Avestia Publishing
Journal of Biomedical Engineering and Biosciences
Volume 3, Year 2016
ISSN: 2564-4998
DOI: 10.11159/jbeb.2016.002

Automated Breast Tumour Detection in Ultrasound Images Using Support Vector Machine and Ensemble Classification

Passant Wahdan[1], Amani Saad[1], Amin Shoukry[2]
[1]Computer Engineering Department, Arab Academy for Science and Technology
Alexandria, Egypt
passantwahdan89@gmail.com; amani.saad@aast.edu
[2]Computer Science and Engineering Department, Egypt-Japan University of Science and Technology
Alexandria, Egypt
amin.shoukry@ejust.edu.eg

Abstract - *Breast cancer is the second leading cause of death in women after heart diseases. A well-known statement in cancer society is "Early detection means better chances of survival". In the past few years several techniques were developed to detect breast tumors in early stages. A proposed system is designed for breast tumors detection using ultrasound images. Ultrasound is used because it is less expensive and less invasive than X-rays used in mammography and computerized tomography. It can provide a second opinion for a physician to detect breast tumors. The proposed system consists of three main steps: pre-processing, feature extraction and classification. Gaussian blurring, anisotropic diffusion and histogram equalization are used to reduce additive noise, speckle noise and to enhance the image quality respectively. The second step is feature extraction and dimensionality reduction. PCA is used to reduce the dimensions of the feature vector. The third and final step is the classification step. A comparison is conducted between support vector machine and bagging ensemble classifier as different classification techniques. The third step is deployed to classify the images into image with/without tumors.*

Keywords**:** breast cancer, ultrasound image, image preprocessing, feature extraction, dimensionality reduction, classification.

Date Received: 2014-10-14
Date Accepted: 2016-03-20
Date Published: 2016-05-03

1. Introduction

Different imaging techniques have been developed to detect breast cancer in early stages to assist in obtaining better chance of recovery. These techniques include thermography, mammography and ultrasound imaging. Ultrasound is the main focus of this research. Ultrasound is more reliable than mammography for women under forty. Using ultrasound imaging, one can find the exact location, shape and size of a tumor while in thermography only the presence of the tumor is indicated. However, the main problem with ultrasound imaging is the noise caused by imperfect instruments, the data acquisition process as well as other natural interfering phenomena as shown in [1], which complicates the detection process. In order to solve this problem, filtering and enhancement are needed.

The objective of this paper, is to develop a breast tumor detection system using ultrasound images. It includes three stages. The first one aims at reducing additive and speckle noises and enhancing the contrast. The second one aims at selecting a good set of features and reducing the dimensionality of the feature vector. The final stage uses a binary classifier to decide whether the input image includes a tumor or not.

The rest of the paper is organized as follows; Section 2 presents background material. Section 3 describes the proposed system, Section 4 discusses the experimental work and, finally, Section 5 presents the conclusions and future work.

2. Background

Cancer is one of the leading causes of death worldwide. One in eight deaths worldwide is due to cancer as demonstrated in [2]. According to the American Cancer Society, the probability of getting breast cancer is 1:2000 at the age from twenty to twenty nine, 1: 229 from thirty to thirty nine, 1:68 from forty to forty nine, 1:37 from fifty to fifty nine, and 1:26 from sixty to sixty nine. In 2010, worldwide breast cancer web site declared that nearly 1.5 million people were diagnosed of breast cancer. 89% of women diagnosed with breast cancer are still alive five years after their diagnosis. Dramatically, one-third of breast cancer death can be decreased if detected and treated early, this means that nearly 400, 000 lives could be saved every year presented in [3].

The detection process is usually divided into three phases: the preprocessing, the feature extraction and selection and the classification phases. Different techniques are used in each phase. The preprocessing phase deals with different kinds of noise such as amplifier (Gaussian), Salt and pepper, Poisson and Speckle noise. Different kinds of filters can be applied to remove noise such as Mean, Median, Gaussian and anisotropic diffusion filters as shown in [7], and [8]. Morphological operators and Histogram equalization are examples of enhancement techniques used in [9].

