Automatic Diagnosis and Classification of Abnormalities in Digital X-ray Mammograms

Abdelali Elmoufidi [1], Khalid El Fahssi [1], Said Jai-andaloussi[1], Abderrahim Sekkaki [1], Gwenole Quellec [2], Mathieu Lamard [2,3], Guy Cazuguel [2,4]

[1]Department of Mathematics and Computer Sciences, Faculty of sciences, Hassan II University, Casablanca, Morocco.
[2]Inserm, UMR 1101, Brest, F-29200 France.
[3]Univ Bretagne Occidentale, Brest, F-29200 France.
[4]Institut Mines-Telecom, Telecom Bretagne, UEB, Dpt ITI, Brest, F-29200 France.
Abdelali.Elmoufidi09@univcasa.ma

Abstract: Mammography remains the most effective tool for the early detection of breast cancer and Computer-Aided Diagnosis (CADx) is usually used as a second opinion by the radiologists. The main objective of our study is to introduce a method to generate and select the features of suspicious lesions in mammograms and classifying them by using support vector machine, in order to build a CADx system to discriminate between malignant and benign parenchyma. Our method has been verified with the well-known Mammographic Image Analysis Society (MIAS) database and we have used the Receiver Operating Characteristics (ROC) to measure the performance of our method. The experimental results show that our method achieved an overall classification accuracy of 96.36%, with 96.77% sensitivity and 95.83% specificity in the training phase and achieved an overall classification accuracy of 94.29%, with 94.11% sensitivity and 94.44% specificity in the testing phase.

Key–Words: Mammography, Breast, Computer Aided Diagnosis, Support Vector Machine, ROC analysis.

1 Introduction

Breast cancer is one of the leading causes of mortality among women in the worldwide. Recent statistics have shown that one in ten women in Europe and one in eight in the United States develop breast cancer during their lifetime [1],[2],[3]. Early detection and diagnosis of breast cancer is the most important factors affecting the possibility of recovery from the disease. For that, the mammography represents the best and most accurate tool in detecting breast cancer [4],[5],[6],[7]. In order to improve the accuracy of interpreting mammograms, a variety of CAD systems that perform computerized mammogram analysis have been proposed. These systems are usually employed as a supplement to the radiologists' assessment. Thus, their role in modern medical practice is considered to be significant and important in the early detection and diagnosis of breast cancer. Generally, the procedure to develop a CAD system for the detection and the diagnosis of suspicious regions in mammograms takes place in two phases: The first one is a Computer-aided detection (CADe) contains two steps: 1) preprocessing step, 2) Image Analysis. And the second one is a Computer-aided diagnosis (CADx) also contains two steps: 1) Extraction and selection of features of ROIs, 2) the Classification of ROIs detected in the first phase [8],[9].

1) Pre-Processing: the purpose of this stage is to prepare the image for the next stage of operations; 2) Image Analysis: the purpose of this stage is to analyze the image and extract the necessary information; 3) Features Extraction and selection of ROIs: In this stage, we can find, match, and identify specific patterns, shapes, density and texture; 4) Classification of ROIs: The purpose of this stage is to classify the mammogram to malignant or benign [9]. In this paper, we have proposed fully automatic and robust CADx for diagnosis of suspicious lesions in a mammogram.

We have started by detecting and extracted the features of ROIs, and we have finished by classified the ROIs extracted to malignant or benign parenchyma, so the classification of mammograms to malignant or benign mammogram. The proposed algorithm is a very accurate technique for diagnosing breast cancer by using mammography images. The obtained quantitative and qualitative results demonstrate the efficiency of this method and confirm the possibility of using it in improving the CADx system. Paper organization: The setup of the paper is organized as follows: An introduction is given in section I; Section II discusses related work; Section III presents materials and method; Section IV describes our proposed research; The results and performance are presented in section V; Section VI includes a conclusion; References are given at the end.

