LDA, on the dataset using all features including CRD. For KNN, firstly, we must determine the value of K, which represents the number of neighbours that controls the class label of the new instance. For this, we started with a large K = 30 and decreased it down to K = 1 (Fig. 6). The best result was obtained when the value of k = 3.

The MFFNN is more complicated than KNN. In the MFFNN, we must take into account the optimal number of neurons in the hidden layer. For this, firstly, we trained and tested an MFFNN using a large number of hidden neurons and reduced the number gradually. Every time, we compared the results with the previous results until the best results were achieved. The best results were obtained using 15 hidden neurons. Secondly, despite the goodness of the results obtained with the 15 hidden neurons, we applied the previously described network pruning method based on



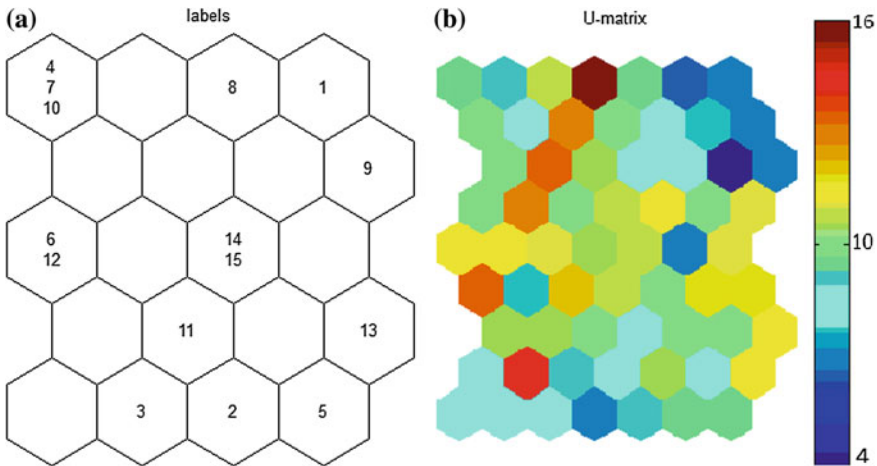
**Fig. 6** The accuracies for different k values in KNN clustering. (X-axis represents the value of K and Y-axis represents the accuracy)



SOM clustering of correlated weighted hidden neuron activation patterns [21] to the MFFNN network just developed. The purpose was to determine whether there is a simpler network than the 15 hidden neuron network that would still give good results. Weighted activation of each hidden neuron for the original inputs constituted an input vector, each representing an individual hidden neuron, into SOM.

To build the SOM, we started with selecting the SOM topology. As the maximum possible clusters is 15 which is the number of hidden neurons, a  $4 \times 5$  rectangular map was built and trained. Then we divided the map neurons into clusters using WARD clustering which resulted in 9 different clusters (Fig. 7). According to Samarasinghe [21], the number of different clusters represents the number of optimum hidden neurons. To verify the performance of the 9 hidden neurons against 15 neurons, we trained and tested a new MFFNN with 9 neurons using the same dataset. The output results obtained from 9 hidden neuron MFFNN were compared with the results from 15 hidden neuron MFFNN and the values of accuracy, sensitivity and specificity were found equal in both networks. The classification accuracy was 95.4 % with 100 % sensitivity and 90.9 % specificity. The results show that the SOM reduced the number of hidden neurons without any effect on the classification performance and reduced the complexity of the neural networks.