Several classification techniques related to breast tumors have been developed for detecting and differentiating between cancerous and benign tumors using ultrasound images. In [4], a Computer Aided Diagnosis (CAD) system has been developed to classify breast tumors using Support Vector Machines (SVM). Another system, described in [5], also aimed at classifying breast tumors in ultrasound images using a hybrid classifier based on a multilayer perceptron network and genetic algorithms. Later, in [6], another CAD system is described to detect and segment the tumor regions. The detection algorithm works in two stages: tumor localization and tumor boundary delineation. In the first stage, an AdaBoost classifier using Haar-like features is applied followed by an SVM classifier.

3. Proposed System

The main focus of this research is detecting the presence of tumours in ultrasound images. We propose a system in which image pre-processing and enhancement techniques are applied in order to improve the detection accuracy. A block diagram representing the proposed system is shown in figure 1.

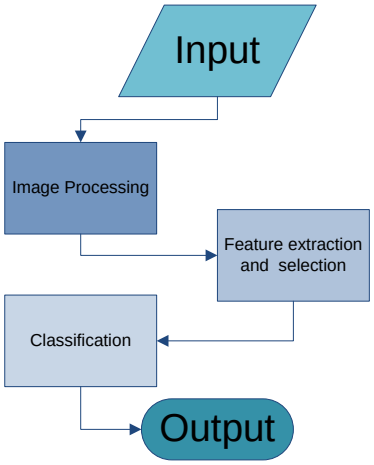


Figure 1. Block diagram of the proposed system.

Figure 2. Example of an original ultrasound image of a breast tumor.

3.1. Pre-processing Stage

The main concern in the pre-processing stage is applying de-noising and enhancement techniques without destroying the useful information in the input image. The pre-processing stage is divided into two steps: filtering and enhancement. Gaussian filters are used to get rid of additive noise. Also, anisotropic diffusion filter is used to overcome the major drawbacks of conventional spatial filters and improve the image quality by preserving important boundary information this is concluded from [10]. To further improve the image quality, histogram equalization is used for image enhancement as illustrated in [11].

3.1.1. Gaussian filter

The Gaussian filter is a non-uniform low pass filter. It removes high-frequency components from the image without affecting the important data in it, as

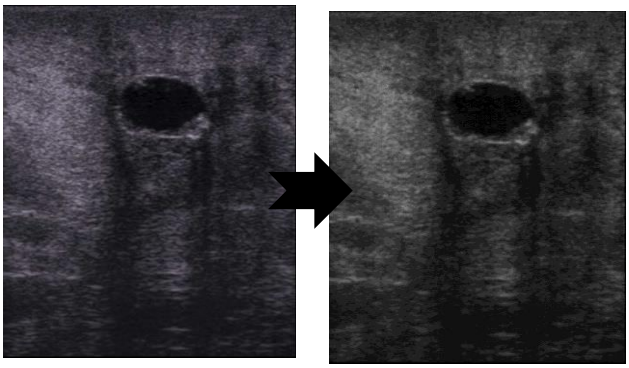
shown in Figure 3. However, it is not particularly effective at removing salt and pepper noise.

Gaussian smoothing is used in order to enhance image structures at different scales. Mathematically, a Gaussian filter modifies the input signal by convoluting it with a Gaussian function, this transformation is also known as the Weierstrass transform. The Gaussian function is:

$$G(x_1, \dots, x_k) = \frac{1}{\sqrt{(2\pi)^k |\sum|}} \exp(-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu)) \quad (1)$$

Where x is a real k-dimensional column vector, $|\sum|$ is the determinant of $\sum$ (the covariance matrix) and μ is the mean vector.

Gaussian filtering has its basis in the human visual perception system. It has been found that neurons create a similar filter when processing visual images. The Gaussian function is used in numerous research areas as mentioned in [12].

