Classification of breast masses in ultrasound images using self-adaptive differential evolution extreme learning machine and rough set feature selection

Kadayanallur Mahadevan Prabusankarlal
Palanisamy Thirumoorthy
Radhakrishnan Manavalan

Classification of breast masses in ultrasound images using self-adaptive differential evolution extreme learning machine and rough set feature selection

Kadayanallur Mahadevan Prabusankar^{a,b,*}, Palanisamy Thirumoorthy^c, and Radhakrishnan Manavalan^d

^aBharathiar University, Research and Development Centre, Department of Electronics and Instrumentation, Coimbatore, India

^bK.S. Rangasamy College of Arts and Science (Autonomous), Department of Electronics and Communication, Tiruchengode, India

^cGovernment Arts College, Department of Electronics and Communication, Dharmapuri, India

^dArignar Anna Government Arts College, Department of Computer Applications and Information Technology, Villupuram, India

Abstract. A method using rough set feature selection and extreme learning machine (ELM) whose learning strategy and hidden node parameters are optimized by self-adaptive differential evolution (SaDE) algorithm for classification of breast masses is investigated. A pathologically proven database of 140 breast ultrasound images, including 80 benign and 60 malignant, is used for this study. A fast nonlocal means algorithm is applied for speckle noise removal, and multiresolution analysis of undecimated discrete wavelet transform is used for accurate segmentation of breast lesions. A total of 34 features, including 29 textural and five morphological, are applied to a k -fold cross-validation scheme, in which more relevant features are selected by quick-reduct algorithm, and the breast masses are discriminated into benign or malignant using SaDE-ELM classifier. The diagnosis accuracy of the system is assessed using parameters, such as accuracy (Ac), sensitivity (Se), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), Matthew's correlation coefficient (MCC), and area (A_z) under receiver operating characteristics curve. The performance of the proposed system is also compared with other classifiers, such as support vector machine and ELM. The results indicated that the proposed SaDE algorithm has superior performance with $Ac = 0.9714$, $Se = 0.9667$, $Sp = 0.975$, $PPV = 0.9666$, $NPV = 0.975$, $MCC = 0.9417$, and $A_z = 0.9604$ compared to other classifiers. © 2017 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: 10.1117/1.JMI.4.2.024507]

Keywords: computer-aided diagnosis system; wavelet transform; extreme learning machine; feature selection; classification; breast ultrasound.

Paper 16239RRR received Nov. 1, 2016; accepted for publication May 25, 2017; published online Jun. 16, 2017.

1 Introduction

Currently, the breast ultrasound is the most important modality, adjunct to mammography for patients with palpable masses. It has the ability to distinguish cystic from solid lesions and unveil hidden lesions in adolescent women with dense breast tissue.¹ Radiologists analyze the breast ultrasound images by observing textural and morphological characteristics of breast lesions. The diagnostic accuracy depends on the expertise and training of the radiologists. Moreover, the ultrasound equipment used to capture images may influence image interpretation and hence the clinical recommendations.² To overcome these difficulties, computer-aided diagnosis (CAD) systems^{3,4} have emerged as a “second reader” for analyzing medical images. Employing computational intelligence in breast ultrasound images increases radiologists' accuracy in the discrimination of malignant masses from benign masses. For more conclusive diagnosis, the radiologists can consider the CAD outcome as a second opinion.⁵

In the past two decades, there have been significant advances in the soft computing-based detection and diagnosis of breast lesions in ultrasound images. Generally, an ultrasound computer-aided system for breast cancer detection and diagnosis consists of five stages: (1) preprocessing stage, (2) segmentation stage, (3) feature extraction stage, (4) feature selection stage, and

(5) classification stage. References 6–8 used a semiautomatic contour procedure for the segmentation of lesions in breast ultrasound images. They used texture features,⁷ morphological features,⁸ and combined textural as well as morphological features⁶ for distinguishing breast tumors by Fisher's linear discriminant analysis (FLDA) classifier. In our previous CAD algorithm,⁹ the nonlocal means (NLM) algorithm is used to remove speckle noise while a hybrid algorithm, which involves particle swarm optimization (PSO) and fuzzy c-means (FCM) clustering, is used for segmentation. The support vector machine (SVM) classifier is used with combined textural and morphological features. In Refs. 10 and 11, ultrasound breast tumors are segmented automatically by the level set method. The autocovariance texture features and morphologic features are extracted from the segmented lesions. The genetic algorithm is used to detect significant features,¹¹ and SVM is used for classification. Based on the definitions in the mass category of breast imaging reporting and data system developed by American College of Radiology, eight features, including shape, orientation, margin, lesion boundary, echo pattern, and posterior acoustic feature classes, are proposed.¹² A binary logistic regression model is used to classify breast masses into benign and malignant. In Ref. 13, two morphological and two texture parameters are used as combined features for FLDA classifier. In Ref. 14,

*Address all correspondence to: Kadayanallur Mahadevan Prabusankar, E-mail: kmsankar@gmail.com

ultrasound breast lesions are segmented from the background by maximizing an average radial gradient index for regions grown from the detected points. The round robin analysis is used to evaluate the quality of the classification of lesions by a Bayesian neural network classifier. The method proposed in Ref. 15 automatically estimated an initial contour based on a manually identified point. A two-stage active contour method refined the initial contour iteratively and performed self-examination and correction on the segmentation result. The linear discriminant analysis (LDA) classification along with step-wise feature selection and twofold cross validation is used to characterize the mass as malignant or benign. However, most of these works have not focused on all the five stages of a CAD system.

The objective of our work is to propose a framework for the classification of breast masses in B-mode ultrasound images. The NLM method is used for despeckling ultrasonograms and it is followed by a robust segmentation method using multi-resolution analysis (MRA) of undecimated discrete wavelet transform (UDWT). The combinations of textural and morphological features are extracted and they are applied to k -fold cross-validation scheme for classification. The more relevant features are selected by quick-reduct (QR) algorithm at each fold, and the SaDE-extreme learning machine (ELM) algorithm classifies the breast ultrasound images into benign or malignant.

2 Methods

2.1 Speckle Removal using Nonlocal Means Algorithm

A major challenge for any ultrasound-based diagnosis system is to quantify the characteristics of breast lesions due to intrinsic limitations, such as speckle noise, low contrast, heterogeneity, or artifacts. The acquired image through the ultrasound equipment is corrupted by speckle noise, which hinders substantial information in the image. The existence of the speckle ruins the image quality and impacts the diagnosis accuracy.¹⁶ Typically, speckle reduction is accomplished by applying various filters.¹⁷ However, these filters also remove finer edge details, which are essential for producing an accurate contour of the tumor for diagnosis.¹⁷ The NLM filter, proposed by Buades et al.,¹⁸ exploits the repetitive patterns of natural or texture images. The NLM filter removes the speckle noise while retaining the image features as much as possible to increase the diagnostic accuracy.¹⁸

The NLM is based on the fact that any pixel of the image has similar pixels that are not necessarily located in a spatial neighborhood. The NLM filter replaces each pixel (or block) with a weighted average of other pixels (or blocks) with similar neighborhoods. The weights in the NLM filter depend on the difference of intensity values and not on the spatial distance between target candidates (pixels or blocks). For a simplified speckle model,¹⁹ $g(n, m) \simeq f(n, m)u(n, m)$, where $g(n, m)$ and $f(n, m)$ are the observed and original versions of a breast ultrasound image, respectively, and the multiplicative component of the speckle noise is $u(n, m)$. The indices n and m denote the axial and lateral indices of the image, respectively.

Consider a noisy grayscale image $v = [v(i)]_{i \in \Omega}$, defined over a domain $\Omega \subset \mathbb{R}^2$, where the $v(i)$ is the noisy observed intensity at position i . The block-wise NLM²⁰ divides the image into overlapping blocks where the restored intensity $NL(v)(B_{ik})$ at position B_{ik} is computed as a weighted average of the block intensities of the image

$$NL(v)(B_{ik}) = \sum_{j \in I} \omega(B_{ik}, B_j) v(B_j), \quad (1)$$

where B_{ik} is the block, centered around pixels ik . For a pixel i included in several blocks B_{ik} , several estimations of the restored intensity $NL(v)(i)$ are obtained in different $NL(v)(B_{ik})$.

The drawback of the NLM algorithm is the consumption of considerable processing time due to its huge computational intensity. Although several improvements have been made to accelerate the processing speed of the algorithm, the calculation of weights is the part which consumes the most computational time. Many preselection methods are introduced to speed up the processing by eliminating dissimilar patches before the weighted averaging process.^{20–22} However, the preselection processes improved the preservation of detail rich regions, while the flat regions are slightly degraded.²³ In fact, the flat regions contain a large number of similar pixels, which tends to improve the denoising performance of NLM. An alternative solution is to use a clustering-based preclassification²⁴ process where the calculation of weights is performed within each cluster instead of for the entire image. The function for calculating weights is given as²⁴

$$\omega_R(B_{ik}, B_j) = \frac{1}{Z_R(B_j)} e^{\frac{-d_R(B_{ik}, B_j)}{h^2}}, \quad (2)$$

where $\omega_R(B_{ik}, B_j)$ depends on the rotationally invariant moment distance measure $d_R(B_{ik}, B_j)$ and $Z_R(B_j) = \sum_{B_j} \omega_R(B_{ik}, B_j)$.

The lack of repetitive patterns in an image leads to unavailability of sufficient candidates for weighted averaging, which degrades the denoising performance. To increase the chance of obtaining more similar patches, the rotationally invariant block matching (RIBM) is utilized.²⁵ The RIBM is a modification of the traditional block matching algorithm to find similar regions in an image. This algorithm is able to detect similar patches even if they appear in several rotated or mirrored instances.²⁵ Moreover, moment invariants-based angle estimation is also performed for RIBM, which leads to better calculation of weights during the NLM process. In comparison with Euclidean-based distance measure, the rotationally invariant moment-based distance measure is more robust against noise.²⁶ It requires the rotation angle between two corresponding blocks B_{ik} and B_j to be estimated.²⁴ The distance measure is the sum of intensity differences between the two corresponding pixels in B_{ik} and B_j .