2 Related Work

Many methods have been proposed for the diagnosis of abnormalities in mammography images. i.e, K. Ganesan, et al. [10] provided an overview about recent developments and advances in the field of Computer-Aided Diagnosis (CAD) of breast cancer using mammograms. M. Veta, et al [11] Published a review entitiled "Breast cancer histopathology image analysis" introduce the steps of image analyses. A.Jalalian, et al. [8] presented the approaches which are applied to develop CAD systems on mammography and ultrasound images. The diagnosis of regions of interest (ROIs) is a capital step in a development CAD system. Hence, a number of methods have been used to feature extraction and classification. For example, Nasseer et al. [12] developed an algorithm for Classification of Breast Masses in Digital Mammograms using Support Vector Machines. Cascio D. et al.[13] Used an approach for Mammogram Segmentation by Contour Searching and Massive Lesion Classification with Neural Network. Jacob Levman et al. [14] proposed a method entitiled "Classification of Dynamic Contrast-Enhanced Magnetic Resonance Breast Lesions by Support Vector Machines (SVM)" for classified the breast lesions using SVM. L.Jelen et al. [15] developed a method for Classification of breast cancer malignancy using cytological images of fine needle aspiration biopsies. J. Malek, et al. [16] proposed a system for Automated Breast Cancer Diagnosis Based on GVF-Snake Segmentation, Wavelet Features Extraction and Fuzzy Classification. The CAD system proved to be powerful tools that could assist medical staff in hospitals and lead to better results in diagnosing a patient.

3 Materials and Method

To develop and evaluate our proposed method we have used the Mammographic Image Analysis Society (MIAS) database [18], and Support Vector Machine (SVM) for classifying the suspicious regions to benign or malignant parenchyma.

3.1 Database

In this work, to develop and evaluate the proposed method we have used the Mammographic Image Analysis Society (MIAS) database [18]. The mammograms have a size of 1024×1024 pixels in Portable Greymap (PGM) format, and resolution of 200 micron. Each pixel in the images is represented as an 8-bit word with a pixel intensity of range [0, 255], where the images are in grayscale format. This database is composed of 322 mammograms of right and left breast, from 161 patients, where 207 mamograms diagnosed as normal and 115 mammograms as abnormal (22 images of CIRC, 19 images of SPIC, 19 images of ARCH, 15 images of ASYM, 26 images of CALC and 14 images of MISC) 52 mammograms malignant and 63 benign.

3.2 Support Vector Machine (SVM)

Support vector machine (SVM) classification algorithm, developed from the machine learning community is a discriminative classifier formally defined by a separating hyperplane. The hyperplane is determined in such a way that the distance from this hyperplane to the nearest data points on each side, called support vectors, is maximal [17]. SVM was used to diagnose breast cancer. For example, an approach with wavelet SVM was discussed in [19] .The details about SVM and its application to breast cancer diagnosis were discussed in [20],[21], which uses similar kernel.

4 Feature generation and extraction

Below a list of eighteen features selected for using as input parameters of SVM for training and testing our

system.
1) Mean Value : The mean (μ) of the pixel values in the segmented ROI represents the average of all the pixels in the segmented ROI.

$$\mu = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} I(i,j) \tag{1}$$

Where: I(i,j) is the pixel value at point (i,j) in a ROI of size MxN.
2) Standard Deviation : The standard deviation (σ) is the estimate of the mean square deviation of a grey pixel value I(i,j) from its mean value μ. It describes the dispersion within a local region, as shown in the following equation:

$$\sigma = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(I(i,j)-\mu)^2} \tag{2}$$

3) Entropy : The Entropy (H) can also be used to describe the distribution variation in a ROI. Entropy is defined as:

$$H = -\sum_{k=1}^{L-1} P_k * log_2(P_k) \tag{3}$$

Where: P_k is the probability of the k^{th} grey level, L is the total number of grey levels.
4) Skewness : The Skewness (S) characterizes the degree of asymmetry of a pixel distribution in the ROI around its mean. It is a pure number that characterizes only the shape of the distribution.

$$S = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[\frac{I(i,j)-\mu}{\sigma}]^3 \tag{4}$$