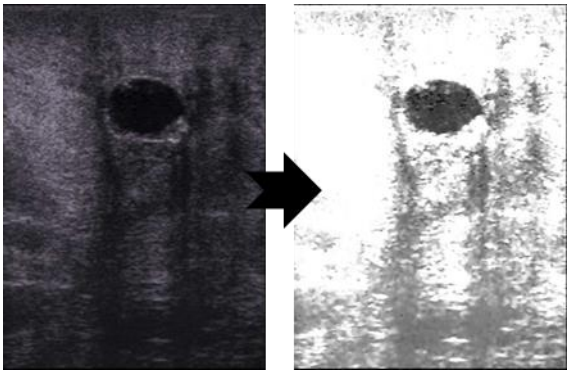


(a) (b)

Figure 3. (a) Original image with a tumor
(b) The same image after applying Gaussian filter.

3.1.2. Anisotropic Diffusion

Anisotropic diffusion reduces the speckle noise and also blurs the image without compromising its quality as shown in Figure 4. The main idea in anisotropic diffusion is to smooth the homogenous areas of the image while enhancing the edges. This creates a piecewise constant image from which the segmentation boundaries can be easily obtained which was brought to light in [4]. The anisotropic diffusion is implemented using the derivation of Speckle Reducing Anisotropic Diffusion (SRAD) proposed in as presented in [7].



(a) (b)

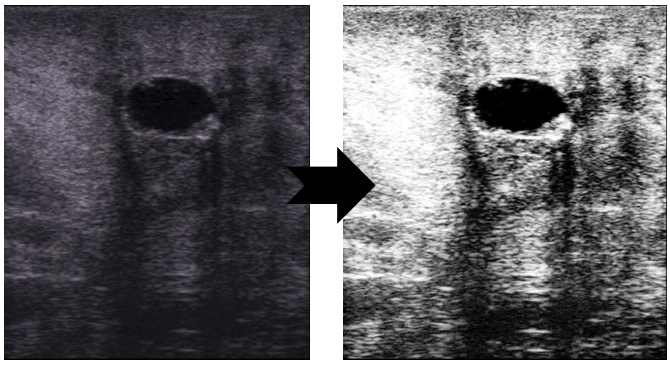
Figure 4. (a) Original image with a tumor
(b) The same image after applying Anisotropic Diffusion.

3.1.3. Histogram Equalization

The aim of image enhancement is to improve the image quality, or to provide better input for other automated image processing techniques. Histogram equalization is known to adjust image intensities to enhance contrast as shown in Figure 5. This helps in reducing differences among images from various ultrasonic systems. Figure 6 shows the histogram of an image before and after equalization. The equalized histogram for pixel SK is defined as follows:

$$S_K = \sum_{j=0}^{K} P_r(r_j) \quad (2)$$

Where r_j represent the gray level of the pixel to be enhanced, K= 0,1,2,3,..., L-1 where L is the total number of possible gray levels in the image and P_r is the probability shown in [12].



(a) (b)

Figure 5. (a) Original image with a tumor
(b) The same image after applying Anisotropic Diffusion.