The centroids are used to estimate the angle of rotation between blocks B_j and B_{ik} and it is defined by assuming that pixels within a block are addressed with a coordinate system C_i that has its origin at the block's center. If $\vec{C}_i$ is the normalized vector corresponding to the centroid of B_{ik} and $m_{B_{ik}B_j}(u)$ is a function that flips the sign of the first component of the vector u , the seventh moment of Hu (ϕ_7) is then used to compute the value of m . The ϕ_7 is invariant under rotation but changes its sign under mirroring. Once the mirroring is compensated using the following function, it always remains stable even if the numerical value suffers from noise

$$m_{B_{ik}B_j}(u) := \begin{cases} (-u_1, u_2)^T & \phi_7(B_j) \cdot \phi_7(B_{ik}) < 0 \\ (u_1, u_2)^T & \text{else} \end{cases}. \quad (3)$$

The final formulation is simplified by considering the gray value of f at the coordinates that are produced by adding the relative

coordinates c_i to the block's center B_{ik} with $f_{B_{ik}}(c_i)$. The new similarity measure²⁴ is derived as

$$d_R(B_{ik}, B_j) = \int_{B_{ik}} [f_{B_{ik}}(c_i) - f_{B_j}(c_j)]^2 dc_i. \quad (4)$$

To transfer the above equation into discrete case, the integral is replaced by a sum and the similarity measure becomes

$$d_R(B_{ik}, B_j) = \sum_{c_i \in B_{ik}} [f_{B_{ik}}(c_i) - I(f_{B_j}, c_j)]^2 dc_i, \quad (5)$$

where I denotes a bilinear interpolation function. For each point of c_i in block B_{ik} , its corresponding point c_j in block B_j is obtained after rotation and interpolation.

2.2 Segmentation using Multiresolution Analysis of Undecimated Discrete Wavelet Transform

The segmentation stage divides an image into two or more non-overlapping regions to separate the required objects from the background.^{17,27} Accurate segmentation is essential for the earliest detection and discrimination of breast tumors by a diagnosis system. The boundaries of the lesion margins contain valuable morphological information regarding the nature of the tumor, which is essential for accurate diagnosis and further characterization. In addition, the segmentation algorithm should encounter various image inherent artifacts, such as shadows.

The wavelet transform is a suitable technique for signal analysis because of its multiscale nature and robustness to noise. Wavelet transform has the property of preserving local texture complexities across scales in an image, which is useful for differentiating between two regions.²⁸ The conventional discrete wavelet transform (DWT) does not produce optimum results in applications, such as filtering and segmentation, due to the loss of the translation-invariance property. This resulted in artifacts of the reconstructed image after the modification of the wavelet coefficients.^{29,30} The UDWT offers huge advantages compared to conventional DWT such as: (1) it has no decimation stage, (2) it does not produce aliasing, and (3) it is shift-invariant.³¹ Also, during MRA, the image energy between different levels does not vary if the original image is shifted prior to decomposition. The MRA makes the UDWT algorithm robust against noise^{29,30} and it becomes more suitable for ultrasound image segmentation.

The multiresolution representation of the input breast ultrasound image $\mathbf{X}$ is obtained by applying UDWT to generate a set of subimages $\mathbf{X}_N = \{\mathbf{X}^0, \mathbf{X}^1, \mathbf{X}^2, \mathbf{X}^3, \dots, \mathbf{X}^r, \dots, \mathbf{X}^{rm}\}$, where the subscript r shows the resolution index and rm shows the maximum possible resolution without losing any meaningful image details.^{29,30} During MRA, each level of UDWT decomposition of an image generates four subbands: $\mathbf{X}_{ll}$, $\mathbf{X}_{lh}$, $\mathbf{X}_{hl}$, and $\mathbf{X}_{hh}$. For level k , the bands are $\mathbf{X}_{k,ll}$, $\mathbf{X}_{k,lh}$, $\mathbf{X}_{k,hl}$, and $\mathbf{X}_{k,hh}$ and for the next level $(k+1)$, the baseband $\mathbf{X}_{k,ll}$ is further decomposed to produce $\mathbf{X}_{k+1,ll}$, $\mathbf{X}_{k+1,lh}$, $\mathbf{X}_{k+1,hl}$, and $\mathbf{X}_{k+1,hh}$. During reconstruction, the inverse UDWT is applied to bands $\mathbf{X}_{k,ll}$, $\mathbf{X}_{k,lh}$, and $\mathbf{X}_{k,hl}$ while band $\mathbf{X}_{k,hh}$ is set to zero. The $\mathbf{X}^0$ corresponds to the input image $\mathbf{X}$ with the lowest resolution index and is strongly affected by noise, but contains maximum details of the image. Significant reduction of noise is observed with reduced image details when the resolution index r increases. The appropriate resolution level r , which contains essential details of the tumor, is automatically selected through mean

structural similarity (MSSIM)³² between input image $\mathbf{X}$ and resolution image $\mathbf{X}^r$.

A feature vector for each pixel (i, j) is created through sampling the dataset $\mathbf{X}_N$ obtained from the MRA process. The four neighborhood pixels of a pixel (i, j) are used for sampling. Each neighborhood $\mathbf{n}_{(i,j)}$ is composed of pixels $\{(i-1, j), (i, j-1), (i, j), (i, j+1), (i+1, j)\}$. The intraresolution feature vector $\mathbf{v}_{i,j}^r$ is computed for each resolution r as follows:²⁹

$$\mathbf{v}_{i,j}^r = \{\mathbf{X}^r[\mathbf{n}_{(i,j)}^1], \dots, \mathbf{X}^r[\mathbf{n}_{(i,j)}^n]\}. \quad (6)$$

The interresolution feature vector $\mathbf{v}_{i,j}$ with a size of $4 \times rm \times 1$ for each pixel is then computed using intraresolution feature vectors ($\mathbf{v}_{i,j}^r$) of each resolution image, starting from $\mathbf{X}^0$ to $\mathbf{X}^{rm}$. The interresolution feature vector $\mathbf{v}_{i,j}$ is computed as

$$\mathbf{v}_{i,j} = \mathbf{v}_{i,j}^0, \mathbf{v}_{i,j}^1, \dots, \mathbf{v}_{i,j}^r, \dots, \mathbf{v}_{i,j}^{rm}. \quad (7)$$

To reduce the computational cost, the dimensionality of each feature vector is reduced using the principle component analysis (PCA).^{29–33} For the sake of convenience, the interresolution feature vector $\mathbf{v}_{i,j}$, the resolution index, and maximum resolution index are denoted as $\mathbf{v}^{(p)}$, p , and $P(1 \leq p \leq P)$, respectively. The dimensionality of each feature vector $\mathbf{v}^{(p)}$ is reduced by creating an eigenvector space using PCA in the following steps:³³ (1) calculate the average value Ψ of $\mathbf{v}^{(p)}$, (2) compute the distance vector $\Delta^{(p)}$ for each feature vector $\mathbf{v}^{(p)}$ as $\Delta^{(p)} = \mathbf{v}^{(p)} - \Psi$, (3) seek a set of D orthonormal vectors $\mathbf{e}_k$ (eigenvectors) along with their scalars λ_k (eigenvalues), (4) calculate the covariance matrix $\mathbf{C}$ as, $\mathbf{C} = \frac{1}{P} \sum_{p=1}^P \Delta^{(p)} \Delta^{(p)T}$, where T is the transpose operator and the matrix $\mathbf{C}$ determines the D eigenvectors as well as eigenvalues, (5) sort the eigenvectors in descending order with respect to their eigenvalues, such that $\lambda_k \geq \lambda_{k+1}$, and (6) finally, convert each feature vector $\mathbf{v}^{(p)}$ of dimension D into $\hat{\mathbf{v}}^{(p)}$ feature vector of dimension K as $\hat{\mathbf{v}}^{(p)} = (\mathbf{v}^{(p)} - \Psi)\mathbf{E}$, where $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_K]$ is an eigenvector matrix and K is automatically estimated using eigenvalues as $K = \arg \min_{k=1,2,\dots,P} \lambda_k / \lambda_1 \leq 0.1$.

The reduced set of feature vectors is segmented into two disjoint classes using clustering. The spatial regularized FCM is used for clustering in which the local spatial information is incorporated into a standard FCM algorithm. The neighborhood effect acts as a regularizer and such regularization is useful in segmenting noise corrupted medical images.^{30,34,35} The algorithm produces two disjoint classes ($c = 2$) by clustering the feature vector data $\hat{\mathbf{v}}^{(p)}$ and assigning each feature vector to the appropriate class using fuzzy memberships.

Let $\hat{\mathbf{v}}_{wc}$ be the cluster mean where N feature vectors are to be partitioned into c clusters and $\hat{\mathbf{v}}^{(p)}$ represents features data. The objective function of spatial regularized FCM is defined as follows:

$$J_m = \sum_{w=1}^c \sum_{k=1}^N \mu_{wk}^m \|\hat{\mathbf{v}}^{(p)} - \hat{\mathbf{v}}_{wc}\|^2 + \frac{\alpha}{N_R} \sum_{w=1}^c \sum_{k=1}^N \mu_{wk}^m \left[\sum_{\hat{\mathbf{v}}^{(p)} \in N_k} \|\hat{\mathbf{v}}_r^{(p)} - \hat{\mathbf{v}}_{wc}\|^2 \right], \quad (8)$$

where μ_{wk} represents the membership of vector $\hat{\mathbf{v}}^{(p)}$ in the w 'th cluster, $\hat{\mathbf{v}}_{wc}$ is the centroid of the w 'th cluster, $\|\cdot\|$ is a norm

metric, and m is a constant that controls the fuzziness of the resulting partition. The membership function μ_{wk} and cluster center $\hat{v}_{wc}$ are updated for each iteration, until the FCM converges to a solution for $\hat{v}_{wc}$ representing the local minimum. The $\hat{v}_r^{(p)}$ is the neighbor of $\hat{v}^{(p)}$, N_k is a set of neighbors within a window around $\hat{v}^{(p)}$, and N_R is the cardinality of N_k . The parameter α controls the effect of the neighbor's term. The significance of a regularizing term using α is inversely proportional to the amount of noise present in the image.³⁵

2.3 Extraction of Textural and Morphological Features

It is essential to explore the feature, or set of features, to obtain better quantifications of the characteristics of breast lesions.³⁶ Identification and selection of significant features lead to accurate diagnosis.³⁷ Textural features are the most efficient in the classification of tumors in breast ultrasound images. However, these features are effective with images obtained from a specific ultrasound system, and diagnostic accuracy reduces with images acquired from different systems and/or with different system settings. On the other hand, the morphological features extracted from the tumor are almost independent of system gain settings or different ultrasound machines.³⁸ Moreover, the borders of benign lesions are smoother than the borders of malignant ones; breast tumors can be assessed based on their morphology.^{9,11} The infiltrative nature of malignant breast tumors produces an irregular pattern of impedance discontinuities, which is perceived as irregular, spiculated, and ill-defined boundaries in ultrasound images. However, benign tumors are characterized by round, smooth, and well-defined boundaries.³⁶ Thus, the morphological information obtained from the boundaries of tumors gives important details for the classifiers. In this work, the combined performance of textural and morphological features for the classification of breast lesions is assessed.

