RESEARCH ARTICLE

Editorial Process: Submission:12/13/2017 Acceptance:10/08/2018

Automated Detection and Classification of Microcalcification Clusters with Enhanced Preprocessing and Fractal Analysis

V Gowri[1], K R Valluvan[2] ,V Vijaya Chamundeeswari[3]*

Abstract

This paper addresses the automated detection of microcalcification clusters from mammogram images by enhanced preprocessing operations on digital mammograms for automated extraction of breast tissue from background, removing artefacts occurring during image registration using X-rays, followed by fractal analysis of suspicious regions. Identification of breast of either left or right and realigning them to a standard position forms a primitive step in preprocessing of mammograms. As the next step in the process, pectoral muscles are separated. Suspicious regions of microcalcifications are identified and are subjected to further analysis of classifying it as benign or malignant. Texture features are representative of its malignancy and fractal analysis was carried out on extracted suspicious regions for its texture features. Principal Component Analysis was carried out to extract optimal features. Ten features were found to be an optimal number of reduced texture features without compromising on classification accuracy. Scaled conjugate Gradient Back propagation network was used for classification using reduced texture features obtained from PCA analysis. By varying hidden layer neurons, accuracy of results achieved by proposed methods is analysed and is calculated to reach maximum accuracy with an optimal level of 15 neurons. Accuracy of 96.3% was achieved with 10 fractal features as input to neural network and 15 hidden layer neurons in neural network designed. The design of architecture is finalised with maximised accuracy for labelling microcalcification clusters as benign or malignant.

Keywords: Artificial neural network- breast cancer- computer-aided diagnosis- microcalcification

Asian Pac J Cancer Prev, **19 (11)**, 3093-3098

Introduction

Breast cancer, being a highly probable risk of cancer causing deaths among females, early diagnosis of breast cancer is looked upon as a saviour. Early detection of breast cancer from mammogram is an important factor in reducing fatality rate. Mammogram is a widely used and reliable screening technology, helping in detection of breast cancer (Avalos-Rivera and Pastrana-Palma, 2016). The presence of microcalcifications in breast is an early indication of breast cancer (Dhawan and Royer, 1988). Identifying the presence of microcalcifications and localization of microcalcifications in breast tissue is normally carried out by radiologists in screening process. Cheng et al., (2003) attempted to carry out the task of identification of microcalcifications and its localization as a computer aided process.

Suspicious regions encompassing microcalcifications are chosen for analysis or feature extraction in many papers (Miranda and Felipe, 2015; Avalos-Rivera and Pastrana-Palma, 2016; Eltoukhy et al., 2019). Statistical features were extracted for the manually identified region termed as 'region of interest', having lesions or microcalcifications. Features based on intensity, texture and shape, or geometric properties of microcalcifications, were extracted to aid in its classification process by Artificial Neural Networks (Arevalo et al., 2015).

Liu et al., (2001) and Eltonsy et al., (2007) proved that multiresolution of mammograms helps in improving the effectiveness of classification process. Gray level Co-occurrence matrix, Gabor wavelets and Contourlets were implemented for texture feature extraction (Tai et al., 2013; Eichkitz et al., 2015). Rashed et al., (2003) carried out multi-resolution analysis of mammogram using wavelets. Comparison of multi wavelet, wavelet, Harelick and shape features were carried out by Soltanian-Zadeh et al., (2004).

Texture features, through literature works, have been proved to play a significant role in recognizing or labelling of microcalcification clusters as benign or malignant. Principal Component Analysis is employed in this proposed work for the selection of more relevant features to represent the type of microcalcification clusters (Arikidis et al., 2006). Scaled Conjugate gradient method had better training results for real valued cases. It is an extension to enhance performance of back propagation algorithm. This method applied on real world applications show improved results over complex gradient descent

*[1]Department of Information Technology, [3]Department of Computer Science and Engineering, Velammal Engineering College, Chennai, [2]Department of ECE, Velalar College of Engineering and Technology India. *For Correspondence: vijaychamu@gmail.com*

algorithm (Popa, 2015).