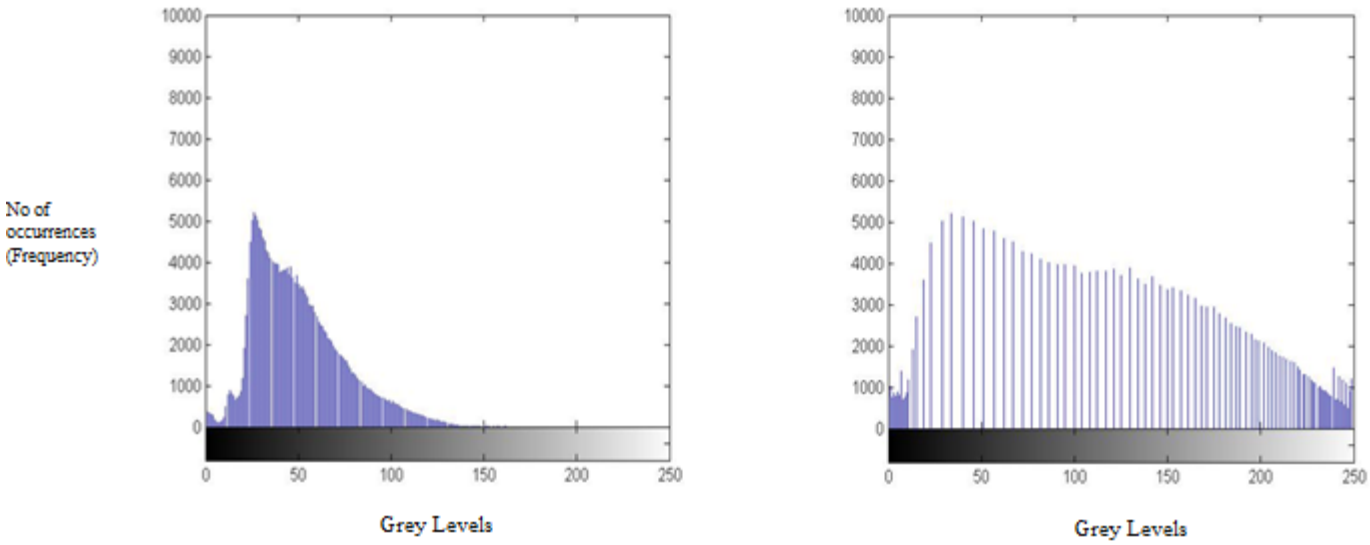


Figure 6. (a) Image histogram before equalization. (b) Image histogram after equalization.

3.2. Feature Extraction and Selection

Textural parameters calculated from the Gray Level Co-occurrence Matrix (GLCM), of a preprocessed input image, helps in understanding the image content as presented in [13]. The GLCM is a powerful measure used in texture classification. In order to classify the image properly, the following sixteen features are extracted from the GLCM as mentioned in [14].

1. Autocorrelation	5. Energy	9. Sum variance	13. Information measure of correlation
2. Contrast	6. Entropy	10. Sum entropy	14. Inverse difference
4. Cluster Shade	7. Homogeneity	11.Difference variance	15. Inverse difference normalized
3.Cluster Prominence	8. Sum of squares variance	12. Difference entropy	16. Inverse difference moment normalized

Principal component analysis (PCA) is used to reduce the dimensionality of the feature space of the data set, while retaining as much as possible of the variation present in it. PCA is achieved by transforming the data set into a new set of uncorrelated variables which are ordered so that the first few retain most of the variation present in all of the original variables as indicated in [15]. Using PCA, 4 factors were found to contain most of the variation present in the original dataset. The use of a reduced set of uncorrelated features enhances the final classification stage.

3.3. Classification

Two classification approaches have been used in the present work, support vector machine and ensemble classifier using bagging technique.

3.3.1. Support Vector Machine

SVM is a supervised learning technique that seeks an optimal hyper-plane to separate two classes of samples. Mapping the input data into a higher dimensional space is done by using Kernel functions with the aim of obtaining a better distribution of the data. Then, an optimal separating hyper-plane in the high-dimensional feature space can be easily found as shown in [16]. An example of an optimal Hyper-plane is shown in Figure 7.

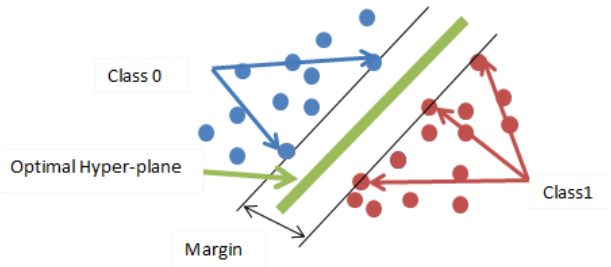


Figure 7. Optimum Hyper-plane for Support Vector Machine.