A total of 34 features are extracted using textural and morphological techniques, which includes a seven gray-level run-length matrix (GLRLM),³⁹ a 22 gray-level co-occurrence matrix (GLCM),⁴⁰ and five morphological features.⁹

2.4 Rough Set-Based Feature Selection

Some general guidelines⁴¹ to identify an effective feature that leads to accurate diagnosis are discrimination, reliability, independence, and optimality. A combination of the best performed features does not make the system effective, but the challenge is to reduce the redundancy by the identification and selection of more relevant and effective features.^{27,9} Moreover, the classification performance also depends on the size of the feature vector. An insufficient number of training samples for a limited number of training data results in a "curse of dimensionality" problem, which leads to a degraded classification performance.⁴² An efficient feature selection algorithm is needed in this stage to get rid of redundant and irrelevant features. For effective classification, a new feature set should be framed with a low-dimensional dataset.⁴³ A rough set-based quick-reduct algorithm is proposed in this work to select more prominent features for the classifier.

The rough sets theory is introduced by Pawlak.⁴⁴ An elementary set is the set of indiscernible objects, and the union of such elementary sets is a crisp set. Any set other than elementary and crisp sets is referred to as a rough set. A rough set is the approximation of a vague concept by a pair of precise concepts,

namely, lower and upper approximations. The lower approximation describes domain objects, which certainly belong to the subset. The upper approximation describes the objects that possibly belong to the subset. In relation to a given set of attributes, a set is rough if both the approximations (lower and upper) are not equal.⁴⁵

In a rough set, the data are organized in a decision table where each row and column corresponds to objects and attributes, respectively. A class label in the dataset defines the class to which each row belongs. The class label is known as decision attribute D and the remaining attributes are called condition attributes C .^{46,47} Let U be any finite universe of discourse and R be any equivalence relation defined on U , thus the (U, R) is the collection of all equivalence classes, namely, the approximation space. Let $\{W_1, W_2, W_3, \dots, W_n\}$ be the elements of (U, R) , called the knowledge base. The lower and upper approximations for any subset A of U are defined as

$$\underline{RA} = \cup \{W_i | W_i \subset A\}, \quad (9)$$

$$\overline{RA} = \cup \{W_i | W_i \cap A \neq \Phi\}, \quad (10)$$

where the ordered pair $(\underline{RA}, \overline{RA})$ is the rough set. After these approximations of A are defined, the reference universe U is divided into three regions, namely, positive region $\text{POS}_R(A)$, negative region $\text{NEG}_R(A)$, and boundary region $\text{BND}_R(A)$

$$\text{POS}_R(A) = \underline{RA}, \quad (11)$$

$$\text{NEG}_R(A) = U - \overline{RA}, \quad (12)$$

$$\text{BND}_R(A) = \overline{RA} - \underline{RA}. \quad (13)$$

Therefore, it is insignificant that if $\text{BND}(A) = \phi$, then A is precise. Generally, the attributes are irrelevant to the target and when they are made redundant by other attributes they are considered as unnecessary. In actual practice, both attributes can exist, and the concept of feature selection is used to reduce these attributes to a minimum.⁴³ A decision table has many reducts and any one reduct is sufficient to replace the original table. An optimality measure associated with the attributes is used for the selection of the reduct. A reduct with a minimal number of attributes is selected as the best reduct.⁴⁸

The basic idea of rough sets is the reduction of attributes in the dataset. The reduction is based on the contained information in each attribute such that there is a relation between similar objects and its decision class.⁴³ The attributes that do not have a direct influence on the corresponding decision class are considered to be redundant and they can be eliminated. Rough sets generate decision boundaries using lower and upper approximations of sets to classify objects. Based on several factors, such as the degree of dependency and the discrimination factors, a reduct algorithm, called a QR, is introduced.⁴⁹

2.4.1 Quick-reduct algorithm

The QR reduces the attributes by comparing equivalence relations produced by the sets of attributes. After the reduction, the reduced set offers the same predictive capability as the original set. A reduct is a subset of minimal cardinality $R_{\min}$ of the conditional attribute set C such that $\gamma_R(D) = \gamma_C(D)$. In a given dataset, the R of all reducts is given as^{49,43}

$$R = \{X: X \subseteq C; \gamma_x(D) = \gamma_c(D)\}. \quad (14)$$

The intersection of all sets in R is the core. A subset with minimum cardinality is then searched for a single element of the minimal reduct set $R_{\min} \subseteq R$

$$R_{\min} = \{X: X \in R; \forall Y \in R; |X| \leq |Y|\}. \quad (15)$$

2.5 Classification of Breast Masses using Self-Adaptive Differential Evolution-Extreme Learning Machine

Artificial neural networks and SVM are the methods which are explored the most by researchers⁵⁰ for applications such as classification. The learning speed of feed-forward neural networks is far slower than required because of two important reasons: (1) the extensive use of gradient-based learning algorithms to train neural networks and (2) such learning algorithms are used to tune network parameters iteratively. Huang et al.^{50,51} developed a method called ELM for single-hidden layer feed-forward neural networks (SLFNs). ELM is well known for its fast training speed by means of randomly generating hidden node parameters and calculating the output weights with least square algorithm. At an extremely fast learning speed, the ELM provides the best generalization performance. In ELM, the number of hidden nodes is assigned *a priori*, the parameters are randomly chosen, and they remain unchanged during the learning phase. The random assignment may introduce nonoptimal input weights and hidden biases. These nonoptimal nodes contribute less in minimizing the cost function and compared to conventional tuning-based algorithms, the ELM usually requires more hidden nodes.^{52,53}

Evolutionary algorithms (EAs) are widely used as a global searching method for the optimization of neural network parameters. Among them, differential evolution (DE)⁵⁴ is a simple but powerful population-based stochastic direct searching technique used for network parameter selection.⁵⁵ The DE is directly used as a training algorithm for feed-forward networks⁵⁵ where all the network parameters are encoded into one population vector and the error function between the approximate output and the expected output of the network is used as the fitness function to evaluate all the populations. However, network training by the DE approach alone⁵⁶ yielded slow convergence speed and so, the DE is combined with the Levenberg–Marquardt (LM) method to train feed-forward networks. In DE-LM, DE is used to train the network, choose suitable network parameters, and then the LM method is used to find the optimal parameters. The evolutionary ELM (E-ELM)^{57,58} takes the advantages of ELM and DE to remove redundancy among hidden nodes, resulting in a superior performance in compact networks. It uses the DE algorithm to optimize the network input parameters and the ELM algorithm to calculate the network output weights. It ensures a more compact network size than ELM and exhibits a superior generalization performance than algorithms such as LM, DE-LM, and ELM. However, in these entire DE-based neural network training algorithms, the trial vector generation strategies and the control parameters in DE need to be manually chosen. For example, the control parameters in E-ELM^{57,58} are manually selected based on an empirical suggestion, and the simple random generation method is used to produce the trial vector. Also, the performance of the DE algorithm highly depends on the selected trial vector generation strategy and

control parameters, and inappropriate choices of strategies and control parameters may result in premature convergence or stagnation. Therefore, for different applications, fixing the trial vector generation strategy and the control parameters in the training of SLFNs leads to different network generalization performances.

An algorithm called self-adaptive differential evolution (SaDE) is proposed by Qin and Suganthan.⁵⁹ In the SaDE algorithm, two learning strategies of DE are selected as candidates based on their previous success rates. To generate new potential solutions, these two learning strategies are applied to individuals in the current population. The strategies are to be chosen based on the probability proportional to their previous performances. The critical parameters of the DE algorithm, CR and F , are changed adaptively to handle problems of different classes. One more critical parameter of DE, the population size NP , is manually set to tackle problems with different complexity.⁵⁹ The learning strategies and associated parameters of the SaDE algorithm are adaptive during the evolving procedure for solving many numeric problems.

Cao et al.⁵³ combined SaDE with ELM where the hidden node learning parameters of ELM are optimized by the SaDE algorithm. The SaDE-ELM algorithm produced promising solutions for classification and regression problems by self-adapting the trial vector generation strategies and control parameters in a strategy pool by learning from their previous experiences. The network output weights are calculated using Moore–Penrose (MP) generalized inverse. The combination of SaDE and ELM produced promising results in applications, such as forecasting water inrush in coal mines⁶⁰ and classification of hyperspectral images.⁶¹

2.5.1 Differential evolution algorithm

An EA, called DE, is a powerful tool for global numeric optimization. In DE, for minimizing a target function $\min_{\theta \in R^D} f(\theta)$ with respect to the parameter vector $\theta \in R^D$, a population of NP individual parameter vectors is evolved toward the global optimum.^{53,54} At the generation G , the i 'th candidature parameter vector is as follows:

$$\theta_{i,G} = [\theta_{i,G}^1, \theta_{i,G}^2, \dots, \theta_{i,G}^D] \quad i = 1, 2, \dots, NP. \quad (16)$$

The general procedures of DE are as follows:

Step 1: Initialization

Initialize a set of NP individual parameter vectors $\theta_{i,G}$ for covering the parameter space as much as possible

$$\theta_{i,G} = \theta_{\min} + \text{rand}(0,1) \cdot (\theta_{\max} - \theta_{\min}), \quad (17)$$

where $\theta_{\min} = [\theta_{\min}^1, \theta_{\min}^2, \dots, \theta_{\min}^D]$ and $\theta_{\max} = [\theta_{\max}^1, \theta_{\max}^2, \dots, \theta_{\max}^D]$ are the minimum and maximum parameter bounds, respectively.