Where: I(i,j) is the pixel value at point (i,j), μ is the mean and σ is the standard deviation.
5) Kurtosis : The Kurtosis (K) measures the flatness of a distribution relative to a normal distribution. The definition of kurtosis is:

$$K = \{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[\frac{I(i,j)-\mu}{\sigma}]^4\} - 3 \tag{5}$$

6) Uniformity : The Uniformity (U) is a texture measure based on histogram and is defined as:

$$U = \sum_{k=0}^{L-1} P_k^2 \tag{6}$$

Where: P_k is the probability of the k^{th} grey level. Because the k^{th} have values in the range [0,1] and their sum equals 1, U is maximum in which all grey levels are equal, and decreases from there.
7) Sum Entropy : The Sum Entropy (SE) is calculated as a logarithmic function of the ROI in consideration.

$$SE = -\sum_{i=2}^{2N_g} p_{x+y}(i) log\{p_{x+y}(i)\}. \tag{7}$$

8) Sum Average : The Sum average (SA) is found from the ROI in consideration and the size of the gray scale

$$SA = \sum_{i=2}^{2N_g} i p_{x+y}(i) \tag{8}$$

9) Difference variance : The Difference variance (DV) is a variance measure between the ROI intensities calculated as a function of the SE calculated previously

$$DV = \sum_{i=2}^{2N_g} (i - SE)^2 p_{x-y}(i) \tag{9}$$

10) Difference entropy : The Difference Entropy (DE) is an entropy measure which provides a measure of no uniformity while taking into consideration a different measure obtained from the original image

$$DE = -\sum_{i=2}^{2N_g} p_{x-y}(i) log\{p_{x-y}(i)\}. \tag{10}$$

11) Inverse Difference Moments : The Inverse Difference Moment (IDM) is a measure of the local homogeneity.

$$IDM = \sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j). \tag{11}$$

12) Area : The area (A) is calculated as the sum of the number of all pixels (x) of segmented ROI.

$$A = \sum_{x \in ROI} 1. \tag{12}$$

13) Perimeter : The perimeter (P) is the length of a polygonal approximation of the boundary (B) of ROI:

$$P = \sum_{x \in B} 1. \tag{13}$$

14) Convexity : The Convexity C(S) is calculated as the ratio of the ROI area and its convex hull (Zunic and Rosin, 2002), the convex hull is the minimal area of the convex polygon that can contain the ROI:

$$C(S) = \frac{A}{Area(CH(S))}. \quad (14)$$

Where: S is a ROI, CH(S) is its convex hull and A is the ROI's area.
15) Compactness : The compactness (C) is a measure of ROI's shape, which indicates how much the ROI is compact, and it is defined as:

$$C = \frac{P^2}{4\pi A}. \quad (15)$$

Where : P is the ROI's perimeter, A is the area of the segmented ROI. The 4π factor is added to the denominator such that the compactness of a complete circle is 1.
16) Aspect Ratio : The Aspect Ratio (AR) corresponds to the aspect ratio of the smallest window fully enclosing the ROI in both directions (see Fig.1.), and it is defined as:

$$AR = \frac{D_y}{D_x}. \quad (16)$$

Where: Dy, Dx are the height and width of the previously mentioned window (see Fig.1.).

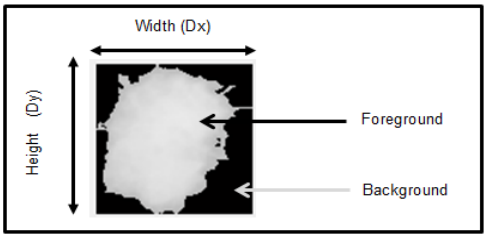

Figure 1: An example of ROI's window from which features will be extracted.