The paper is organised as follows

Materials and Methods section describes the database used, algorithm for processing methodology, preprocessing operations involving removal of artefacts, removal of labels, separation of breast tissue, computing and realigning orientation of breast tissue and separation of suspicious regions containing microcalcification clusters. It is followed by description of fractal analysis and extraction of texture features. The application of Principal Component analysis for reduction of features and the training and testing of scaled conjugate gradient back propagation network for classification of microcalcification clusters as benign or malignant are explained. Discussion section involves results obtained by preprocessing the mammogram image, analysis of Lee filter, and effects of varying hidden layer neurons with classification accuracy.

Materials and Methods

Database used

The Mammographic Image Analysis Society (MIAS) maintains mammogram database. The mammogram database consists of 322 images, representing 161 breast pairs in mediolateral oblique view. The images have been reviewed by a consultant radiologist to identify abnormalities and truth data is available with the database. Images are numbered conveniently from 1 to 322. The images are presented as consecutive left-right pairs of each patient, so that odd numbers indicate a left breast image and even numbers a right breast image with right breast following the left breast in order. For illustrations in this paper, 'mdb005' and 'mdb006' are used. 'mdb005' represents left breast image and 'mdb006' represents right breast tissue of the same patient considered.

Results

Algorithm for Processing Methodology

1. Mammogram of patient is considered as input for analysis
2. Preprocessing of breast image
   a. Artefact removal by 'Template Matching Algorithm'
   b. Enhanced Lee filter of window size 5 is applied to suppress noise
   c. Support Vector Machine classifier is applied to get a binary image
   d. Gamma filter is applied to get distinct boundary of breast region
3. Realign orientation of breast tissue
4. Separation of mass or suspicious region
   a. Histogram equalization
   b. Exponential Operation
   c. Filter using Otsu's thresholding and half of standard deviation to get binary image
5. Fractal Texture Analysis of suspicious region to extract [24 x1]fractal texture vector
6. Feature Subset Selection to reduce fractal texture vectors , [24x1] to [10 x1]

Figure 1. Mammogram Image with Patient Id

Figure 2. Mammogram Obtained, Removing Labels

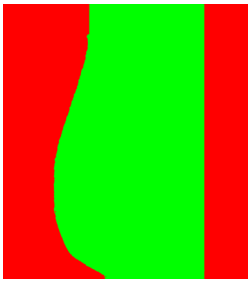

Figure 3. Gamma Filtered SVM Classifier Output

7. Principal Component Analysis is applied to get principal components from fractal texture vector
8. Classifier Network for labeling suspicious region as benign, malignant, normal or no lesion.

Preprocessing of Images to extract Breast Tissue

Mammogram capturing breast image, typically include labels with patient id and descriptions. These labels have either alphabets or numbers enclosed in rectangular boxes. These artefacts were removed by 'template matching algorithm'. Database for template matching consists of alphabets, numerals and a rectangular box. By applying template matching algorithm, the locations of box containing patient id is identified and masked. Figure 1 represents the mammogram image. Figure 2 represents the image with patient id removed.

Enhanced Lee filter is applied to suppress noise while enhancing image detail and sharpness. Multiplicative noise can be removed effectively using this filter using Enhanced Lee filter with a window size of 5. Cross

Table1. Window Size Versus Breast Boundary Detection Accuracy

Window size	Breast Boundary detection accuracy (%)
3x3	84
5x5	98
7x7	88
9x9	76

Figure 4. Boundary Line Separating Breast Tissue and Background Superimposed on Mammogram

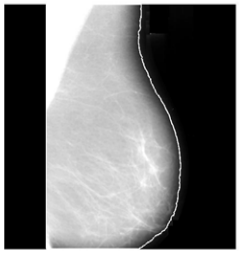

Figure 5. Boundary Line on 'mdb006', Image of Right Breast Tissue

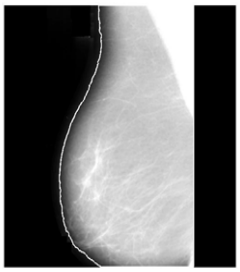

Figure 6. Aligned Breast Tissue

correlation analysis is carried out to enhance breast tissue and background. SVM classifier is applied to give binary output of breast tissue and background. Output of SVM classifier is in binary form and is applied as a mask to the original image to extract breast tissue making an uniform background, so that intensity variations in background will not affect the breast tissue processing for further analysis. Figure 3 represents gamma filtered output of SVM classified output. Gamma filter is applied to further suppress the noise and enhance boundary. Resultant is the boundary line traced and hence separation of breast tissue from the background. Figure 4 represents breast tissue separated with suppressed/ masked background. Figure 5 represents the boundary line detection on right breast.