Step 2: Mutation

A self-organizing scheme that takes the difference vector of randomly chosen population vectors is applied to convert an existing vector into a new mutant vector. Through a certain mutation strategy,⁵⁴ for each individual parameter vector $\theta_{i,G}$ at current generation, a mutant vector $v_{i,G}$ is generated. Some of the popular mutation strategies (St) are listed as follows:

St₁: DE/rand/1

$$v_{i,G} = \theta_{r_1^i,G} + F \cdot (\theta_{r_2^i,G} - \theta_{r_3^i,G}). \quad (18)$$

St₂: DE/rand-to-best/2

$$v_{i,G} = \theta_{r_1^i,G} + F \cdot (\theta_{\text{best},G} - \theta_{r_1^i,G}) + F \cdot (\theta_{r_2^i,G} - \theta_{r_3^i,G}) \\ + F \cdot (\theta_{r_4^i,G} - \theta_{r_5^i,G}). \quad (19)$$

St₃: DE/rand/2

$$v_{i,G} = \theta_{r_1^i,G} + F \cdot (\theta_{r_2^i,G} - \theta_{r_3^i,G}) + F \cdot (\theta_{r_4^i,G} - \theta_{r_5^i,G}). \quad (20)$$

St₄: DE/current-to-rand/1

$$v_{i,G} = \theta_{i,G} + K \cdot (\theta_{r_1^i,G} - \theta_{i,G}) + F \cdot (\theta_{r_2^i,G} - \theta_{r_3^i,G}). \quad (21)$$

In all these equations, the indices $r_1^i, r_2^i, r_3^i, r_4^i, r_5^i$ are mutually exclusive integers. These indices are different from the index i and are randomly generated within the range $[1, 2, \dots, NP]$. The positive amplification factor F controls the scaling of the difference vectors and is usually selected within the range $0 \leq F \leq 2$. The control parameter K is randomly generated within the region $0 \leq K \leq 1$. Among the different vector generation strategies,^{54,62} the “DE/rand/1” is suitable for solving multimodal problems due to its stronger exploration capability. However, this strategy usually demonstrates slow convergence speed.^{54,62} The “DE/rand-to-best/2” strategy relying on the best solution found so far converges rapidly and performs well when dealing with unimodal problems. However, this strategy is likely to get stuck at a local optimum for multimodal problems and may lead to premature convergence. The two-difference-vectors-based strategies, “DE/rand-to-best/2” and “DE/rand/2,” could lead to a better perturbation than one-difference-vector-based strategies, but they also require a high computational cost. “DE/current-to-rand/1” is a rotation-invariant strategy. It has been shown that this strategy is efficient in solving multiobjective optimization problems.⁶²

Step 3: Crossover

To increase the diversities of the perturbed parameter vectors, a crossover procedure is used. At generation G , with respect to each mutant vector $v_{i,G} = [v_{i,G}^1, v_{i,G}^2, \dots, v_{i,G}^D]$, a trial vector $u_{i,G} = [u_{i,G}^1, u_{i,G}^2, \dots, u_{i,G}^D]$ is generated according to the following crossover equation:

$$u_{i,G}^j = \begin{cases} v_{i,G}^j, & \text{if } (\text{rand}_j \leq CR) \text{ or } (j = j_{\text{rand}}) \\ \theta_{i,G}^j, & \text{Otherwise} \end{cases}, \quad (22)$$

where CR is the crossover rate which controls the fraction of the parameter values obtained from the mutant vector with a positive value between $0 \leq CR < 1$. The rand_j is the j 'th evaluation of a uniform random number

generator with outcome in $[0, 1]$. The j_{rand} is an integer (randomly chosen) from $[1, D]$, which ensures that the minimum one parameter in $u_{i,G}$ is different from the target vector $\theta_{i,G}$.

Step 4: Selection

For each target vector and its corresponding trial vector, a selection step is performed using a fitness function. The one with the lower value of the fitness function is kept as the population of the next generation.

Steps 2 to 4 are repeated until the objective is met or the maximum number of iterations is reached.

2.5.2 Self-adaptive differential evolution algorithm

The strategies used and parameter settings determine the performance of the DE algorithm.⁶³ To achieve superior performance on a given problem, all available learning strategies have to be tried during the mutation phase. Moreover, fine tuning of respective critical control parameters, such as CR , F , and NP , is also essential. Such procedures consume a huge computational time.⁵⁹ A better solution is to use the (SaDE) algorithm that can adapt the learning strategies and the control parameters automatically during the evolution.^{59,62} The objective of the SaDE algorithm is to select a learning strategy among the available strategies probabilistically, for the current population.

In this work, the St₁, St₂, St₃, and St₄ are four candidate learning strategies available to be chosen during the evolution. The probability of applying a strategy from available strategies, St₁, St₂, St₃, and St₄, to each individual in the current population are p_1, p_2, p_3 , and p_4 , respectively. The initial probabilities for all four strategies are made equal and set to be equal 0.25 ($p_1 = p_2 = p_3 = p_4 = 0.25$). For the population of size NP , a vector of size NP can be generated randomly with uniform distribution in the range $[0, 1]$.

If the l 'th element value of the vector is smaller than or equal to p_1 , the strategy St₁ is applied to the l 'th individual in the current population or it moves to the next available strategies. After the evaluation of all trial vectors, the numbers of new trial vectors which enter into the next generation G (generated by different strategies) are recorded as ns_1, ns_2, ns_3 , and ns_4 , respectively. And the numbers of trial vectors discarded are recorded as nf_1, nf_2, nf_3 , and nf_4 . These values are accumulated within a specific number of G that is called a learning period (LP). Then, the probability p of a strategy l is updated as

$$p_l = \frac{St_l}{\sum_{l=1}^4 St_l}, \quad (l = 1, 2, 3, 4). \quad (23)$$

The probabilities of applying the four strategies are updated after LPs, and values acquired in the previous stage by ns_1 to ns_4 and nf_1 to nf_4 are reset. This procedure gradually evolves the most suitable learning strategy for the given problem.

Among the three critical control parameters of DE, the CR , F , and NP , the NP is kept as a user-defined value to tackle problems with different dimensionalities. The F corresponds to the convergence speed, and it is allowed to take any value in the range $(0, 2)$ randomly with normal distributions of mean (0.5) and standard deviation (0.3) for different individuals in the current population. Throughout the evolution process, this scheme could keep a local search with small values of F and

a global search with large values of F for the generation of effective mutant vectors.

The CR is much more sensitive to properties such as the multimodality, and a good choice of CR leads to superior performance under various learning strategies. However, a poor choice of CR leads to a deteriorated performance under any learning strategy. Therefore, the value of CR is dynamically adapted within a certain generation interval by accumulating the previous learning experience. The value of CR is assumed to be normally distributed in a range with mean CR_m and standard deviation 0.1. Initially, the value of CR_m is set at 0.5, and different normally distributed values of CR are generated for each individual in the current population. These values remain unchanged for several generations and then under the same normal distribution, a new set of CR values is generated. The CR values along with trial vectors that are entering into the next generation successfully are recorded at each generation.

The value of CR has been modified a number of times with center CR_m and standard deviation 0.1 after several generations. Using all the recorded CR values, the mean of normal distribution of CR is then recalculated corresponding to successful trial vectors. When the procedure is repeated with this new mean of normal distribution and standard deviation, the range of the proper CR value for the current problem has been automatically learned.

2.5.3 Extreme learning machine algorithm

The ELM algorithm chooses and fixes the weights between input neurons and hidden neurons randomly based on some continuous probability density function. The weights between hidden neurons and output neurons of the SLFN are determined analytically. The ELM algorithm proceeds as follows:^{50,51,64–66}

1. For a given set of training samples (X_i, Y_i) , select the activation function $g(x)$ and hidden neuron number N ,
2. assign input weights w_i and bias values b_i randomly for each neuron,
3. calculate the output matrix H of the hidden layer, and
4. calculate the output weights β_i as

$$\hat{\beta} = H^\dagger T, \quad (24)$$

where $H^\dagger$ is the MP generalized inverse of H . The H , β , and T are defined as

$$H(w_1, \dots, w_{\tilde{N}}, b_1, \dots, b_{\tilde{N}}, x_1, \dots, x_N) = \begin{bmatrix} g(w_1 \cdot x_1 + b_1) & \cdots & g(w_{\tilde{N}} \cdot x_1 + b_{\tilde{N}}) \\ \vdots & \cdots & \vdots \\ g(w_1 \cdot x_N + b_1) & \cdots & g(w_{\tilde{N}} \cdot x_N + b_{\tilde{N}}) \end{bmatrix}_{N \times \tilde{N}}, \quad (25)$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_N^T \end{bmatrix}_{\tilde{N} \times m} \quad \text{and} \quad T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}_{N \times m}. \quad (26)$$

The solution produced [Eq. (24)] accomplishes the minimum square training error and exhibits superior performance on patterns.^{50,51,64,65}

2.5.4 Self-Adaptive Differential Evolution-Extreme Learning Machine algorithm

The SaDE algorithm overcomes the limitations of ELM, such as manual selection of vector generation strategy and fixed values of control parameters for SLFNs. The algorithm optimizes the input weights and hidden node biases and helps the ELM to adaptively derive the output weights.⁵³

Given a set of training data and L hidden nodes with an activation function $g(\cdot)$, the SaDE-ELM algorithm consists of the following steps:

Step 1: Initialization

Initialize a set of NP vectors as the populations of the first generation. Each vector includes all the network hidden node parameters

$$\theta_{k,G} = [a_{1,(k,G)}^T, \dots, a_{L,(k,G)}^T, b_{1,(k,G)}, \dots, b_{L,(k,G)}], \quad (27)$$

where a_j and b_j ($j = 1, \dots, L$) are randomly generated, G is the generation, and $k = 1, 2, \dots, NP$.