17) Area to background percentage : The Area Ratio (R_Area) is specified by dividing the area of the segmented ROI in pixels by the area of the same window given in Fig.1, which is written as:

$$R\_Area = \frac{Area\_ROI(in\ pixels)}{Area\_window(in\ pixels)}. \quad (17)$$

Where: $Area\_window$ = Dx*Dy , Dx is the width's ROI and Dy is the height's ROI. The value of R_Area will range from 0 to 1. So, It takes small values for ROI with appendices and branches emitted from it, and larger values for more compacted and rounded objects.
18) Perimeter Ratio: The Perimeter Ratio (R_Perim) presents the ratio between the perimeter of the segmented ROI to the perimeter of the same rectangular window of fig.1, this can be written as:

$$R\_Perim = \frac{Perimeter\_ROI(in\ pixels)}{Perimeter\_window(in\ pixels)}. \quad (18)$$

5 Our proposed research

In this paper, we have implemented a method for automatic diagnosis of suspicious lesions in mammograms. Our proposed method is divided into two major blocks, namely: (1) Extraction and selection of technical features for each region of interest, and (2) classification of ROIs extracted to benign or malignant parenchyma.
One among the novelties of our algorithm, that in the case of detection of multiple regions of interest, we are going to separate the ROIs detected one by one and extracted the features of each one separately, and then the diagnosing. In the end, if all ROIs belong in the same mammogram are benign, then the mammogram is benign. Otherwise, the mammogram is malignant. In addition, our algorithm is able to diagnosing the different objects in the mammogram: the masses, the calcifications and the micro-calcifications. The obtained quantitative and qualitative results demonstrate the efficiency of this method and confirm the possibility of its use in improving the computer-aided diagnosis (CADx).

5.1 Diagnosis of Regions of Interest (ROIs)

The next three figures display the details of the discussed method. 1) The button "download" is for downloading a new mammogram, 2) the button "Preprocessing" is to apply a preprocessing on original mammogram, 3) the button "Apply LBP" is to apply LBP algorithm on the image after preprocessing step, 4) the button "Extract ROIs" is to extract all objects detected as regions of interest. If we obtain just one region of interest, only the button "ROI 1" is going to enable. If we obtain two regions of interest, the two buttons "ROI 1" and "ROI 2" are going to enable, and

so on. 5) The button "ROI 1" is to extract the region of interest number one. The button "Clac-features" is to calculate the features of the region of interest selected in the previous step. 6) The button "add-feature" is to add the features in our database. 7) In the end, the button "Classify" is to classify the ROI selected to malignant or benign. if the ROI selected is malignant a red button appears on the screen containing the text malignant, if the ROI selected is benign a green button appears on the screen contains the text benign. In addition, if we obtain many regions of interest, we are going to classify them one by one and if all the ROIs are benign, the mammogram is benign. If at least one ROI is classified malignant, the mammogram is malignant.

5.1.1 Experimental results

Example 1 One suspicious lesion detected. Normally, represents a malignant lesion

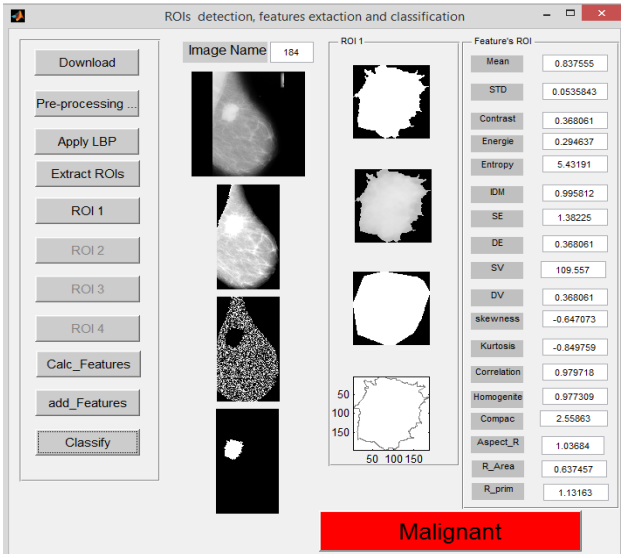

Figure 2: The mammogram correctly diagnosed as malignant.

Example 2 One suspicious lesion detected. Normally, represents a benign lesion

Example 3 Two suspicious lesion detected, which one is malignant and the second one represents a false positive.