Computing Orientation of Breast Tissue and Realigning

Once the breast tissue was separated from the background, region properties of breast tissue, viz., Major axis length, Minor Axis length and Orientation, are calculated. Orientation is calculated as the angle

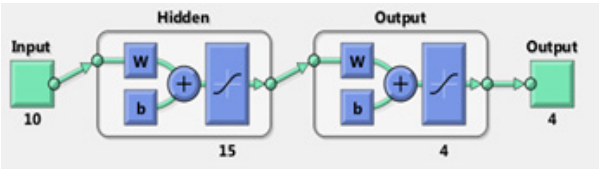

Figure 7. Neural Network for Classifying Microcalcifications Into Four Types of Lesions

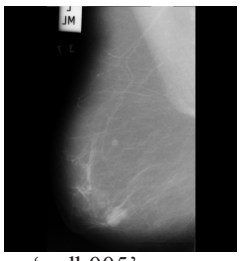

Figure 8. (a) Image 'mdb005'

Figure 8. (b) Template Images (Sample), Representing Alphabets M.L and a Rectangular Box

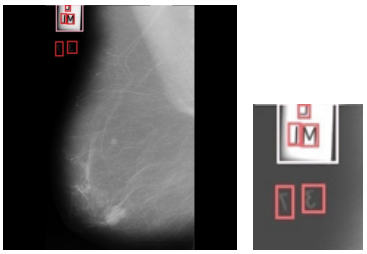

Figure 8. (c) Represents Labels and Artifacts Identified in Input Image 'mdb005'. Artifacts identified area are brightness adjusted to visualize hidden numbers, 3 and 7.

Figure 8. (d) Represents the Masked Image , Devoid of Labels and Artifacts

between horizontal line and major axis. In addition to this, identifying tip of the breast aligning them uniformly in the whole of image database is also required. Breast tip position can be identified by scanning x, and y coordinates of the boundary line separating breast tissue and background. If all images are aligned so that tip of breast occurs in right side of the image, it makes identification easier. On the basis of orientation, image is either flipped or maintained as such for realignment. Study of orientation values on 50 test images belonging to both left and right breast tissue helps in identifying and confirming the role of orientation. Figure 6 represents the flipped breast tissue. With the help of orientation and correcting the alignment of breast tissue with the negation of orientation, helps in realigning the breast tissue.

Separation of Suspicious regions

Histogram Equalization was applied on breast tissue to enhance contrast. Exponential operation on breast tissue further enhances dynamic range. With this process, microcalcification regions are visually enhanced and separable. To make this process automatic, filter using

Table 2. Number of Neurons in Hidden Layer Versus Classification Accuracy

Hidden layer Neurons	Classification Accuracy
6	82.0
8	84.0
10	84.0
12	88.0
14	92.4
15	96.3
16	91.0
18	88.5

Otsu's thresholding and a half of standard deviation was applied to get mass regions emphasized. Mass regions thus obtained are separated by developing a mask applied on breast tissue masking all other regions except mass region. Mass regions or suspicious regions having microcalcification clusters are obtained in this manner automatically.

Fractal Texture Analysis

When the mass regions were identified as mentioned in step I and II, texture features are extracted using fractal analysis. Texture features are intended to capture the granularity and repetitive patterns of regions within an image. Texture feature extraction was carried out using segmentation based Fractal Texture Analysis (SFTA). The input image is decomposed into a set of binary image using two thresholded binary decomposition algorithm. From the binary images, fractal dimensions of the resulting regions were computed that describes the image. In this work, input image of suspicious region was decomposed into set of four images. Each image is described by a [24 x 1] vector of fractal texture features (Cardona et al., 2014).