The outputs of different classifiers obtained from the first stage are shown in Table 5 (refer to numbers without brackets). The MFFNN was the superior



**Fig. 7** SOM representing 15 hidden neurons. **a** Distribution of 15 neurons over the map; groups of hidden neurons (4, 7, 10), (6, 12) and (14, 15) each shared an SOM neuron (and formed three Ward clusters). **b** U-matrix for the 15 hidden neurons; neurons (9, 1, 8) were found close to each other (the blue color on the top right of the matrix) and considered to be one cluster by WARD clustering that divided the SOM neurons into 9 clusters

**Table 5** Performance of different classifiers (SN is sensitivity, SP specificity and Ac accuracy)

Classifier	Training			Testing		
	SN (%)	SP (%)	Ac (%)	SN (%)	SP (%)	Ac (%)
KNN	85.7 (80)	92.8 (92.8)	89.6 (85.7)	81.9 (90.9)	90.9 (90.0)	86.4 (90.9)
NC	82.8 (82.8)	75.6 (75.6)	80.5 (80.5)	100 (90.9)	63.6 (63.6)	81.8 (77.3)
MFNN	100 (100)	100 (100)	100 (100)	100 (81.8)	90.9 (81.8)	95.4 (81.8)
LDA	82.8 (80)	85.7 (83.3)	84.4 (81.8)	90.9 (90.9)	81.8 (72.7)	86.4 (81.4)

Numbers without brackets are for all features (including CRD) and those within brackets are for all features excluding CRD)

classifier with 100 % sensitivity, 90.9 % specificity, 8.3 % False Positive (FP) rate, 100 % True Negative (TN) rate and 95.4 % classification accuracy. MFNN has 100 % results in the training stage. NC is the worst classifier. KNN and LDA had similar accuracies but LDA was a better discriminator of malignant cases.

In the second stage, we applied the same classifiers (KNN, NC, MFNN and LDA) on the same dataset using all features except CRD. The output results of different classifiers are shown in Table 5 (numbers within brackets). By comparing the results of different classifiers obtained from the first and second stages we found that, the sensitivities of KNN and LDA in the training phase were improved by 5.7 and 2.8 %, respectively, by adding CRD. Also, the specificity of LDA was increased from 83.3 to 85.7 %; furthermore the overall accuracies of LDA and KNN have increased. In the testing phase, the sensitivity of MFNN and NC were improved by 18.2 and 9.1 %, respectively, by adding CRD. Also the specificity of MFNN and LDA were increased to 90.9 and 81.8 % from 81.8 and 72.7 %, respectively. Adding the CRD increased the overall accuracy of the three classifiers, MFNN, NC and LDA.

Only in the testing phase of KNN classifier that we found adding CRD decreased the classification accuracy. This is because of the disadvantages KNN that leads to an increase in the number of misclassified cases in the testing phase. However, the overall accuracy of the classifier (training and testing) is increased by adding the new CRD feature where the overall accuracy of KNN using all features except CRD is  $((22*90.9/100) + (77*85.7))/99$  % = 86.9 %; whereas, the overall accuracy of KNN using all features is 88.8 %.

From the above comparison, we found that adding the new feature CRD enhanced the diagnostic accuracy of the BC-CAD. The best results were obtained by MFNN with 95.4 % accuracy, 90.9 % specificity and 100 % sensitivity in the testing phase. These results were compared with Wei et al. (2007) [11] study that used the same morphological features of masses except our new feature CRD as inputs to the CAD. The comparison found the accuracy of the current system 3.7 % better than their system. Furthermore, a significant improvement of the proposed BC-CAD over the system in [11] was in the 100 % sensitivity achieved by it compared to only 88.89 % achieved by the system in [11] which means that 11.11 % of cancer cases misdiagnosed in [11] could be diagnosed correctly by the current system and hence more lives could be saved.

Also, we compared the results of the current system with Chang et al. (2005) [12] results and we found that the current diagnostic system has 3.7 % higher accuracy and 9 % higher sensitivity, a significant improvement. Furthermore, the results of the current system was slightly better than the results obtained by Chen et al. [13] where the best accuracy for this system was 94.8 % compared with 95.4 % classification accuracy of the current system. But Chen et al. [13] system was more complicated than the current system because all the features used in their system were based on image processing algorithms that are computationally exhaustive and time consuming. Also, it is difficult for a radiologist to extract such features which make these features unusable in a direct diagnostic assessment made by the radiologist. In contrast, all the features used in the current system are extracted from the radiologist report, except DW ratio and our new feature CRD that are easy to extract and be used by a radiologist, without requiring any image processing. By these comparisons we found that the current system has superior diagnostic accuracy to previous ultrasound based BC-CAD systems.

