2018 17th IEEE International Conference on Machine Learning and Applications

ROI Detection in Mammogram Images using Wavelet-Based Haralick and HOG Features

Sena Busra Yengec Tasdemir[1], Kasim Tasdemir[2], Zafer Aydin[3]
Computer Science Depertmant
Abdullah Gul University
Kayseri, Turkey
email: [1] sena.yengec@agu.edu.tr, [2] kasim.tasdemir@agu.edu.tr, [3] zafer.aydin@agu.edu.tr

***Abstract*—Digital mammography is a widespread medical imaging technique that is used for early detection and diagnosis of breast cancer. Detecting the region of interest (ROI) helps to locate the abnormal areas, which may be analyzed further by a radiologist or a CAD system. In this paper, a new classification method is proposed for ROI detection in mammography images. Features are extracted using Wavelet transform, Haralick and HOG descriptors. To reduce the number of dimensions and eliminate irrelevant features, a wrapper-based feature selection method is implemented. Several feature extraction methods and machine learning classifiers are compared by performing a leave-one-image-out cross-validation experiment on a difficult dataset. The proposed feature extraction method provides the best accuracy of 87.5% and the second-best area under curve (AUC) score of 84% when employed in a random forest classifier.**

Keywords—ROI detection, Haralick Features, Wavelet Decomposition, Random Forest Classifier.

I. Introduction

Breast cancer is the most common cancer type among women. Early detection of breast cancer is important for developing effective treatment strategies and reducing the mortality rate. Mammography is the most reliable method for detecting breast cancer. However, the signs of the cancer are very subtle at the early stages. Therefore, expert radiologists can misdiagnose an important proportion of the cases. To reduce this error rate, computer aided detection (CAD) systems have been developed. A CAD system typically consists of three main stages: pre-processing, feature extraction and classification. Choosing the right method for each of these steps are important for the accuracy of a CAD system. Furthermore, to facilitate the decision process, detecting regions of interest (ROI) automatically can help to localize the cancer tissue better instead of extracting features directly from the original high-resolution mammography image.

Several feature extraction methods have been proposed in the literature for analyzing the mammography images [7,8]. Among those texture-based approaches such as LBP [1] [2], GLCM [3] [4], HOG [5], Wavelet+GLCM [6], Haralick [7] [8], Wavelet+HOG [9] have been applied to mammogram classification. In this paper, a new feature extraction method is proposed for ROI detection in mammography images, which first computes the one level two-dimensional discrete Wavelet transform (2D-DWT) and then derives Haralick and HOG features from the Wavelet representation. In the next step, the most important features are selected using a wrapper-based feature selection method. Finally, a random forest classifier decides whether the input image is a ROI or not.

II. Materials and Methods

A. Dataset

Mammographic images were obtained from the pilot dataset of the Digital Mammography Dream Challenge, which was held on 2017 across the globe [10]. Detailed description on mammography images used during this challenge can be found in [11]. The pilot set consists of 500 images, 34 of which are labeled as cancer positive and 466 as cancer negative. This data set can be downloaded from [12] by registering with the Synapse platform. In this study, 31 ROIs from the cancer positive images of the pilot set are marked by Assoc. Prof. Fahrettin Kılıç, who is an expert on radiology (Fig. 2). This is followed by a manual cropping and labeling process, which selects a rectangular region around each ROI and labels it as positive. A similar cropping has been applied to extract 31 randomly selected rectangular regions from the cancer negative images as well as from non-ROI sections of cancer positive images. These are labeled as negative. As a result, a ROI dataset is constructed that contains 31 positive and 31 negative images (Fig. 1). Each ROIs size is approximately 73x68.

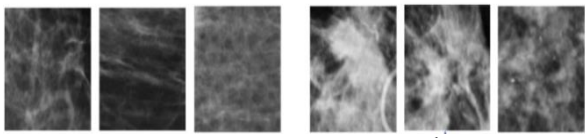


Figure 1. Mammogram ROI dataset: (a) Three samples from non-ROI images labeled as negative (b) Three samples from ROI images labeled as positive

B. ROI Detection

The proposed method includes two major tasks: feature extraction and classification. Feature extraction is applied on small-sized rectangular images derived from the original mammogram images. In the next step, a classifier decides whether the feature set that represents the selected region belongs to a ROI (positive) or not (negative). A block diagram summarizing the steps of the proposed method is presented in Fig. 3.