Principal Component Analysis

Fractal analysis of microcalcification containing region has [24 x 1] feature vector. Before applying a learning algorithm or neural network for training and testing purposes, it is imminent to check for optimization of feature vectors. If there is too much irrelevant and redundant information present or the data is noisy and unreliable, then learning during training phase is more difficult. Feature subset selection is

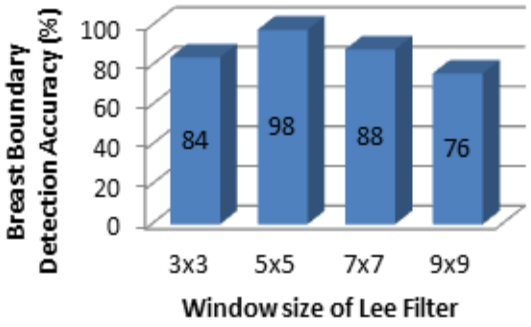

Figure 9. Represents Breast Boundary Detection Accuracy Versus Window Size of Lee Filter

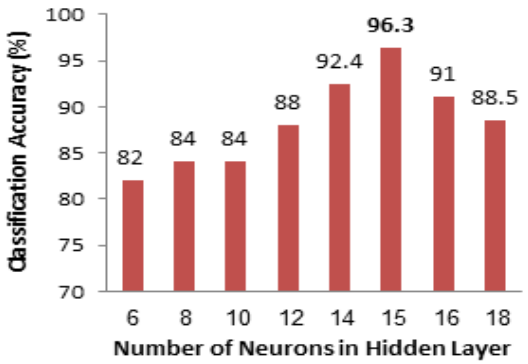

Figure 10. Classification Accuracy for Varying Number of Neurons in Hidden Layer

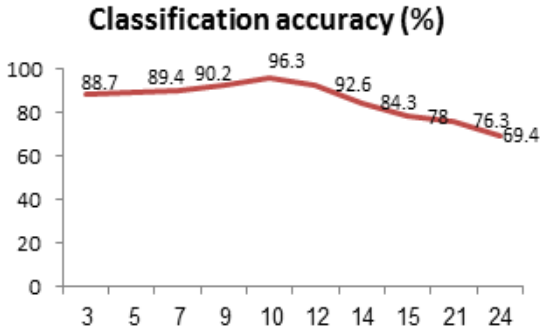

Figure 11. Represents Plot of Classification Accuracy Obtained by Varying the Number of Principal Components

the process of identifying and removing irrelevant and redundant information as much as possible (Hall, 1999). Process of subset selection reduces the dimensionality of the problem, and allows accuracy of classification to be improved. Principal Component Analysis is chosen to compute reduced feature vectors. By varying number of feature vectors, and applying classification process for training, classification accuracy is plotted. From this analysis, optimal number of feature vectors is identified to be ten.

Classifier Network For Labelling Benign And Malignant Microcalcifications

Scaled Conjugate Gradient Back propogation network

Table 3. Number of Features Selected by Principal Component Analysis and Its Classification Accuracy

Principal Components	Classification accuracy (%)
3	88.7
5	89.4
7	90.2
9	92.6
10	96.3
12	92.6
14	84.3
15	78
21	76.3
24	69.4

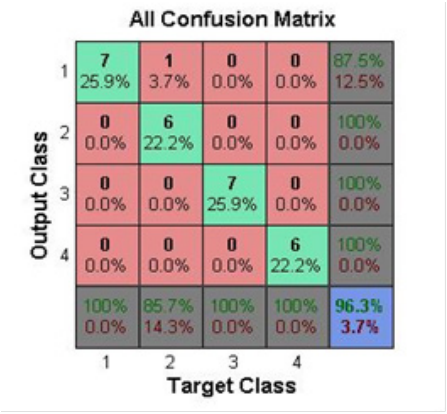

Figure 12. Confusion Matrix of Classifier Output

was used for training and testing purposes. Training was carried out using 50 breast images having 20 malignant, 15 benign and 15 normal tissues. Reduced texture feature vector from fractal analysis and PCA reduction was computed for all fifty sample suspicious regions containing benign, malignant and normal tissues. For the neural network, input layer consisted of 10 neurons and an output layer consisting of four neurons, to represent four types of lesions or microcalcifications. Number of neurons in hidden layer of this network found to have an impact on classifier accuracy. By varying number of hidden layer neurons, classification accuracy was plotted. Then, number of hidden layer neurons was fixed at 15 to give a classifier accuracy of 96.3%. Figure 7 represents the neural network trained for classification of microcalcifications as benign, malignant, normal and no lesions.

Discussion

Preprocessing of Input Image

Test image considered for the analysis is labeled as 'mdb005' in MIAS mammogram database. Generally, mammogram images include labels carrying patient information. It usually includes patient id and description whether image belong to left or right breast, enclosed in a rectangular box. From the analysis of various mammogram images, it is concluded that artifacts generally include a rectangular section, within which alphabets and numerals were printed. In some cases, alphabets and numerals alone are printed even outside the rectangular label section. Artifacts are removed by implementation of 'template matching' algorithm.