## 4 Conclusions

We have presented in this chapter a new ultrasound based CAD system for early detection of breast cancer, incorporating a novel effective geometric feature—Central Regularity Degree (CRD). From a thorough investigation of the ultrasound images, we extracted this new geometric feature related to the shape of the mass in images, and it was inspired by the fact that most malignant masses are irregular. The CRD reflects the degree of regularity of the middle part of the mass. To demonstrate the effect of CRD in differentiating malignant from benign masses and the resulting improvement in the diagnostic accuracy of breast cancer based on ultrasound images, this study evaluated the diagnostic accuracy of four different classifiers when CRD was added to five known effective mass features: one geometric feature which is Width-Depth ratio (WD); two morphological features: shape and margin; level of blood flow and age.

Multilayer Feed Forward Neural Networks (MFFNN), K Nearest Neighbour (KNN), Nearest Centroid (NC) and Linear Discriminant Analysis (LDA) were used for classification and accuracy, sensitivity and specificity measures were used for evaluation. Ninety-nine breast sonograms—46 of which were malignant and 53 benign were evaluated. The results revealed that CRD is an effective feature discriminating between malignant and benign cases leading to improved accuracy of diagnosis of breast cancer. MFFNN obtained the best results, where the accuracies in the training and testing phases using all features except CRD were 100 and 81.8 %, respectively. On the other hand, adding CRD improves the accuracy of training and testing phases to 100 and 95.5 %, respectively. Therefore, the overall improvement by adding CRD was about 14 %, a significant improvement. Also, the new CRD feature makes the current system better than the previous CADs that used the morphological features of ultrasound images by a considerable margin.