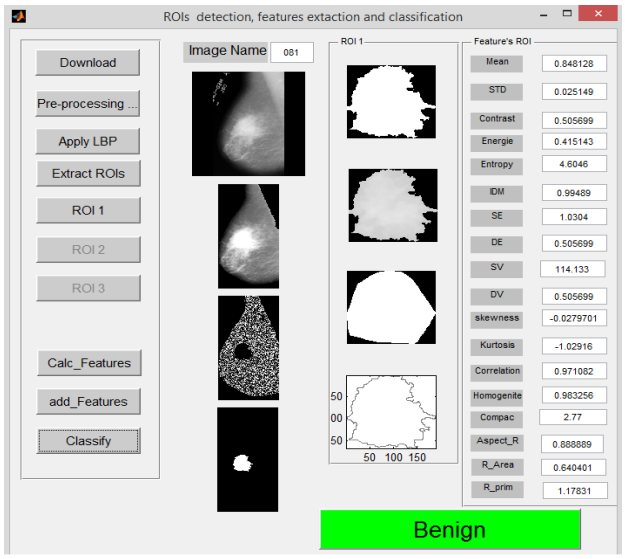

Figure 3: The mammogram correctly diagnosed as benign.

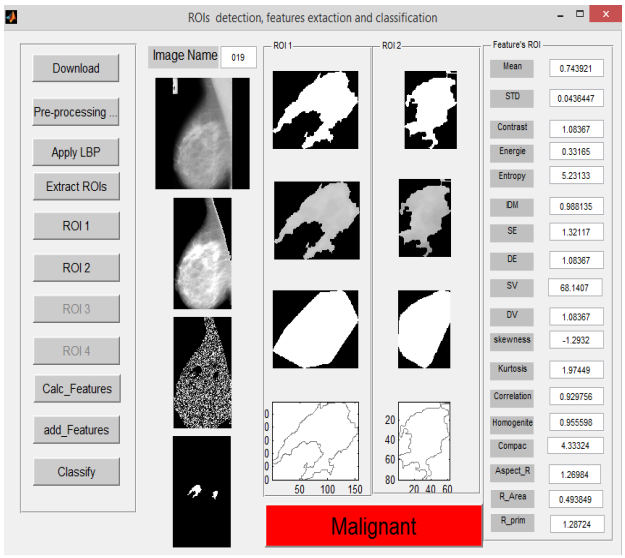

Figure 4: The mammogram correctly diagnosed as malignant

6 Results and performance

The global diagnosis method has tested on 115 images from the online available MIAS database. The detail about MIAS database is given above. The evaluated procedure is flow: the database is divided into two parts: the first one for training contains (55 images) approximately 1/2 from total images (115 images) se-

lected aleatory, the second one for testing contains the rest of database (60 images) the detail of the database distribution between training and testinig is given below:

Table 1: Number of images used to train SVM Classifier.

Image	Training	Testing	Total
Malignant	24	28	52
Benign	31	32	63
Total	55	60	115

6.1 Performance diagnosis, evaluation

In the classification case of ROIs to benign or malignant mass, a positive case means correct classification of ROIs to benign or malignant while a negative case means incorrect classification of ROIs as such a type. The definitions of the fractions are as below:

True Positive (TP) means breast classified as benign that proved to be benign; False Positive (FN) means breast classified as benign that proved to be malignant; False Negative (FP) means breast classified as benign that proved to be malignant; True Negative (TN) means breast classified as malignant that proved to be malignant.
We have tested the performance of the SVM classifier by calculating and analysis of accuracy, sensitivity and specificity for malignant and benign classification. These are defined and calculated as follows:

Accuracy: number of correct classified mass/number of total mass:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100\%. \quad (19)$$

Sensitivity: number of correct classified benign mass/number of total benign mass :

$$Sensitivity = \frac{TP}{TP + FN} * 100\%. \quad (20)$$

Specificity: number of correct classified malignant mass/number of total malignant mass:

$$Specificity = \frac{TN}{TN + FP} * 100\%. \quad (21)$$

Where: B=Benign; M=Malignant;