C. Feature Extraction

Feature extraction aims to capture the visual information content of an image by mapping and reducing the pixel data to another domain in order to facilitate the decision-making process. In this paper, three feature extraction methods are employed: 2D-DWT, Histogram of Oriented Gradients (HOG) and Haralick descriptors.

1) Multi-resolution Analysis using 2D-DWT

The Discrete Wavelet transform can be implemented as a combination of a down-sampler and a filter bank that de-

978-1-5386-6805-4/18/$31.00 ©2018 IEEE
DOI 10.1109/ICMLA.2018.00023

composes a 1D or a 2D signal (i.e. image) into sub-bands at different resolution levels [12]. The digital filter bank contains high-pass (g) and low-pass (h) filters. Because the sub-images contain the high and low frequency information, the Wavelet transform can extract useful texture features from mammographic ROIs. The two-level Wavelet decomposition of a two-dimensional signal is shown in Fig. 4.

To compute 2D-DWT, one-dimensional DWT is applied on each column and the row of the input image separately. Fig. 5 illustrates the 2D-DWT of a given mammographic ROI. The right image in this figure illustrates the decomposition of the original image on the left into four sub-bands in the frequency domain, which represent low-low (LL), low-high (LH), high-low (HL), high-high (HH) components corresponding to approximation, horizontal, vertical, and diagonal, respectively. Three of the sub-band images, HH, HL and LH contains the detail information for different orientations and resolutions of the ROIs while the LL involves the coarse approximation at a particular resolution level. The LL sub-band can be decomposed further into sub-bands repeatedly to increase the resolution of the Wavelet. In this paper, the one level 2D-DWT in PyWavelets library of Python has been used [13] to compute the Wavelet transform. In the next step, all of the four sub-images are used to extract the Haralick features while only the approximation image is used to extract the HOG features.

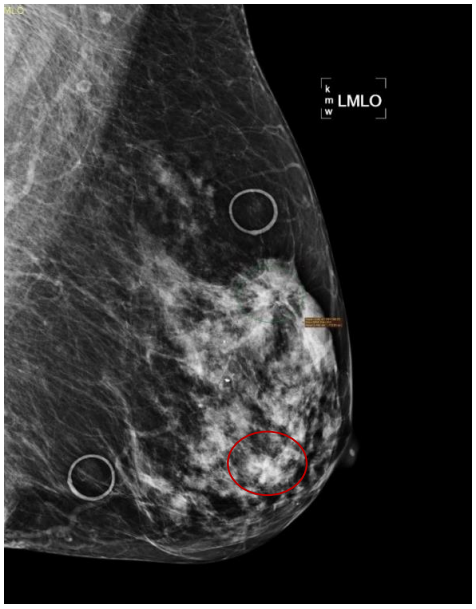


Figure 2. A ROI marked as a red circle in a mammogram image. White circles represent mole markers

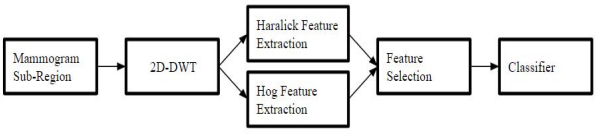


Figure 3. Block diagram of the proposed ROI detection method

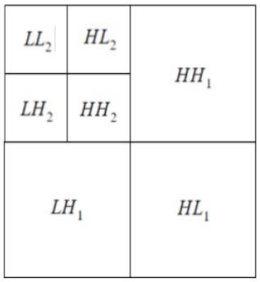
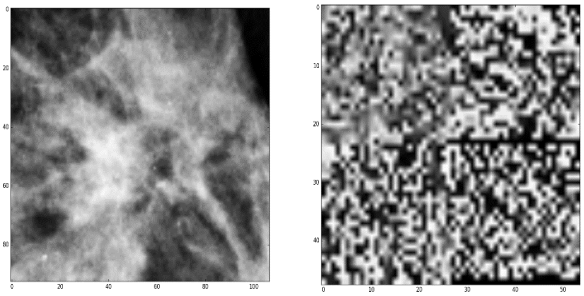


Figure 4. 2-D Wavelet decomposition of an image signal



a b

Figure 5. Mammographic ROI (a) and one-level 2D-DWT of the ROI (b)