## References

1. R.J. Mason, C.V. Broaddus, T. Martin, T. King, D. Schraufnagel, J.F. Murray, J.A. Nadel, Benign tumors, in *Murray & Nadel's Textbook of Respiratory Medicine* (Saunders Elsevier, Philadelphia, 2005)
2. R.N. Panda, D.B.K. Panigrahi, D.M.R. Patro, Feature extraction for classification of microcalcifications and mass lesions in mammograms. *Int. J. Comput. Sci. Netw. Secur.* **9**, 255–265 (2009)
3. M.B. Popli, Pictorial essay: sonographic differentiation of solid breast lesions, *Indian J. Radiol. Imaging* **12**, 275–279 (2002)
4. G. Rahbar, A.C. Sie, G.C. Hansen, J.S. Prince, M.L. Melany, H.E. Reynolds, V.P. Jackson, J. W. Sayre, L.W. Bassett, Benign versus malignant solid breast masses: US differentiation 1. *Radiology* **213**, 889–894 (1999)
5. H. Madjar, E.B. Mendelson, *The Practice of Breast Ultrasound Techniques: Findings, differential Diagnosis*, 2 edn. (Thieme Verlag, New York, 2008)
6. Y.L. Huang, D.R. Chen, Y.R. Jiang, S.J. Kuo, H.K. Wu, W.K. Moon, Computer-aided diagnosis using morphological features for classifying breast lesions on ultrasound. *Ultrasound Obstet. Gynecol.* **32**, 565–572 (2008)
7. R. Sivaramakrishna, K.A. Powell, M.L. Lieber, W.A. Chilcote, R. Shekhar, Texture analysis of lesions in breast ultrasound images. *Comput. Med. Imaging Graph.: Off. J. Comput. Med. Imaging Soc.* **26**, 303–307 (2002)
8. W.J. Kuo, R.F. Chang, D.R. Chen, C.C. Lee, Data mining with decision trees for diagnosis of breast tumor in medical ultrasonic images. *Breast Cancer Res. Treat.* **66**, 51–57 (2001).
9. F. Zakeri, H. Behnam, N. Ahmadijad, Classification of benign and malignant breast masses based on shape and texture features in sonography images. *J. Med. Syst.* **36**, 1621–1627 (2012). 2012/06/01
10. X. Shi, H.D. Cheng, L. Hu, Mass detection and classification in breast ultrasound images using fuzzy SVM. *Proceedings of the Joint Conf. on Information Sciences, Taiwan* (2006)
11. W.C. Shen, R.F. Chang, W.K. Moon, Y.H. Chou, C.S. Huang, Breast ultrasound computer-aided diagnosis using BI-RADS features. *Acad. Radiol* **14**(8), 928–939 (2007)
12. R.F. Chang, W.J. Wu, W.K. Moon, D.R. Chen, Automatic ultrasound segmentation and morphology based diagnosis of solid breast tumors. *Breast Cancer Res. Treat.* **89**, 179–185 (2005). 2005/01/01
13. C.M. Chen, Y.H. Chou, K.C. Han, G.S. Hung, C.M. Tiu, H.J. Chiou, S.Y. Chiou, Breast lesions on sonograms: computer-aided diagnosis with nearly setting-independent features and artificial neural networks 1. *Radiology* **226**, 504–514 (2003). February 1
14. J. Heinig, R. Witteler, R. Schmitz, L. Kiesel, J. Steinhard, Accuracy of classification of breast ultrasound findings based on criteria used for BI-RADS. *Ultrasound Obstet. Gynecol.* **32**, 573–578 (2008)
15. A.S. Hong, E.L. Rosen, M.S. Soo, J.A. Baker, BI-RADS for sonography: positive and negative predictive values of sonographic features. *Am. J. Roentgenol.* **184**, 1260–1265 (2005). April 1
16. G.N. Lee, D. Fukuoka, Y. Ikedo, T. Hara, H. Fujita, E. Takada, T. Endo, T. Morita, Classification of benign and malignant masses in ultrasound breast image based on geometric and echo features, in *Digital Mammography, Proceedings*, vol. 5116, ed. by E.A. Krupinski (Springer, Berlin, 2008), pp. 433–439
17. J.W. Tian, Y. Wang, J.H. Huang, C.P. Ning, H.M. Wang, Y. Liu, X.L. Tang, The digital database for breast ultrasound image. Presented at the proceedings of the 11th joint conference on information sciences, Paris (2008)
18. S. Johnson, Hierarchical clustering schemes. *Psychometrika* **32**, 241–254 (1967)
19. A. Al-yousef and S. Samarasinghe, Ultrasound based computer aided diagnosis of breast cancer: evaluation of a new feature of mass central regularity degree, in *Proceedings of the*

- 19th International Congress on Modelling and Simulation (MODSIM'11)*. Perth, Australia, 12–16 December 2011 (2011), pp. 1063–1069
20. S. Samarasinghe, Neural networks for water system analysis: from fundamentals to complex pattern recognition. Presented at the hydrocomplexity: new tools for solving wicked water problems, Paris (2010)
  21. X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, D. Steinberg, Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **14**, 1–37 (2007)
  22. G.A. Marcoulides, The elements of statistical learning: data mining, inference and prediction. *Struct. Equ. Model. Multi. J.* **11**, 150–151 (2004)
  23. S. Balakrishnama, A. Ganapathiraju, Linear Discriminant Analysis—A brief Tutorial. Institute for Signal and information processing (1998). [http://www.music.mcgill.ca/~ich/classes/mumt611/classifiers/lda\\_theory.pdf](http://www.music.mcgill.ca/~ich/classes/mumt611/classifiers/lda_theory.pdf)
  24. D.G. Altman, J.M. Bland, Statistics notes: diagnostic tests 1: sensitivity and specificity. *BMJ* **308**, 1552 (1994). 1994-06-11 00:00:00
  25. D.G. Altman, J.M. Bland, Statistics notes: diagnostic tests 1: sensitivity and specificity. *BMJ* **308**(6943), 1552 (1994)