3.3.2. Bagging

Bagging is a machine learning ensemble meta-algorithm used to improve the stability and accuracy of a classifier. It reduces the variance and helps avoid over-fitting. Bagging is a special case of the model averaging approach. It randomly distorts the data set by re-sampling it. Bagging seems to enhance accuracy when random features are used which was stated in [17]. Reduced-Error Pruning (REP) tree is a fast decision tree learner as shown in [18]. It builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (with back-fitting). Bagging is used to grow an ensemble of trees and let them vote for the most popular class. Figure 8 shows an example of an ensemble classifier.

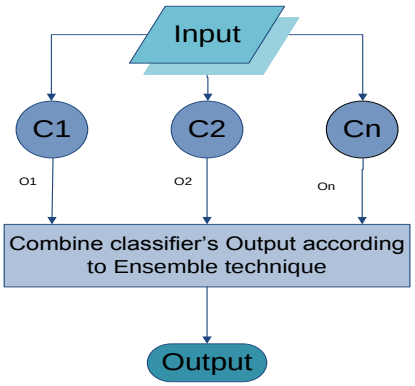


Figure 8. Ensemble Classifier.

Having explained each stage of the proposed model, a block diagram of the system is shown in Figure 9.

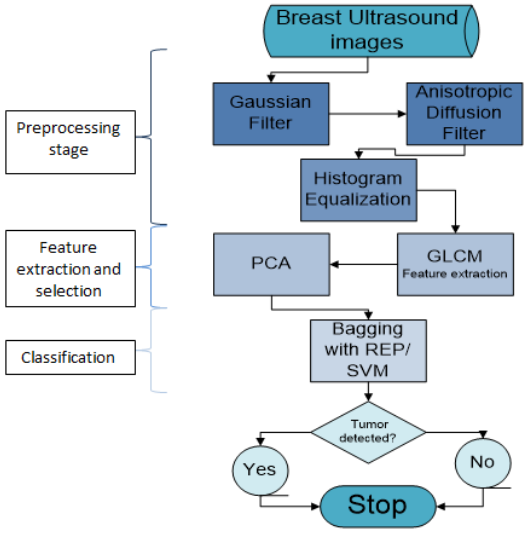


Figure 9. Block diagram presenting the proposed system.

4. Experimental Work

The feature extraction and classification techniques, presented in section 3, have been applied on 106 ultrasound images (55 images clean of tumor and 51 containing tumor) obtained from different medical centers in Alexandria, Egypt. The SVM classifier has been trained on 70% of the data set and tested on the remaining 30%. On the other hand, bagging ensemble has used all the data using bootstrap technique.

Both the filtering and enhancement steps were implemented using matlabR2012a. The results of applying the pre-processing techniques are shown in Figures 10 and 11. PCA has been implemented using LSTAT (an add-on-Excel) which is shown in [19]. Finally, the two classification techniques have been implemented using Weka 3.6.9 which is denuded from [18].

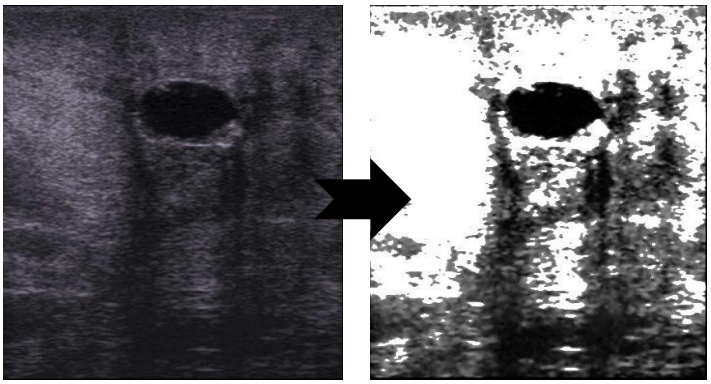


Figure 10. Breast ultrasound image after filtering and enhancement.