2) Histogram of Oriented Gradients (HOG)

HOG features are used by the computer vision community to detect objects and localize them [14]. It is based on the theory that a mass or shape can be distinguished by differential intensity histogram of the local intensity gradients or edge directions. Each mammographic ROI is divided into non-overlapping uniform windows called cells. For each cell, the differentials for the desired orientation are calculated. These are called the gradients G_X and G_Y in x and y directions, respectively, which are formulated in (1) and (2). The gradient which is related to differentials is constructed from group of cells.

$$G_X = \frac{\partial f(x,y)}{\partial x} = \frac{f(x+1,y)-f(x-1,y)}{(x+1)-(x-1)} \quad (1)$$

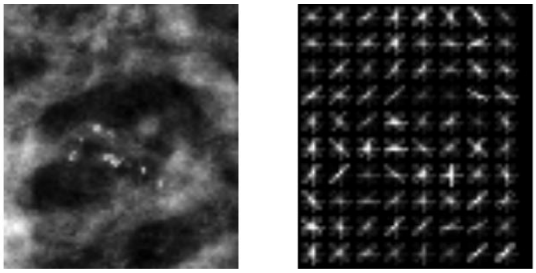
$$G_Y = \frac{\partial f(x,y)}{\partial y} = \frac{f(x,y+1)-f(x,y-1)}{(y+1)-(y-1)} \quad (2)$$

The second step includes calculating the magnitude and orientation of the gradient for each pixel in the image, which can be achieved using (3) and (4). The orientation represents the angle of the gradients which are evenly spread from -180 to 180 degrees (signed) or 0 to 360 degrees (unsigned).

$$|G| = \sqrt{G_X^2 + G_Y^2} \quad (3)$$

$$\theta(x,y) = tan^{-1}\left(\frac{G_Y^2}{G_X^2}\right) \quad (4)$$

Fig. 6 shows the visualization of a malignant ROI and its HOG representation. The HOG feature extraction method extracts 16 features and it is summarized in Algorithm 1.



a b

Figure 6. The illustration of a malignant ROI (a) and its HOG representation (b)

Algorihm 1. HOG feature extraction

1 //input I: The mammographic ROI image
2 //output: 16 HOG features
3 Gx = Sobel filter of input image (vertical);
4 Gy = Sobel filter of input image (horizontal);
5 Magnitude, Angle = Calculate the magnitude and angle of Gx and Gy;
6 Number of Bins = 16;
8 Bin = (Number of Bins * Angle) / 360;
9 Features = Count number of occurrences of each value in array of Bin using Magnitude as weight where minimum length is set equal to the number of bins;
10 Return Features;

3) Haralick's Textural Features

The texture of an image bears many characteristics that can be used to identify a ROI. In this paper, the gray level co-occurrence matrices (GLCM) are used to extract 14 Haralick's texture features using the Mahotas library of Python [15]. GLCM is a technique that can be used to compute the texture information by capturing the spatial relationships of the pixels [16]. The GLCM calculates the spatial relationship by counting the pixel pairs that have the same values for a given distance and direction (Fig. 8). For this purpose, a co-occurrence matrix is computed for each of the four directions (0°, 45°, 90° and 135°) with one-pixel distance. (Fig. 7) The size of the GLCM becomes equal to the range of gray-level value of the given image. This feature extraction method extracts 56 features. Moreover, both Haralick's method and GLCM feature extraction method calculates the co-occurrence matrix. After the calculation Haralick extracts 56 features while GLCM method extracts 4 features.

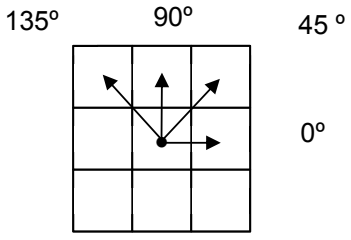


Figure 7. Spatial Information of GLCM

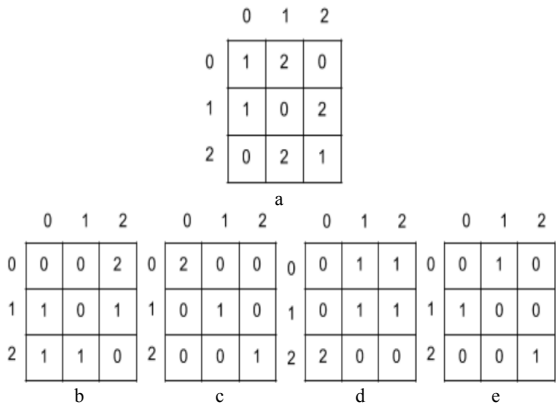


Figure 8. Computation of Matrices with unit distance and four directions. (a) Gray Level Values of the Input Image (b) 0º GLCM c) 45º GLCM d)90º GLCM e) 135º GLCM