Artefacts Removal by Template Matching Algorithm

Template matching is a brute force algorithm to search for a subset image of predefined template within an image. Template matching is a preferred strategy for discovering zones of an image matching a template image. Template matching algorithm uses input image, image from MIAS database as a source image, within which algorithm hopes to identify labels and artifacts, and template images. Templates for all alphabets 'a-z','A-Z', and digits '0-9' and template for rectangular box are employed as template images. Template matching algorithms require template image to be of the same size as in source image. In this paper, Fourier transform of the template is computed so that the matching process is irrespective of the size and orientation of the template in source image. Figure 8 (a) represents the source image, 'mdb005'.

This is the source image given as input to template matching algorithm. Figure 8 (b) represents templates, sample template representing L, M and rectangular box. These are given as template images. ENVI +IDL 4.7 is used to employ template matching algorithm. Loci of template matching areas were located. Figure 8 (c) represents the template locations identified in the source image, 'mdb005'. Then the corresponding regions were masked. The tools, 'Build mask' and 'Apply mask' from ENVI were utilized building the mask and obtaining the masked image. Figure 8 (d) represents the masked image, void of artifacts.

Analysis of Lee filter

The image obtained from mammogram includes speculative noise, specifically, multiplicative noise (Makandar and Halalli, 2015). Enhanced Lee filter is an adaptive filter. Adaptive filtering uses the standard deviation of those pixels within a local box surrounding each pixel to calculate a new pixel value (Vikhe and Thool, 2016). Lee filter is a standard deviation based sigma filter that filters data based on statistics calculated within individual filter windows. Enhanced Lee filter are the adaptation of Lee filter and similarly uses local statistics, specifically, coefficient of variation, with individual filter windows. As a result, each pixel is put into three classes, which are treated as (a) Homogeneous: the pixel value is replaced by average of filter window, (b) Heterogeneous: pixel value is replaced by a weighted average, and (c) Point target: pixel value is not changed.

Unlike a typical low pass filter, Enhanced Lee filter preserve image sharpness and detail while suppressing noise. The masked image is filtered for its multiplicative noise preserving image sharpness and texture using adaptive Lee filter. Window size of Lee filter is varied from 3x3 to 5x5, 7x7, and 9x9 and accuracy of detecting breast tissue from background is calculated. Application of Lee filter, for suppressing noise, followed by SVM classifier, produces a binary output, representing breast tissue and boundary.

With the analysis of varying window size of Lee filter, window size of 5x5 is taken as optimal for suppressing noise in detecting breast boundary algorithm (Figure 9).

In the design of classifier for the neural network, input layer consists of 10 neurons and an output layer consisting of four neurons, to represent four types of lesions or microcalcifications. Number of neurons in hidden layer of this network is found to have an impact on classifier accuracy. By varying number of hidden layer neurons, classification accuracy is plotted (Figure 10). From Table 2, number of hidden layer neurons was fixed at 15 to give an optimal classifier accuracy of 96.3%.

From the breast tissue separated from background, suspicious regions were localized. Texture features are clearly representative of benign or malignancy of the breast tissue, fractal analysis was applied. Fractal analysis resulted in twenty four features. These features,

if used, as such, will have redundant information and may reduce the accuracy of results. Table 3 and Figure 11 represents the classification accuracy obtained by varying number of principal components from PCA application. From the graph depicted in Figure 11, it is clearly obvious that the optimal number of principal components was ten. If lesser number of features is used, then classifier was not able to capture significant information for malignancy detection, resulting in lesser accuracy.

Scaled conjugate Gradient Back propogation network was used for classification using reduced texture features obtained from PCA analysis. By varying hidden layer neurons, accuracy of results achieved by proposed methods is analysed to reach an optimal level of 15 neurons. Figure 12 represents the confusion matrix obtained as performance of classifier. Class 1 represents malignant tissue, class 2 represents benign tissue, class 3 represents normal tissue and class 4 represents no lesion. The classifier has resulted in 96.3% accuracy with testing carried out for 27 sets of regions of breast tissue.