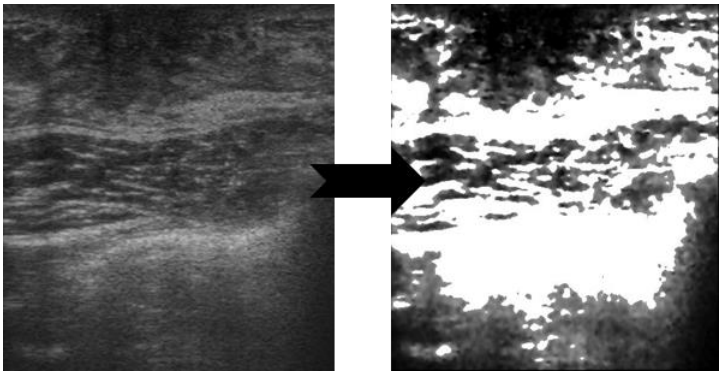


Figure 11. Clean breast ultrasound image after filtering and enhancement.

4.1. Preprocessing Evaluation

A comparison between different combinations of the proposed image preprocessing techniques is presented in Table 1.

Table 1. Performance measures of different combination of image processing techniques.

Performance Measure	Histogram	Gaussian-Histogram	Gaussian	Diffusion	Gaussian-Diffusion	Histogram-Diffusion	All
Accuracy	79.4%	81.3%	82.3%	82.3%	83.3%	83.3%	85.9%
True Positive Rate	79.4%	81.4%	82.4%	82.4%	83.3%	83.3%	85.5%
False positive Rate	22.2%	20.9%	19.4%	19.4%	18.9%	18.9%	12.4%
Precision	80.1%	83.2%	83.4%	83.4%	85.4%	85.4%	87.2%
Recall	79.4%	81.4%	82.4%	82.4%	83.3%	83.3%	85.8%
F-Measure	79.1%	80.9%	82%	82%	82.9%	82.9%	85.9%

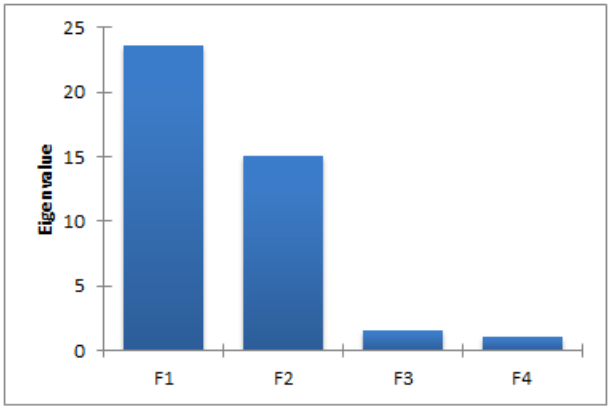


Figure 12. A plot between factors and their Eigen-values

A plot between components (factors) and their corresponding Eigen-values is shown in Figure 12. It presents the four principal components (factors) that were found to contain more than 98% of the variation in the original dataset, as a result of the principal component analysis conducted on the 16 features extracted from the gray level co-occurrence matrix.

4.2. Classification Techniques

4.2.1 Support Vector Machine

The classifier has been trained on 70% of the data and tested on the remaining 30%.The correctly classified instances were 75.5%. The performance measures of the SVM classifier are shown in Table 2. TP and FP-rates correspond to true and false positive rates, respectively. “Positive” represents the class of images defined by a physician as images with tumors while the“Negative”class represents clean images.

4.2.2. Bagging

Bagging implies training a set of REP trees using bootstrapping. The correctly classified instances were 85.9% while the incorrectly classified instances were 14.1%. The performance measures of the Bagging classifier are shown in Table 3. The performance of bagging is expected since bagging trains several REP trees with different data and combine the decisions.

Table 2. The performance measures of the SVM classifier after applying image processing techniques.

	TP Rate	FP Rate	Precision	Recall	F-Measure
Positive Class	70.2%	38.2%	61.1%	70.2%	65.3%
Negative Class	61.8%	29.8%	70.8%	61.8%	66%
Weighted Avg.	65.7%	33.7%	66.4%	65.7%	65.7%

Table 3. The performance measures of the Bagging classifier after applying image processing techniques.