Fig. 8 Describes the computation of the GLCM in a given image. The gray level of the image is three which is the size of the matrices. The co-occurrence matrices for each direction are calculated by counting the pixel pairs for the specified direction. After counting process, numbers have been placed in an appropriate position.

D. Feature Selection

The extracted features are mathematical descriptors, which can be employed by a classifier to make a decision. If the number of features is high, the classification model can suffer from the presence of irrelevant features as well as noise contained in features. Furthermore, as the number of dimensions increases the classification model can be prone to over-fitting, which reduces the prediction accuracy on unseen test examples. To address these problems, feature selection can be used. The major task is to select a feature set, which contains a subset of the original features that are relevant for predicting the output class. In this paper, the wrapper feature selection method with best first search strategy is used for feature subset selection, which is implemented using WEKA (Waikato Environment for Knowledge Analysis) [17]. Wrapper method basically finds the optimal feature subset using a classifier method. In the present study, a random forest classifier with default parameters is used for the wrapper style selection of features.

E. Classifier Methods and Cross-validation Experiment

The following classifiers are implemented in WEKA software and optimized to predict whether a given image is a ROI or not: random forest, support vector machine (SVM) with RBF kernel and AdaBoost with default base learner [18] [19] [20]. To assess the prediction accuracy of each method, a leave-one-image-out cross-validation experiment (LOOCV) is used [21]. In each iteration of the LOOCV, one image is selected as the test data and rest are used as the train set. This process is repeated until all images have been used as the test sample. Before training the classifiers, feature selection is applied on each train set of the LOOCV and the feature subset is selected accordingly for the test set. Similarly, the number of trees parameter of random forest, the number of iterations parameter of AdaBoost and C, gamma parameter pairs of SVM are optimized separately on each training set. The optimization has been done on train

sets using 10-fold cross-validation. The optimum parameters have been used on testing phase.

III. RESULTS AND DISCUSSION

Fig. 9 shows the frequency histogram of the features that are selected the most often across the individual folds of the LOOCV experiment. The features numbered from 17 to 240 are Haralick's Features, the remaining are HOG features. It is found that Haralick's features are selected with higher frequency than the HOG features.

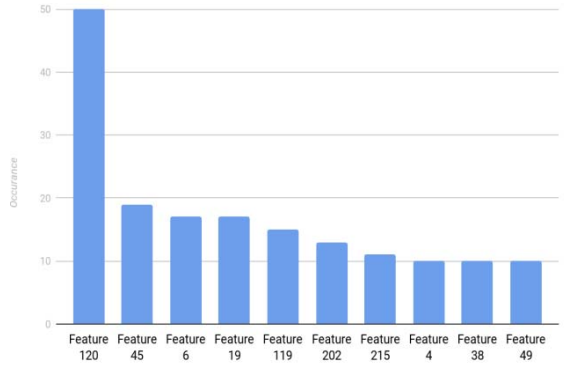


Figure 9. Frequency histogram of selected features across different folds of cross-validation (LOOCV)

The frequently selected features are sum average, angular second moment, correlation, inverse difference moment and variance. For instance, 120th feature represents the sum average of 0º GLCM of the wavelet decomposed high-high component of ROIs.

After selecting the feature subset, the min max normalization has been done, each classifier is trained on train set and prediction is computed on test set. Also, optimum parameters were found for each training set and LOOCV fold, optimum number of trees for Random Forest classifier across the different folds differs but frequently 100 trees were selected. For Adaboost classifier the optimum number of iteration changes for each fold, but mostly selected number of iterations was 20. For SVM the optimum parameters were found when C and gamma values were 2 and 0.5 respectively. This procedure is repeated for all folds of the LOOCV. Table 1 shows various accuracy measures of random forest, SVM and AdaBoost results with proposed methods relevant features and optimum parameters. According to this table, the random forest classifier provided the best accuracy measures.