Step 2: Output weights and root-mean-square error (RMSE) calculations

Calculate the output weight matrix of network and RMSE according to each population vector using the following equations, respectively:

$$\beta = H_{k,G}^\dagger T, \quad (28)$$

$$\text{RMSE}_{k,G} = \sqrt{\frac{\sum_{i=1}^N \left\| \sum_{j=1}^L \beta_j g[a_{j,(k,G)}, b_{j,(k,G)}, x_i] - t_i \right\|^2}{m \times N}}, \quad (29)$$

where $H_{k,G}$ is shown in the below equation and $H_{k,G}^\dagger$ is the MP generalized inverse of $H_{k,G}$

$$H_{k,G} = \begin{bmatrix} g(a_{1,(k,G)}, b_{1,(k,G)}, x_1) \cdots g(a_{L,(k,G)}, b_{L,(k,G)}, x_1) \\ \vdots \\ g(a_{1,(k,G)}, b_{1,(k,G)}, x_N) \cdots g(a_{L,(k,G)}, b_{L,(k,G)}, x_N) \end{bmatrix}, \quad (30)$$

$$\theta_{k,G+1} = \begin{cases} u_{k,G+1} & \text{if } \text{RMSE}_{\theta_{k,G}} - \text{RMSE}_{u_{k,G+1}} > \epsilon \cdot \text{RMSE}_{\theta_{k,G}} \\ u_{k,G+1} & \text{if } |\text{RMSE}_{\theta_{k,G}} - \text{RMSE}_{u_{k,G+1}}| < \epsilon \cdot \text{RMSE}_{\theta_{k,G}} \\ & \text{and } \|\beta_{u_{k,G+1}}\| < \|\beta_{\theta_{k,G}}\| \\ \theta_{k,G} & \text{else.} \end{cases} \quad (31)$$

The population vector of the first generation with the best RMSE is stored as $\theta_{\text{best},1}$ and $\text{RMSE}_{\theta_{\text{best},1}}$.

Step 3: Mutation and crossover

For each target vector in the current generation, the trial vector generation strategy is chosen from a candidate pool constructed by the four strategies according to probability $p_{l,G}$. The $p_{l,G}$ that represents the probability of the strategy l ($l = 1, 2, 3, 4$) should be chosen at generation G . A fixed number of iterations (G) is defined as LP, and the probability $p_{l,G}$ is updated in any one of the following ways. (1) If $G \leq \text{LP}$, each strategy has an equal probability ($p_{l,G} = 0.25$). (2) If $G > \text{LP}$, then,

$$p_{l,G} = \frac{S_{l,G}}{\sum_{l=1}^4 S_{l,G}} \text{ with}$$

$$S_{l,G} = \frac{\sum_{g=G-\text{LP}}^{G-1} ns_{l,g}}{\sum_{g=G-\text{LP}}^{G-1} ns_{l,g} + \sum_{g=G-\text{LP}}^{G-1} nf_{l,g}} + \varepsilon, \quad (l = 1, 2, 3, 4), \quad (32)$$

where $ns_{l,g}$ is the number of trial vectors generated by the l 'th strategy at g 'th generations that successfully enter into the next generation and $nf_{l,g}$ denotes the number of trial vectors generated by the l 'th strategy at g 'th generations that are discarded in the next generation. The values of $ns_{l,g}$ and $nf_{l,g}$ are stored for LP generations. The earliest records are removed and the new numbers in the current generation are stored once the iterations go beyond the initial LP generations. The ε is a small positive constant, which avoids the possibility of a null success rate.

Moreover, for each target vector, the control parameters F and crossover rate CR are generated in random according to normal distributions $\mathcal{N}(0.5, 0.3)$ and $\mathcal{N}(0.5, 0.1)$, respectively. Appropriate values of CR for a given problem usually fall within a small range. The mean value of CR is adjusted gradually according to past CR values that have generated successful trial vectors for the next generation.

Step 4: Evaluation

All the trial vectors $u_{k,G+1}$ generated at the generation $(G + 1)$ are evaluated using Eq. (31) where ε is the pre-defined small positive tolerance rate. The norm of the output weight $\|\beta\|$ is added as another criterion for the trial vector selection since the neural networks tend to produce better generalization performance with smaller weights. Steps 3 and 4 are repeated until the pre-set goal is reached or the maximum learning iterations are completed. The number of populations NP is highly dependent on the complexity of the given application and it is left as a user-specified parameter as discussed in Sec. 2.5.2.

2.6 Framework of the Proposed Approach

The framework of the proposed approach is shown in Fig. 1. At first, speckle noise in the input breast ultrasound image is removed using the NLM algorithm. Second, the exact lesion is obtained through a segmentation method using MRA of UDWT. Third, the textural and morphological features are computed to

form a 34-dimensional feature vector of the image and it serves as the input for the classifier. Finally, the discrimination of breast masses is performed by the SaDE-ELM classifier with the rough set-QR feature selection method.

3 Results

The algorithm is implemented in MATLAB[®] (Mathworks Inc., Natick, Massachusetts) on a PC with 2-GHz Intel[®] Core[™] i3 CPU (Intel Corp., Santa Clara, California), and 8-GB RAM. The breast ultrasound image database consists of 140 images, including 80 benign and 60 malignant cases. Figure 2 shows the pathologically proven details of the images in the database. The images are obtained from "ultrasoundcases"⁶⁷ in digital imaging and communications in medicine format that complies with the HONcode (health on the net foundation) standard, which provides trustworthy health information. The images are acquired through high end B-mode ultrasound systems which include Aloka Hitachi ProSound F75, HI VISION Ascendus, and Hitachi Noblus (Hitachi Medical Systems Europe, Switzerland) from different patients over a period of seven months from March 2015 to September 2015 with the consent of the patients. All study protocols are approved by the ethics committee of the institution (Gelderse Vallei Hospital, Ede, the Netherlands).

A total of 140 regions of interest (ROIs) of breast ultrasound images are generated through image preprocessing and segmentation algorithm. Figure 3 shows a set of benign and malignant breast ultrasound images: (a) is the original image, (b) and (c) show despeckled and segmented images, respectively. A total of 34 features, including seven GLRLM, 22 GLCM, and five morphological features, are extracted separately from each ROI. A k -fold cross-validation^{9,43} scheme is utilized for classification ($k = 10$) using the SaDE-ELM algorithm. A total of 10 partitions are created by dividing the 140 images, with 14 images contained in each partition. Among the k parts of data, ($k - 1$) parts are used for training and one part of data is used for testing. To ensure all parts of data are tested, the process is repeated 10 times. The processing time consumed by our method is also acceptable. The despeckling and segmentation routines consumed 24 and 14 ms, respectively. The entire algorithm consumed 147, 83, and 117 ms to execute using SVM, ELM, and SaDE-ELM classifiers, respectively.

3.1 Validation Metrics

The effectiveness of the features is evaluated in terms of statistical parameters: accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and Matthew's correlation coefficient (MCC).^{17,9} The MCC is a potential metric for evaluating the accuracy of machine learning methods, specifically when the number of positive and negative samples is unbalanced.¹⁷

3.2 Selection of Parameters for Experiment

At each level of a CAD system, the optimal parameter setting is the most important factor which leads to accurate diagnosis of breast masses. All breast ultrasound images used for the experiments have a size of 256×256 pixels. In the NLM algorithm, the size of the search window is fixed at 12×12 and the block size is set at 7×7 . The filtering parameter h is an important parameter and the optimal value of h corresponds to the quantity

Fig. 1 Different stages of the proposed CAD system.

of noise present in the image. Selecting a low value of h produces a noisy image and a high value of h blurs the fine details of the image. In many methods,⁶⁸ the value of h is chosen as $h = C\sigma$, where C is a constant and σ is the standard deviation of noise. As we confine our experiments to breast ultrasound images, the noise cannot be estimated and so we have chosen $h = 15$ as suggested in Refs. 69 and 25 and based on our previous work with a similar database.^{70,30} The window size of the median filter is set at 5×5 , moment variant φ_7 is used, and the number of clusters in the FCM algorithm is set at 1200.

In the segmentation algorithm, the wavelet Daubechies 2 is used for UDWT construction and reconstruction.²⁹ The appropriate resolution level is automatically selected using MSSIM, calculated between X and X' with a threshold ($t_s = 0.6$). In the spatial regularized FCM clustering, number of clusters ($c = 2$),

constant for fuzziness ($m = 2$), the regularizing term ($\alpha = 0.8$), and neighboring term [$N_R = 9$ (3×3 window)] are set as suggested in Ref. 30. The GLRLM is calculated in four orientations (0 deg, 45 deg, 90 deg, and 135 deg).^{39,9} The GLCM is computed in direction $\theta = 45$ deg, distance $d = 2$, and the quantization level is set at 32.^{43,9} For SVM, the RBF kernel is used, C value is set ($C = 100$) as suggested in Refs. 71 and 72 and the γ value is chosen ($\gamma = 0.2$). For ELM, the number of hidden neurons is set at 20. For DE, the following values are used: the population size $NP = 20$, number of function evaluation $FES_{\max} = 100$, and control parameters $CR = 1$ and $F = 0.9$. Four strategies: DE/rand/1, DE/rand-to-best/2, DE/rand/2, and DE/current-to-rand/1 are used. For SaDE-ELM, the population size $NP = 20$ is used. The parameter settings are held constant throughout the entire experiment.

Fig. 2 Details of the images in the breast ultrasound image database.

Fig. 3 Breast ultrasound images with benign and malignant masses: (a) input image, (b) despeckled image, and (c) segmented image.