In conclusion, research works carried out in the analysis of mammograms have applied intensity, texture and morphological operators to classify microcalcifications. In this paper, we have adopted the texture features using fractal analysis. Even though twenty four features obtained from fractal analysis helped in effective classification of synthetic images, for real world application of classifying microcalcifications, the proposed method resulted in 69.4%. To remove redundancy in data and making features lesser and representative of the type of MC, plot of classification accuracy versus principal components was drawn and obtained the optimal number of principal components selected as features was found to be 10. Then, classifier was trained and tested to achieve an accuracy of 96.3%. Automated analysis was achieved by identifying suspicious regions with the help of proposed method of tracing boundary and extraction of suspicious regions or regions containing microcalcifications. Future scope of the work is to identify the type of breast tissue as fat or dense and check classification accuracy of labelling lesions presence in both types of breast tissue. Tailor made architectures can be designed to achieve higher accuracy for detection of types of calcifications in all kinds of breast tissue.

References

Arevalo J, Gonzalez FA, Ramos-Pollan R, Oliverira JL, Lopez MAJ (2015). Convolutional neural networks for mammography mass lesion classification. Engineering in Medicine and Biology Society (EMBC), 37th Annual International Conference of the IEEE, pp 797-800.

Arikidis N, Skiadopoulos S, Sakellaropoulos F, Panayiotakis G, Costaridou L (2006). Microcalcification features extracted from principal component analysis in the Wavelet domain. Artificial Intelligence Applications and Innovations, **204**, pp 730-6.

Avalos-Rivera ED, Pastrana-Palma ADJ (2016). Classifying microcalcifications on digital mammography using morphological descriptors and artificial neural network. *Adv Sci Technol Eng Syst J*, **2**, 233-40.

Cardona HDV, Orozco A, Alvare MA (2014). Automatic recognition of microcalcifications in mammography images through fractal texture analysis. *Adv Vis Comput*, **8888**, 841-50.

Cheng HD, Cai X, Chen X, Lu H, Lou X (2003). Computer-aided detection and classification of microcalcifications in mammograms: a survey. *Pattern Recognit*, **36**, 2967-91.

Dhawan AP, Royer EL (1988). Mammographic feature enhancement by computerized image processing. *Comput Methods Programs Biomed*, **27**, 23-35.

Eichkitz CJ, Davies J, Amtmann J, Schreilechner MG, Groot PD (2015). Gray level co-occurrence matrix and its application to seismic data. *First Break*, **33**, 71-7.

Eltonsy NH, Tourassi GD, Elmaghraby AS (2007). A Concentric morphology model for the detection of masses in mammography. *IEEE Trans Med Imag*, **26**, 880-9.

Eltoukhy MM, Faye I, Samir BB (2010). Breast cancer diagnosis in digital mammogram using multiscale curvelet transform. *Comput Med Imaging Graph*, **34**, 269-76.

Hall MA(1999). Correlation-based feature selection for machine learning. Ph.D thesis, The University of Waikato, Hamilton, New Zealand, pp 1-172.

Liu S, Babbs CF, Delp EJ (2001). Multiresolution detection of spiculated lesions in digital mammograms. *IEEE Trans Image Process*, **10**, 874-84.

Makandar A, Halalli B (2015). Breast cancer image enhancement using median filter and CLAHE. *Int J Sci Eng Res*, **6**, 462-5.

Miranda GHB, Felipe JC(2015). Computer-aided diagnosis system based on fuzzy logic for breast cancer categorization. *Comput Biol Med*, **64**, 334-46.

Popa CA (2015). Scaled conjugate gradient learning for quaternion – valued neural networks. *Adv Intell Sys Comput (AISC)*, **378**, 221-33.

Rashed EA, Ismail IA, Zaki SI (2003). Multiresolution mammogram classification using a wavelet transform decomposition. *Pattern Recognit Lett*, **24**, 973-82.

Soltanian-Zadeh H, Rafiee-Rad F, Pourabdollah-Nejad SD (2004). Comparison of multiwavelet, wavelet, haralick and shape features for microcalcification classification in mammograms. *Pattern Recognit*, **37**, 1973-86.

Tai SC, Chen ZS, Tsai WT, Lin CP, Cheng LL (2013). A mass detection system in mammograms using grey level co-occurrence matrix and optical density features. *Adv Intell Syst Appl*, **2**, 369-76.

Vikhe PS, Thool VR (2016). Mass detection in mammographic images using wavelet processing and adaptive threshold technique. *J Med Syst*, **40**, 1-16.