TABLE I. ACCURACY METRICS OF ROI CLASSIFIERS

	Accuracy Measures						
	Accuracy	***F-Measure***	***FP Rate***	***Specificity***	***Precision***	***Recall***	***AUC Score***
RF	87.1 %	87.5%	16.1%	83.8%	84.8%	90%	84%
SVM	77%	78%	25%	74%	75%	80%	77%
Adaboost	69%	69%	32%	67%	68%	71%	72%

As a third experiment, different feature extraction methods are implemented and compared for ROI detection using a random forest classifier. According to Table 2, the best ROI detection accuracy is obtained when the proposed feature extraction method is employed. This is obtained for all accuracy measures except for AUC, in which the Haralick descriptor achieved a slightly better accuracy. Therefore combining information contained in Haralick and HOG features, which are computed from Wavelet transformation is useful for ROI detection. Fig. 10 shows the ROC curves of the Haralick and the propsed Wavelet-Haralick-HOG feature extraction methods.

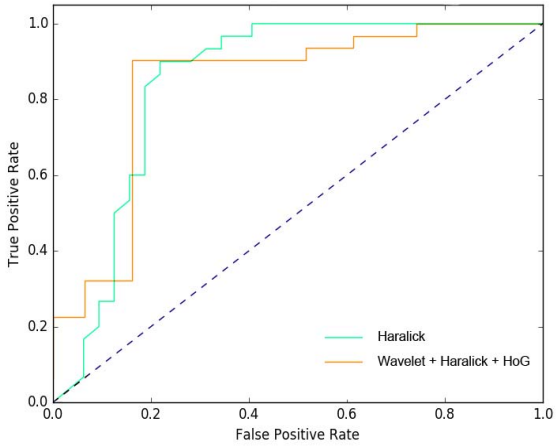


Figure 10. ROC curves of the Haralick and the proposed Wavelet-Haralick-HOG descriptor (feature selection is applied in LOOCV setting for each method)

TABLE II. ACCURACY METRICS OF DIFFERENT FEATURES

	Accuracy Measures						
	Accuracy	***F-Measure***	***FP Rate***	***Specificity***	***Precision***	***Recall***	***AUC Score***
Haralick	77.5 %	75.9%	18.8%	81.3 %	78.6 %	73.3%	85%
HOG	43.5 %	42.6%	56.3%	43.8 %	41.9 %	43.3%	44%
LBP	50%	45.6%	43.7%	56.3 %	48.1 %	43.3%	46%
TAS	56.45%	55.7%	43.8%	56.3 %	54.8 %	56.7%	62%
Wavelet + Haralick	71%	70%	28.1%	71.9 %	70%	70%	79%
Wavelet + HOG	46.8 %	50.7%	62.5%	37.5 %	45.9 %	56.6%	44%
Wavelet + GLCM	58.1 %	58.1%	43.8%	56.3 %	56.3 %	60%	66%

Wavelet + Haralick + HOG (Proposed method)	87.1 %	87.5%	16.1%	83.8 %	84.8 %	90%	84%

IV. Conclusion and Future Work

In this work, a new feature extraction method is proposed for ROI detection in mammogram images. The proposed method first computes the Wavelet transform of the selected image, followed by extracting HOG and Haralick descriptors which compute textural and gradient features. Then the best set of features is selected using a wrapper strategy. When employed in a random forest classifier the proposed feature extraction method achieves the best ROI detection accuracy. This method can be used for automatic detection of ROIs in a CAD system. For this purpose, first, a mammogram image can be subdivided into small square-sized images by applying a sliding window. Then each image can be classified as ROI (positive) or not (negative). This can be explored further as a future work. Furthermore, the performance of the method can be analyzed on other clinical databases in order to verify the feasibility and adaptability of the results. Finally, the proposed method can be combined and compared with other machine learning and deep learning methods on larger image databases for breast cancer detection

Acknowledgment

The authors would like to thank Assoc. Prof. Fahrettin Kılıç for his valuable contribution to dataset preparation.

References

[1] X. Lladó, A. Oliver, R. M. J. Freixenet and J. Martí, "A textural approach for mass false reduction in mammography," *Computerized Medical Imaging and Graphics,* pp. 415-422, 2009.