3.3 Evaluation of Individual Features

The individual feature vectors generated using texture and morphologic methods are applied as inputs of SVM, ELM, DE-ELM, and SaDE-ELM classifiers along with the k -fold validation scheme. The individual feature vectors are set as separate datasets with normalized values range of $[-1,1]$. The output class of the classifier is corresponded to a low value if the tumor is benign or a high value if it is malignant. The performance of individual features is shown in Table 1. As shown, the accuracy values for each of the features in the discrimination of malignant from benign tumors varied from 0.4643 to 0.9357,

with the highest value being achieved by con. The highest sensitivity, NPV, and MCC are also achieved by con and the values are 0.9333, 0.9494, and 0.8691, respectively. However, the morphological feature P_{ratio} produced the highest specificity and PPV with values of 0.95 and 0.9310, respectively.

The best performing features in the Table 1 are selected manually using accuracy values sorted in descending order. The features obtained through P_{ratio} , normalized residual value (nrv), con, and asm_1 are applied to the classifier as a 4×140 matrix. The classification results are shown in Table 2. As shown, the accuracy, sensitivity, specificity, PPV, NPV, MCC values are 0.95, 0.9333, 0.9625, 0.9492, 0.9506, and 0.8978, respectively.

Table 1 The performance of individual features derived through textural and morphological techniques. The features are sorted in descending order according to the accuracy (Ac) value obtained.

Feature type	Features	Ac	Se	Sp	PPV	NPV	MCC
Morphology (5)	Ratio of perimeters	0.9286	0.9000	0.95	0.9310	0.9268	0.8539
	Normalized residual value	0.9143	0.9000	0.925	0.9	0.925	0.825
	Aspect ratio	0.8428	0.8667	0.825	0.7878	0.8919	0.6857
	Circularity	0.7428	0.7333	0.75	0.6875	0.7895	0.4801
	Overlap ratio	0.65	0.6667	0.6375	0.5797	0.7183	0.3011
GLCM (22)	Contrast	0.9357	0.9333	0.9375	0.9180	0.9494	0.8691
	Homogeneity-I	0.9214	0.9167	0.925	0.9016	0.9367	0.8400
	Correlation-I	0.8857	0.8667	0.9000	0.8666	0.9	0.7666
	Entropy	0.8500	0.8500	0.8500	0.8095	0.8831	0.6963
	Correlation-II	0.8214	0.8000	0.8375	0.7868	0.8481	0.6362
	Information measure of correlation-I	0.7714	0.7333	0.8000	0.7333	0.8	0.5333
	Homogeneity-II	0.7357	0.7000	0.7625	0.6885	0.7721	0.4616
	Information measure of correlation-II	0.7143	0.6833	0.7375	0.6613	0.7564	0.4192
	Inverse difference moment normalized	0.7000	0.6667	0.7250	0.6451	0.7436	0.3902
	Cluster prominence	0.6857	0.6333	0.7250	0.6333	0.725	0.3583
	Cluster shade	0.6500	0.6000	0.6875	0.5902	0.6962	0.2869
	Inverse difference normalized	0.6214	0.5833	0.6500	0.5555	0.6753	0.2321
	Sum of squares: variance	0.6071	0.5667	0.6375	0.5397	0.6623	0.2031
	Autocorrelation	0.5928	0.5500	0.6250	0.5238	0.6493	0.1741
	Dissimilarity	0.5786	0.5500	0.6000	0.5077	0.64	0.1488
	Energy	0.5500	0.5333	0.5625	0.4776	0.6164	0.0949
	Maximum probability	0.5357	0.5167	0.5500	0.4627	0.6027	0.0660
	Sum average	0.5143	0.5000	0.5250	0.4411	0.5833	0.0247
	Sum entropy	0.5000	0.4833	0.5125	0.4264	0.5694	-0.0043
	Sum variance	0.4857	0.4667	0.5000	0.4117	0.5555	-0.0330
	Difference variance	0.4714	0.4667	0.4750	0.4	0.5428	-0.0577
	Difference entropy	0.4643	0.4500	0.4750	0.3913	0.5352	-0.0742
GLRLM (7)	Short run emphasis	0.9071	0.9000	0.9125	0.8852	0.9241	0.8109
	Long run emphasis	0.8214	0.7833	0.85	0.7966	0.8395	0.6347
	Gray-level nonuniformity	0.6571	0.6167	0.6875	0.5967	0.7051	0.303
	Run-length nonuniformity	0.6142	0.5833	0.6375	0.5468	0.6711	0.2194
	Run percentage	0.4714	0.4667	0.475	0.4	0.5428	-0.0577
	Low-gray-level run emphasis	0.9	0.8833	0.9125	0.8833	0.9125	0.7958
	High-gray-level run emphasis	0.5214	0.5000	0.5375	0.4477	0.5890	0.0371

Ac = accuracy, Se = sensitivity, Sp = specificity, GLRLM = gray-level run-length matrix, and GLCM = gray-level co-occurrence matrix.

Table 2 The performance of combined features derived through textural and morphological techniques. Four features (P_{ratio} , nrv, con, and asm1) are manually selected according to the accuracy values obtained in individual assessment.

Manually selected features	Accuracy (Ac)	Sensitivity (Se)	Specificity (Sp)	PPV	NPV	MCC
P_{ratio} , nrv, con, asm ₁	0.95	0.9333	0.9625	0.9492	0.9506	0.8978

Table 3 The performance of selected features by RST-QR feature selection.

Selected features by RST-QR	Accuracy (Ac)	Sensitivity (Se)	Specificity (Sp)	PPV	NPV	MCC
P_{ratio} , con, asm ₁ , sre	0.9714	0.9667	0.975	0.9666	0.975	0.9417

3.4 Evaluation of Selected Features by Rough Set-Quick Reduct Algorithm

The entire feature set (34×140) matrix is applied to the rough set-quick reduct (RST-QR) feature selection algorithm to automatically select the best performing features. The selected features by RST-QR algorithm are given in Table 3. The P_{ratio} from the morphological category, the con and asm₁ from GLCM and sre from GLRLM are the features selected by the algorithm. The classification results in Table 3 also indicate that these feature combinations produced better results compared to individual features. A highest accuracy of 0.9714 is produced by these features. They also produced sensitivity, specificity, PPV, NPV, and MCC values of 0.9667, 0.975, 0.9666, 0.975, and 0.9417, respectively, which suggests that the features selected by RST-QR algorithm produces the best values of classification accuracy. The scatter diagram in Fig. 4 shows the values produced by the set of features manually selected and those features selected by the QR-RST algorithm.

3.5 Evaluation of Classification Performance and Receiver Operating Characteristic Analysis

The overall performance of a diagnosis test can be accessed using a receiver operating characteristic (ROC) curve.⁷³ The area A_Z under the ROC curve (AUC) is a quantitative criterion

Fig. 4 Scatter diagram shows the values produced by the set of features obtained by manual selection versus QR-RST selection.

of the overall performance of a diagnosis system. It is also useful to compare the performances of various classification methods in differentiating positive and negative diagnostic findings. The performance of the SaDE-ELM algorithm is compared with the performances of the ELM and SVM algorithms using k -fold cross-validation scheme ($k = 10$). The entire feature set is applied to the validation scheme by keeping the RST-QR algorithm within the validation loop. This allows the set of optimal features to differ along the folds. The software package OriginPro 8 (OriginLab Corp., Northampton, Massachusetts) is used to compute the AUC values. The ROC curves in Fig. 5 show the AUC value produced by SaDE-ELM ($A_Z = 0.9604 \pm 0.0190$) is superior to the AUC values of SVM and ELM algorithms (0.9462 ± 0.0187 and 0.9497 ± 0.0198 , respectively).

3.5.1 Statistical analysis

To assess the effectiveness of the proposed approach further, the analysis of variance (ANOVA) test on classification accuracy is performed. The AUC value of the proposed approach is compared with other approaches. At a confidence level $\alpha = 0.05$, the results in Table 4 show that the proposed approach significantly outperforms the other two approaches.

Fig. 5 Comparison of ROC curves of the SVM, ELM, and SaDE-ELM classification algorithms.

Table 4 ROC analysis. Comparison of AUC values of the three classifiers. ANOVA test shows the AUC value produced by the proposed method is statistically significant than the AUC values of other two methods with a p -value < 0.05 ($\alpha = 5\%$).

Validation metric	SVM	ELM	SaDE-ELM	p value	Statistical significance
A_z	0.9462 ± 0.0187	0.9497 ± 0.0198	0.9604 ± 0.0190	<0.05	Yes
Accuracy (Ac)	0.9071	0.9429	0.9714	<0.05	Yes
Sensitivity (Se)	0.9	0.9333	0.9667	<0.05	Yes
Specificity (Sp)	0.9125	0.95	0.975	<0.05	Yes
PPV	0.8852	0.9333	0.9666	<0.05	Yes
NPV	0.924	0.95	0.975	<0.05	Yes
MCC	0.8109	0.8833	0.9417	<0.05	Yes

4 Discussion

In this work, a breast ultrasound CAD system using self-adaptive differential E-ELM is proposed. The speckle noise in the input ultrasound image is removed using the NLM algorithm. The accurate region of the tumor is extracted using a segmentation method, which uses MRA of UDWT. Twenty-nine textural features are extracted using GLCM and GLRLM techniques and five morphological features are also extracted. A rough set theory-based feature selection algorithm is used to select more relevant features. The SaDE-ELM algorithm is used for classification of breast masses into benign or malignant.