Table 2: Classification accuracy of Benign/Malignant

	Training		Testinig	
	B	M	B	M
Benign	(30)TP	(1)FP	(30)TP	(2) FP
Malignant	(1)FN	(23)TN	(2)FN	(26)TN

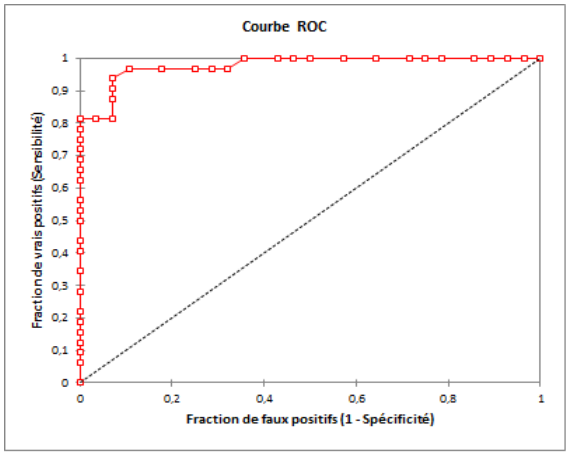

Figure 5: Plot illustrating ROC curve.

Fig.5 shows the ROC curve of our proposed diagnosis method of the testing phase. ROC analysis is based on statistical decision. The diagnosis stage achieved an overall classification accuracy of 93.33%, with 93.75% sensitivity and 92.85% specificity.

6.2 The comparison of our algorithm with existing papers.

Table 3 : Shows the performance comparison of our algorithm and the similar papers in the literature.

7 Conclusion

In this paper, an algorithm for breast mass diagnosis has been implemented under the MATLAB environment for automatic diagnosis of suspicious regions in mammogram by using SVM classifier. The performance of our algorithm has been evaluated by using Receiver Operating Characteristics (ROC). The experimental results show that our method achieved an overall classification accuracy of 96.36%, with 96.77% sensitivity and 95.83% specificity in the training phase and in the testing phase achieved an overall classification accuracy of 94.29%, with 94.11% sensitivity

Table 3: The comparison of the Performance's our method with papers published recently.

Authors	Method used	Accuracy
Veena et al.[22]	CAD Based System for Automatic Detection & Classification	92.13%
Nasseer et al.[12]	Classification of Breast Masses in Digital Mammograms Using SVM	93.069%
Ganesan et al.[17]	One-Class Classification of Mammograms Using Trace Transform Functionals	92.48 %
Our method	Automatic diagnosis & Classification of Abnormalities in Digital Mammograms	93.33%

and 94.44% specificity. The obtained results demonstrate the efficiency of this method and comparable to other methods. Our proposed algorithm can contribute to solving the main problem in mammography image processing such as the diagnosis of masses and calcifications. The efficiency of the proposed method confirms the possibility of its use in improving the CADx system.

References:

[1] Abdelali Elmoufidi Member IEEE et al., "Automatically Density Based Breast Segmentation for Mammograms by using Dynamic K-means Algorithm and Seed Based Region Growing," I2MTC 2015 - International Instrumentation and Measurement Technology Conference , PISA, ITALY, MAY 11-14, 2015.

[2] K. Hu et al., "Detection of suspicious lesions by adaptive thresholding based on multiresolution analysis in mammograms," IEEE Trans on Instrumentation and Measurement, vol. 60, no. 2, pp. 462-472, 2010.

[3] Abdelali Elmoufidi, et al., "Evaluate dynamic K-means algorithm for automatically segmented different breast regions in mammogram based on density by using seed region growing technique", Journal of Theoretical and Applied Information Technology 20th February 2015. Vol.72 No.2 ISSN: 1992-8645.

[4] A. Ferrero Fellow IEEE et al., "Uncertainty evaluation in a fuzzy classifier for microcalcifications in digital mammography," I2MTC 2010 - International Instrumentation and Measurement Technology Conference Austin, TX, 3-6 May 2010.