[2] X. Xianchuan and Q. Zhang, "Medical Image Retrieval Using Local Binary Patterns," in *International Conference on Information Engineering and Computer Science*, 2009.

[3] M. Pratiwi, Alexander, J. Harefa and S. Nanda, "Mammogram classification using gray-level co-occurrence matrix and radial basis function neural network," in *International Conference on Computer Science and Computational Intelligence*, 2015.

[4] K. Kanadam and S. Chereddy, "Mammogram classification using sparse-ROI: A novel representation to arbitrary shaped masses," *Expert Systems with Applications,* pp. 204-213, 2016.

[5] V. Pomponiu, H. Hariharan, B. Zheng and D. Gur, "Improving Breast Mass Detection using Histogram of Orieneted Gradients," in *Medical Imaging*, 2014.

[6] S. Beura, B. Majhi and R. Dash, "Mammogram Classification using two dimensional discrete wavelet transform and gray-level co-occurance matrix for detection of breast cancer," *Neurocomputing,* pp. 1-14, 2015.

[7] K. Bovis and S. Singh, "Detection of Masses in Mammograms Using Texture Features," *IEEE,* pp. 267-270, 2000.

[8] R. Ragayyan, T. Nyugen, F. Ayres and A. Nandi, "Effect of Pixel Resolution on Texture Features of Breast Masses in Mammograms," *Journal of Digital Imaging,* vol. 23, no. 5, pp. 547-553, 2010.

[9] S. Ergin and O. Kilinc, "A new feature extraction framework based on wavelets for breast cancer diagnosis," *Computers in Biology and Medicine,* 2014.

[10] "The Digital Mammography DREAM Challenge," [Online]. Available: https://www.synapse.org/#!Synapse:syn4224222/files/.

[11] "The Digital Mammography Dream Challenge," [Online]. Available: https://www.synapse.org/#!Synapse:syn4224222/wiki/401750.

[12] S. Mallat, "A Theory For Multiresolution Decomposition: The Wavelet Representation," *IEEE Trans. Pattern Anaylsis and Machine Intelligence,* vol. 11, no. 7, pp. 674-693, 1989.

[13] G. Lee, F. Wasilewski, R. Gommers, K.Wohlfahrt, A. O'Leary and H. N. a. Contributors, "PyWavelets - Wavelet Transforms in Python," *2006.*

[14] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.

[15] L. Coelho, "Mahotas: Open source software for scriptable computer vision," *Journal of open research software,* 2013.

[16] R. M. Haralick, K. Shanmugam and I. Dinstein, "Textural Features For Image Classification," *IEEE Trans System and Cybernetics,* Vols. SMC-3, no. 6, pp. 610-621, 1973.

[17] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations ,* vol. 11, no. 1, pp. 10-18, 2009.

[18] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001.

[19] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, no. 3, pp. 273-297, 1995.

[20] Y. Freund and R. Schapire, "A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting," 1995.

[21] R. Duda, P. Hart and D. Stork, Pattern Classification, 2nd Edition, New York: Wiley, 2001.

[22] S. Lui, C. F. Babbs and E. J. Delp, "Multiresolution Detection of Spiculated Lesions in Digital Mammograms," *IEEE Trans. Image Process.,* 2001.

[23] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence,* pp. 273-324, 1997.

[24] A. Oliver, J. Freixenet, J. Martin, E. Perez, J. Pont and E. Denton, "A review of automatic mass detection and segmentation in mammographic images," *Medical Image Analysis,* pp. 87-110, 2010.

[25] M. Eltoukhy, I. Faye and B. Samir, "A statistical based feature extraction method for breast cancer diagnosis in digital mammogram using multiresolution representation," *Computers in Biology and Medicine,* pp. 123-128, 2012.

[26] R. Sivaramakrishna, K. Powell, M. Lieber, W. Chilcote and R. Shekhar, "Texture analysis of lesions in breast ultrasound images," *Computerized Medical Imaging and Graphics,* pp. 303-307, 2002.

[27] M. Berbar, "Hybrid Methods for feature extraction for breast masses classification," *Egyptian Informatics Journal,* 2017

[28] Gonzalez. R, and Woods. R, Digital Image Processing, Upper Saddle River, N.J.: Prentice Hall,2008.