The best performance is shown by four features P_{ratio} , con, asm_1 and sre with a highest A_z value of 0.9604 using the SaDE-ELM classifier. Moreover, these features produced accuracy, sensitivity, specificity, PPV, and NPV values of 97.14%, 96.67%, 97.5%, 96.66%, and 97.5%, respectively. The morphologic feature P_{ratio} is the ratio of perimeters between a hypothetical circle around the tumor region and the boundary of the tumor.⁹ The characteristics of malignant tumors are associated with irregular margins and the perimeter of such tumors is higher compared to the hypothetical circle drawn around the tumor area.⁷⁴ Basically, the malignant tumors are of three different types: (1) large size with small irregularities and a large perimeter, (2) medium size with small irregularities or few lobulations, and (3) small size with round or oval contour.⁶ The con is a significant texture parameter to distinguish between malignant and benign tumors in breast US images.⁷⁵ It quantifies the difference between the internal tumor region and its surrounding tissue produced by infiltrating malignant tumors, which have ill-defined edges and spicular structures. It is possible to separate the malignant tumors from benign ones using texture features, such as heterogeneous echotexture and posterior acoustic shadowing.⁶

The parameter asm_1 gives additional information about tumor characteristics. It is effective in differentiating benign lesions, such as cysts and fibroadenomas, atypical hyperplasias, and tumor recurrences. It is a measure that provides the degree of homogeneousness of a determined texture. Benign lesions with more organized texture patterns produce higher values of asm_1 than malignant ones.⁷⁶

The sre is based on the fact that the texture in an image can be characterized based on run-lengths of gray levels. Most texture information is concentrated in the short-run region.⁷⁷ The short-run matrix shows a better discrimination ability than long run

matrices. Moreover, a set of four parameters selected by RST-QR (one morphological: P_{ratio} and three texture: con, asm_1 , and sre) presented the best performance.

In the literature, Alvarenga et al.⁷ achieved classification values less than 90% (Ac = 84.2%, Se = 87.0%, Sp = 78.8%, PPV = 88.8%, and NPV = 75.9%) using a set of four texture parameters on a database of 244 breast ultrasound images. They reported slightly higher values using morphological features⁸ (Ac = 83.74%, Se = 83.05%, Sp = 85.51%, PPV = 93.63%, and NPV = 66.29%). Increased values are observed when textural and morphological features are combined in Ref. 6 (Ac = 85.37%, Se = 83.62%, Sp = 89.86%, PPV = 95.48%, and NPV = 68.13%) on the same database. In our previous work,⁹ comparatively higher values are achieved (Ac 95.85%, Se 96%, and Sp 91.46%) on a breast ultrasound database of 120 images using combined textural and morphological features. Wu and Moon¹⁰ reported results higher than 91% (Ac = 92.86%, Se = 94.44%, Sp = 91.67, PPV = 89.47%, and NPV = 95.65%) on a database of 210 images. Shen et al.¹² studied 265 pathology-proven cases; using eight features, they achieved an A_z value of 0.97 (Ac = 91.70%, Se = 90.59%, Sp = 92.22%, PPV = 84.62%, and NPV = 95.40%). Horsch et al.¹³ evaluated two morphological and two texture parameters, achieving an A_z value of 0.87 considering a database of 400 cases. All these results indicate that combining texture and morphological parameters could help to increase clinical diagnostic accuracy and reduce the number of biopsies.

DE is a suitable global searching method for optimizing parameters of SLFNs. In ELM, the number of hidden nodes and other randomly chosen fixed parameters creates many non-optimal nodes, which contribute less in minimizing cost functions. The DE is proven to be useful in selecting the hidden node parameters for ELM, and it makes the hidden nodes more powerful in the cost function minimization process. However, the control parameters and learning strategies of DE are highly dependent on a specific task, and a huge amount of computational time is spent selecting a suitable strategy and fine tuning the corresponding parameters. The SaDE algorithm solves these problems efficiently by automatically choosing a learning strategy and self-adapting critical parameters, CR and F, through previous learning experience.

The classification system plays a major part in the accurate diagnosis of breast masses. The literature points out different results from methods that combine texture and morphological

parameters to classify the breast lesions in ultrasound images. Horsch et al.¹³ used FLDA and achieved an A_z value of 0.87 in the ROC analysis. Drukker et al.¹⁴ used a Bayesian neural network classifier and achieved an A_z of 0.91. Shen et al.¹² utilized a binary logistic regression model to classify breast ultrasound images and obtained an A_z of 0.97. An LDA classifier was used by Cui et al.¹⁵ and achieved an A_z of 0.88. Alvarenga et al.⁶ used FLDA to get an A_z of 0.881 in a breast ultrasound image database. In our previous work,⁹ using an SVM classifier, we achieved an A_z value of 0.9444. The current study reported an A_z value of 0.9462 using an SVM classifier, 0.9497 using an ELM classifier, and the highest AUC value ($A_z = 0.9604$) is obtained using the proposed SaDE-ELM classifier.

5 Conclusion

In this article, the classification performance of self-adaptive differential E-ELM with rough set-based feature selection in breast ultrasound images is evaluated. The performance of individual as well as selected textural and morphological features is analyzed using statistical parameters: accuracy, sensitivity, specificity, MCC, PPV, and NPV. To test the performance of the proposed classifier, ROC analysis is also performed. The proposed classifier outperformed SVM and ELM classifiers by producing highest value of area under ROC curve (A_z).

Disclosures

Dr. Prabusankarlal has nothing to disclose. Dr. Thirumoorthy has nothing to disclose. Dr. Manavalan has nothing to disclose.

Acknowledgments

The authors would like to thank Dr T. S. A. Geertsma, MD, Head, Department of Radiology, Gelderse Vallei Hospital, Ede, the Netherlands, for providing various categories of breast ultrasound images.

References

- W. Gomez et al., "Computerized lesion segmentation of breast ultrasound based on marker controlled watershed transformation," *Med. Phys.* **37**(1), 82–95 (2010).
- M. J. Calas et al., "Intraobserver interpretation of breast ultrasonography following the BI-RADS classification," *Eur. Radiol.* **74**(3), 525–528 (2010).
- K. Drukker et al., "Computerized detection and classification of cancer on breast ultrasound," *Acad. Radiol.* **11**, 526–535 (2004).
- K. Drukker et al., "Breast US computer-aided diagnosis workstation: performance with a large clinical diagnostic population," *Radiology* **248**(2), 392–397 (2008).
- M. L. Giger, "Computer-aided diagnosis of breast lesions in medical images," *Comput. Sci. Eng.* **2**(5), 39–45 (2000).
- A. V. Alvarenga et al., "Assessing the combined performance of texture and morphological parameters in distinguishing breast tumors in ultrasound images," *Med. Phys.* **39**(12), 7350–7358 (2012).
- A. V. Alvarenga et al., "Complexity curve and grey level co-occurrence matrix in the texture evaluation of breast tumor on ultrasound images," *Med. Phys.* **34**, 379–387 (2007).
- A. V. Alvarenga et al., "Assessing the performance of morphological parameters in distinguishing breast tumors on ultrasound images," *Med. Eng. Phys.* **32**, 49–56 (2010).
- K. M. Prabusankarlal, P. Thirumoorthy, and R. Manavalan, "Assessment of combined textural and morphological features for diagnosis of breast masses in ultrasound," *Hum.-Centric Comput. Inf. Sci.* **5**(1), 1–17 (2015).
- W. J. Wu and W. K. Moon, "Ultrasound breast tumor image computer aided diagnosis with texture and morphological features," *Acad. Radiol.* **15**, 873–880 (2008).
- W. J. Wu, S. W. Lin, and W. K. Moon, "Combining support vector machine with genetic algorithm to classify ultrasound breast tumor images," *Comput. Med. Imaging Graphics* **36**(8), 627–633 (2012).
- W.-C. Shen et al., "Breast ultrasound computer-aided diagnosis using BI-RADS features," *Acad. Radiol.* **14**, 928–939 (2007).
- K. Horsch et al., "Computerized diagnostic of breast lesions on ultrasound," *Med. Phys.* **29**, 157–164 (2002).
- K. Drukker et al., "Computerized lesion detection on breast ultrasound," *Med. Phys.* **29**, 1438–1446 (2002).
- J. Cui et al., "A new automated method for the segmentation and characterization of breast masses on ultrasound images," *Med. Phys.* **36**, 1553–1565 (2009).
- I. Njeh et al., "Speckle noise reduction in breast ultrasound images: SMU (SRAD median unsharp) approach," in *8th Int. Multi-Conf. on Systems, Signals and Devices (SSD'11)*, pp. 1–6 (2011).
- H. D. Cheng et al., "Automated breast cancer detection and classification using ultrasound images: a survey," *Pattern Recognit.* **43**(1), 299–317 (2010).
- A. Buades, B. Coll, and J. M. Morel, "A review of image denoising algorithm, with a new one," *Simulation* **4**, 490–530 (2005).
- O. V. Michailovich and A. Tannenbaum, "Despeckling of medical ultrasound images," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **53**(1), 64–78 (2006).
- P. Coupé et al., "An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images," *IEEE Trans. Med. Imaging* **27**(4), 425–441 (2008).
- J. Wang et al., "Fast non-local algorithm for image denoising," in *Proc. IEEE Int. Conf. Image Processing*, pp. 1429–1432 (2006).
- P. Chao et al., "A fast NL-means method in image denoising based on the similarity of spatially sampled pixels," in *Proc. IEEE Int. Workshop on Multimedia Signal Processing*, pp. 1–4 (2009).
- P. Coupé et al., "Nonlocal means-based speckle filtering for ultrasound images," *IEEE Trans. Image Process.* **18**(10), 2221–2229 (2009).
- R. Yan et al., "Improved nonlocal means based on pre-classification and invariant block matching," *IEEE J. Display Technol.* **8**(4), 212–218 (2012).
- S. Grewenig, S. Zimmer, and J. Weickert, "Rotationally invariant similarity measures for nonlocal image denoising," *J. Visual Commun. Image Representation* **22**, 117–130 (2011).
- Q. Li, H. Zhang, and T. Wang, "Scale invariant feature matching using rotation-invariant distance for remote sensing image registration," *Int. J. Pattern Recognit. Artif. Intell.* **27**(2), 1354004 (2013).
- K. M. Prabusankarlal, P. Thirumoorthy, and R. Manavalan, "Computer aided breast cancer diagnosis techniques in ultrasound: a survey," *J. Med. Imaging Health Inf.* **4**(3), 331–349 (2014).
- D. R. Chen et al., "Diagnosis of breast tumors with sonographic texture analysis using wavelet transform and neural networks," *Ultrasound Med. Biol.* **28**(10), 1301–1310 (2002).
- T. Celik and T. Tjahjadi, "A novel method for sidescan sonar image segmentation," *IEEE J. Oceanic Eng.* **36**(2), 186–194 (2011).
- K. M. Prabusankarlal, P. Thirumoorthy, and R. Manavalan, "Segmentation of breast lesions in ultrasound images through multi-resolution analysis using undecimated discrete wavelet transform," *Ultrason. Imaging* **38**(6), 384–402 (2016).
- J. L. Starck, J. Fadili, and F. Murtagh, "The undecimated wavelet decomposition and its reconstruction," *IEEE Trans. Image Process.* **16**(2), 297–309 (2007).
- Z. Wang et al., "Image quality assessment: from error measurement to structural similarity," *IEEE Trans. Image Process.* **13**(1), 600–612 (2004).
- C. Ding and X. He, "K-means clustering via principal component analysis," in *Proc. of the ACM Twenty-First Int. Conf. Machine Learning*, p. 29 (2004).
- W. Cai, S. Chen, and D. Zhang, "Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation," *Pattern Recognit.* **40**(3), 825–838 (2007).
- K. S. Chuang et al., "Fuzzy c-means clustering with spatial information for image segmentation," *Comput. Med. Imaging Graphics* **30**(1), 9–15 (2006).
- W. C. A. Pereira et al., "A non-linear morphometric feature selection approach for breast tumor contour from ultrasonic images," *Comput. Biol. Med.* **40**(11), 912–918 (2010).