	TP RATE	FP RATE	PRECISION	RECALL	F-MEASURE
POSITIVE CLASS	80%	6.5%	94.1%	80%	86.5%
NEGATIVE CLASS	93.5%	20%	78.2%	93.5%	85.1%
WEIGHTED AVG.	85.8%	12.4%	87.2%	85.8%	85.9%

4.3. Discussion

There are many types of binary classifiers that can be used. Two classifiers have been selected: Support Vector Machine and an ensemble classifier using bagging to train a set of REP trees. The bagging ensemble classifier with REP tree was found to give the best accuracy of 85.9% while the Support Vector machine achieved an accuracy of 75.5% after applying all image and data processing techniques suggested before. Figure 13 shows an example of an image with tumor passing all the preprocessing techniques.

By comparing the performance measures of the bagging with REP classifier before and after the preprocessing stage and the dimensionality reduction of the data set, we conclude that by applying the proposed methodology the performance measures did improve by 8.5%.

Several combinations of image preprocessing techniques have been implemented to reach the optimal performance measures. Although by the naked eye, after applying histogram equalization the image seemed like it would give better performance measures it only gave 79.4% accuracy. However, Gaussian filter gave better accuracy of 82.3%. The combination of Gaussian, Anisotropic Diffusion and Histogram Equalization gave the best performance measures of 85.9%.

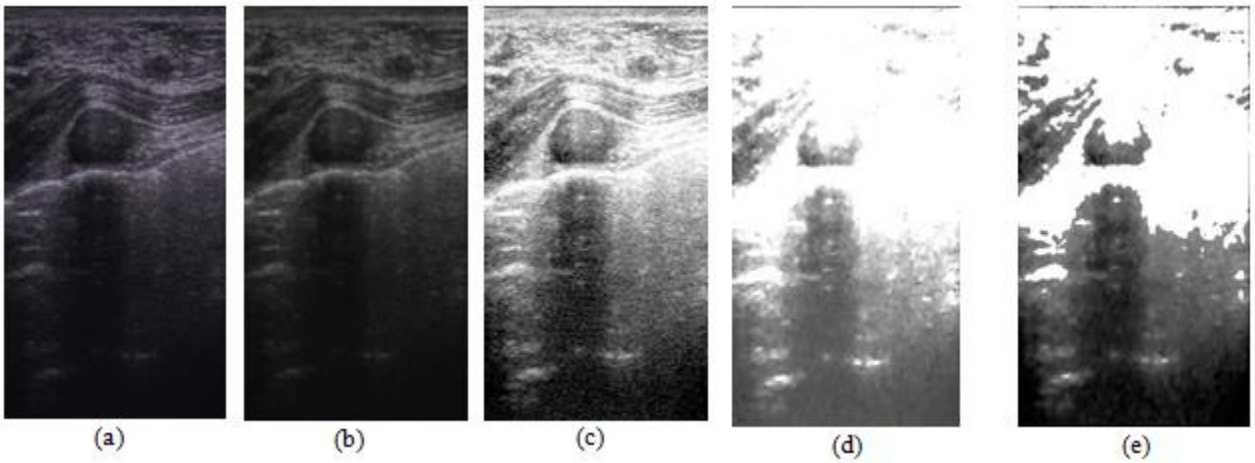


Figure 13. (a) Original Image (b) Image after Gaussian (c) Image after Histogram (d) Image after Diffusion (e) Image after all image pre-processing techniques.

5. Conclusions and Future Work

A system has been proposed to serve as a second opinion for a physician to detect breast tumors in ultrasound images. This system consists of three stages: Preprocessing, Feature extraction and Selection and Classification into clean images and images with tumors. The obtained result have shown the superiority of bagging ensemble classifier over the SVM classifier.

In the future, new cases should be added to the data set. More image processing techniques will be added to the system to improve its accuracy. We plan to train the designed classifiers using additional features as well as other classifiers as members of the ensemble classifier. More functions such as tumor localization, segmentation and classification into benign and malignant tumors will be added. The system can be extended to fuse inputs from different data sources: Ultrasound, X-rays etc… to increase the detection rate.