[5] Abdelali Elmoufidi et al.,"Detection of Regions of Interest in Mammograms by Using Local Binary Pattern, Dynamic K-Means Algorithm and Gray Level Co-occurrence Matrix," 2014 Fifth International Conference on Next Generation Networks and Services (NGNS'14) 28-30 May 2014, Casablanca, Morocco.

[6] Abdelali Elmoufidi et al, "Detection of Regions of Interest in Mammograms by Using Local Binary Pattern and Dynamic K-Means Algorithm," International Journal of Image and Video Processing: Theory and Application Vol. 1, No. 1, 30 April 2014 ISSN: 2336-0992.

[7] A. Elmoufidi Member IEEE, K. El Fahssi, S. Jai-Andaloussi, A. Sekkaki, G. Quellec, M. Lamard, G. Cazuguel., "Automatically Diagnosis of Suspicious Lesions in Mammograms", Mathematical Models and Computational Methods, ISBN: 978-1-61804-350-4.

[8] A.Jalalian et al., "Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound," Clinical Imaging, 37 2013 420426.

[9] S. Shirmohammadi and A. Ferrero, "Camera as the Instrument: The Rising Trend of Vision Based Measurement," IEEE Instrumentation and Measurement Magazine, Vol. 17, No. 3, June 2014, pp. 41-47.

[10] K. Ganesan et al., "Computer-Aided Breast Cancer Detection Using Mammograms," IEEE Reviews in biomedical engineering, vol. 6, 2013.

[11] M. Veta, et al. "Breast cancer histopathology image analysis: a review.", IEEE transactions on bio-medical engineering, vol. 61, no. 5, pp. 140011, May 2014.

[12] Nasseer M. Basheer et al., "Classification of Breast Masses in Digital Mammograms Using

Support Vector Machines," International Journal of Advanced Research in Computer Science and Software Engineering ISSN: 2277 128X, Volume 3, Issue 10, October 2013.

[13] D. Cascio et al., "Mammogram Segmentation by Contour Searching and Massive Lesion Classification with Neural Network," Institute of Electrical and Electronic Engineering (IEEE), 2006.

[14] Jacob Levman, "Classification of Dynamic Contrast-Enhanced Magnetic Resonance Breast Lesions by Support Vector Machines", IEEE Transactions On Medical Imaging, Vol. 27, No. 5, May 2008.

[15] L. Jelen et al., "Classification of breast cancer malignancy using cytological images of fine needle aspiration biopsies," int. j. appl. math. comput. sci., 2008, vol. 18, no. 1, 7583 doi: 10.2478/v10006-008-0007-x.

[16] Jihene Malek et al., "Automated Breast Cancer Diagnosis Based on GVF-Snake Segmentation, Wavelet Features Extraction and Fuzzy Classification," J Sign Process Syst. DOI: 10.1007/s11265-008-0198-2

[17] K. Ganesan et al., "One-Class Classification of Mammograms Using Trace Transform Functionals", IEEE Transactions on Instrumentation and Measurement, Vol. 63, No. 2, February 2014.

[18] J. Suckling et al., "The Mammographic Image Analysis Society digital mammogram database," Exerpta Medica, International Congress Series 1069 pp. 375-378., 1994.

[19] M. Shen et al., "A prediction approach for multichannel EEG signals modeling using local wavelet SVM", IEEE Trans. Instrum. Meas., vol. 59, no. 5, pp. 14851492, May 2010.

[20] H. X. Liu, et al., "Diagnosing Breast Cancer Based on Support Vector Machines", J. Chem. Inf. Comput. Sci. 2003, 43, 900-907.

[21] L. Wei, et al., "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications," IEEE Trans. Med. Imag., vol. 24, no. 3, pp. 371380, Mar. 2005.

[22] Veena, et al., "CAD Based System for Automatic Detection et Classification of Suspicious Lesions in Mammograms," International Journal of Emerging Trends et Technology in Computer Science (IJETTCS) ISSN 2278-6856 , Volume 3, Issue 4 July-August 2014.