37. H. Li et al., "Computerized radiographic mass detection. I. Lesion site selection by morphological enhancement and contextual segmentation," *IEEE Trans. Med. Imaging* **20**(4), 289–301 (2001).
38. Y. L. Huang et al., "Computer-aided diagnosis using morphological features for classifying breast lesions on ultrasound," *Ultrasound Obstet. Gynecol.* **32**(4), 565–572 (2008).
39. K. R. Krishnan and R. Sudhakar, "Automatic classification of liver diseases from ultrasound images using GLRLM texture features," in *Soft Computing Applications*, pp. 611–624, Springer, Berlin, Heidelberg (2013).
40. A. Al-Janobi, "Performance evaluation of cross-diagonal texture matrix method of texture analysis," *Pattern Recognit.* **34**(1), 171–180 (2001).
41. H. Li et al., "Computerized radiographic mass detection. I. Lesion site selection by morphological enhancement and contextual segmentation," *IEEE Trans. Med. Imaging* **20**(4), 289–301 (2001).
42. K. Ganesan et al., "Computer-aided breast cancer detection using mammograms: a review," *IEEE Rev. Biomed. Eng.* **6**, 77–98 (2013).
43. K. Thangavel and R. Manavalan, "Soft computing models based feature selection for TRUS prostate cancer image classification," *Soft Comput.* **18**(6), 1165–1176 (2014).
44. Z. Pawlak, "Rough sets," *Int. J. Comput. Inf. Sci.* **11**(5), 341–356 (1982).
45. G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Machine Learning: Proc. of the Eleventh Int. Conf.*, pp. 121–129 (1994).
46. H. L. Chen et al., "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," *Expert Syst. Appl.* **38**(7), 9014–9022 (2011).
47. R. W. Swinowski and A. Skowron, "Rough set methods in feature selection and recognition," *Pattern Recognit. Lett.* **24**(6), 833–849 (2003).
48. M. Modrzejewski, "Feature selection using rough sets theory," in *Machine Learning: ECML-93*, pp. 213–226, Springer, Berlin, Heidelberg (1993).
49. Y. Chen, D. Miao, and R. Wang, "A rough set approach to feature selection based on ant colony optimization," *Pattern Recognit. Lett.* **31**(3), 226–233 (2010).
50. G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: a new learning scheme of feedforward neural networks," in *Proc. of Int. Joint Conf. on Neural Networks*, Budapest, Hungary, pp. 985–990 (2004).
51. G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing* **70**(1), 489–501 (2006).
52. K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *Q. Appl. Math.* **2**, 164–168 (1944).
53. J. Cao, Z. Lin, and G. B. Huang, "Self-adaptive evolutionary extreme learning machine," *Neural Process. Lett.* **36**(3), 285–305 (2012).
54. R. Storn and K. Price, "Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces," *J. Global Optim.* **11**(4), 341–359 (2004).
55. J. Ilonen, J. I. Kamarainen, and J. Lampinen, "Differential evolution training algorithm for feedforward neural networks," *Neural Process. Lett.* **17**, 93–105 (2003).
56. B. Subudhi and D. Jena, "Differential evolution and Levenberg Marquardt trained neural network scheme for nonlinear system identification," *Neural Process. Lett.* **27**, 285–296 (2008).
57. Q.-Y. Zhu et al., "Evolutionary extreme learning machine," *Pattern Recognit.* **38**(10), 1759–1763 (2005).
58. J. W. Cao, Z. P. Lin, and G.-B. Huang, "Composite function wavelet neural networks with differential evolution and extreme learning machine," *Neural Process. Lett.* **33**(3), 251–265 (2011).
59. A. K. Qin and P. N. Suganthan, "Self-adaptive differential evolution algorithm for numerical optimization," in *IEEE Congress on Evolutionary Computation*, Vol. 2, pp. 1785–1791 (2005).
60. Z. Zhao and M. Hu, "Multi-level forecasting model of coal mine water inrush based on self-adaptive evolutionary extreme learning machine," *Appl. Math. Inf. Sci. Lett.* **2**(3), 103–110 (2014).
61. J. Ku, Z. Cai, and X. Yang, "Self-adaptive differential evolution extreme learning machine for the classification of hyperspectral images," in *Proc. Int. Conf. Mechatronics, Control and Electronic Engineering (MCE'14)*, pp. 709–713, Atlantis Press (2014).
62. A.-K. Qin, V.-L. Huang, and P.-N. Suganthan, "Differential evolution algorithm with strategy adaptation for global numerical optimization," *IEEE Trans. Evol. Comput.* **13**(2), 398–417 (2009).
63. R. Gämperle, S. D. Müller, and P. Koumoutsakos, "A parameter study for differential evolution," *Adv. Intell. Syst. Fuzzy Syst. Evol. Comput.* **10**, 293–298 (2002).
64. N. Y. Liang et al., "Classification of mental tasks from EEG signals using extreme learning machine," *Int. J. Neural Syst.* **16**(1), 29–38 (2006).
65. G.-B. Huang et al., "Extreme learning machine for regression and multi-class classification," *IEEE Trans. Syst. Man Cybern. B* **42**(2), 513–529 (2012).
66. G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing* **70**(1–3), 489–501 (2006).
67. "Ultrasoundcases," <http://ultrasoundcases.info/category.aspx?cat=67> (December 2015).
68. Y. Gu et al., "Ultrasound echocardiography despeckling with non-local means times series filter," *Neurocomputing* **124**, 120–130 (2014).
69. S. Zimmer, S. Didas, and J. Weickert, "A rotationally invariant block matching strategy improving image denoising with non-local means," in *Proc. 2008 Int. Workshop on Local and Non-Local Approximation in Image Processing*, pp. 135–142 (2008).
70. K. M. Prabusankarlal, R. Manavalan, and R. Sivaranjani, "An optimized non local means filter using automated clustering based preclassification through gap statistics for speckle reduction in breast ultrasound images," *Appl. Comput. Inf.* (2017).
71. Y. L. Huang, K. L. Wang, and D. R. Chen, "Diagnosis of breast tumors with ultrasonic texture analysis using support vector machines," *Neural Comput. Appl.* **15**(2), 164–169 (2006).
72. W. J. Wu, S. W. Lin, and W. K. Moon, "Combining support vector machine with genetic algorithm to classify ultrasound breast tumor images," *Comput. Med. Imaging Graphics* **36**(8), 627–633 (2012).
73. E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics* **44**(3), 837–845 (1988).
74. Y. L. Huang et al., "Computer-aided diagnosis using morphological features for classifying breast lesions on ultrasound," *Ultrasound Obstet. Gynecol.* **32**(4), 565–572 (2008).
75. K. G. Kim et al., "Computerized scheme for assessing ultrasonographic features of breast masses," *Acad. Radiol.* **12**, 58–66 (2005).
76. C. M. Azevedo et al., "Computer-assisted analysis of breast tumors texture on sonographic images of patients submitted to breast conserving surgery," *Radiol. Bras.* **42**, 363–369 (2009).
77. X. Tang, "Texture information in run-length matrices," *IEEE Trans. Image Process.* **7**(11), 1602–1609 (1998).

Kadayanallur Mahadevan Prabusankarlal received his MSc degree in applied electronics from PSG College of Arts and Science College, Coimbatore, in 1998. He obtained his ME degree in computer science and engineering from VM University, Salem, in 2007. Currently, he is the head of the Electronics and Communication Department at K.S. Rangasamy College of Arts and Science, Tiruchengode, India. He obtained his PhD in electronics at Bharathiar University, Coimbatore, in 2017. He has produced extensive publications in many SCI-indexed journals. His research areas include medical image processing and pattern recognition.

Palanisamy Thirumoorthy received his MSc degree in applied electronics from PSG College of Arts and Science College, Coimbatore, in 1992. He obtained his ME degree in computer science and engineering from VMU, Salem, in 2007. He received his PhD from Bharathiar University, Coimbatore, in 2010. Currently, he is working as an assistant professor of electronics and communication at Government Arts College, Dharmapuri, India. His research areas include thin films applications, digital image processing, and mobile computing.

Radhakrishnan Manavalan obtained his MSc degree in computer science from St. Joseph's College of Bharathidasan University, Trichy, India, in 1999. He obtained his PhD from the Department of Computer Science, Periyar University in 2013. He is the head of Computer Applications and Information Technology Departments at Arignar Anna Government Arts College, Villupuram. He has published articles in more than 120 international journals and conferences. His areas of interest are medical image processing and analysis, soft computing, mobile computing, pattern recognition, and theory of computation.