References

[1] Y. Guo, "Computer-Aided Detection of Breast Cancer Using Ultrasound Images," Utah State University, Logan, 2010.

[2] M. Garcia, A. Jemal, E. Ward, M. Center, Y. Hao, R. Siegel and M. Thun, "Global Cancer Facts & Figures," American Cancer Society, Atlanta, 2007.

[3] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward and D. Forman, "Global Cancer Statistics," *CA: Cancer Journal for Clinicians*, vol. 61, no. 2, pp. 69-90, 2011.

[4] Y.-L. Huang, K.-L. Wang and D.-R. Chen, "Diagnosis of breast tumors with ultrasonic texture analysis using support vector machines," *Neural Computing & Applications*, vol. 15, no. 2, pp. 164-169, 2006.

[5] A. V. Alvarenga, W. C. A. Pereira, A. F. C. Infantosi and C. M. Azevedo, "Classifying Breast Tumours on Ultrasound Images Using a Hybrid Classifier and Texture Features," in *Intelligent Signal Processing*, Alcala de Henares, 2007.

[6] P. Jiang, J. Peng, G. Zhang, E. Cheng, V. Megalooikonomou and H. Ling, "Learning Based Automatic Breast Tumor Detection and Segmentation in Ultrasoung Images," in *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, Barcelona, 2012.

[7] O. V. Michailovich and A. Tannenbaum, "Despeckling of medical ultrasound images," 2006.

[8] V. S. Frost, A. S. Josephine, K. S. Shanmugan and J. C. Holtzman, "A Model for Radar Images and Its Application to Adaptive Digital Filtering of Multiplicative Noise," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-4, no. 2, pp. 157- 166, 2009. D. Gadkari, Image Quality Analysis Using GLCM, 2014.

[9] R. Hummel, "Image enhancement by histogram transformation," *Computer Graphics and Image Processing*, vol. 6, no. 2, pp. 184-195, 1977.

[10] M. S. Minavathi, S. Murali and M. S. Dinesh, "Classification of Mass in Breast Ultrasound Images using Image Processing Techniques," *International Journal of Computer Applications*, vol. 42, no. 10, pp. 29-36, 2012.

[11] Q. Wang, L. Chen and D. Shen, "Fast histogram equalization for medical image enhancement," in *Conference Proceedings IEEE Eng Med Biol Soc*, 2008.

[12] R. C. Gonzales and P. Wintz, Digital image processing (2nd ed.), Boston: Addison-Wesley Longman Publishing Co., Inc., 1987.

[13] D. Gadkari, "Image Quality Analysis Using GLCM," M.S. thesis, University of Central Florida, Orlando, 2004.

[14] R. M. Haralick, K. Shanmugam and I. Dinstein, "Textural Feartures for Image Classification," IEEE Transactions on Systems Man and Cybernetics, vol. SMC3, no. 6, pp. 610-621, 1973.

[15] H. Abdi and L. J. Williams, "Principal Component Analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433-459, 2010.

[16] H. D. Cheng, J. Shan, W. Ju, Y. Guo and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: A survey," *Pattern Recognition*, vol. 43, no. 1, p. 299-317, 2010.

[17] J. d. Campo-Ávila, N. Moreno-Vergara and M. Trella-López, "Analizying Factors to Increase the Influence of a Twitter User," in *Highlights in Practical Applications of Agents and Multiagent Systems: 9th International Conference on Practical Applications of Agents and Multiagent Systems*, Berlin, Springer Science & Business Media, 2011, pp. 69-76.

[18] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. Witten, "The WEKA Data Mining Software: An Update," Department of Computer Science, University of Waikato, Hamilton, New Zealand, 2009.

[19] Addinsoft, "XLSTAT: Your Data Analysis Solution," Addinsoft, 2016. [Online]. Available: https://www.xlstat.com/en/.