Classification of Breast Cancer Detection Using Artificial Neural Networks

Sulochana Wadhwani
Department of Electrical,
Madhav Institute of Technology and Science
Gwalior, M.P. – 474005
E-mail: sulochana_wadhwani1@rediffmail.com

A.K.Wadhwani
Department of Electrical,
Madhav Institute of Technology and Science
Gwalior, M.P. – 474005
E-mail:wadhwani_arun@rediffmail.com

Monika Saraswat
Department of Electrical,
Madhav Institute of Technology and Science
Gwalior, M.P. – 474005
E-mail: monikasaraswat64@yahoo.co.in

Abstract

Breast cancer is the second most lethal cancer for women in the world today. Early detection of the breast cancer can reduce mortality rate. The breast Tumor are of two types first is Benign and second is Malignant, Benign Tumor which is non-cancerous and not life threatening and Malignant Tumor are cancerous and life threatening. Detection of breast cancer can be achieved using Digital Mammography and classification of breast tumor whether is Malignant or Benign. This paper presents a research on breast cases of 68 samples proven breast tumor are analyzed and classified into benign and malignant categories using ANN. Two different feature extraction methods used here are GLCM and Intensity based features. A supervised classifier system based on neural network is used. The performance of the each feature extraction method is evaluated on Digital mammogram images. The experimental results suggest that GLCM method outperformed the other one.

Keywords— Artificial Neural Network (ANN), GLCM, Intensity based features, Mammograms.

1. Introduction:

Breast cancer is the most frequent cancer in women worldwide. The disease is curable if detected early enough. Screening is carried out on the basis of mammograms, which use x-ray images to reveal lumps in the breast. Calcium deposits can also indicate the existence of a tumor. However, the deposits are often only a few tenths of a millimetre in size and so deeply embedded in dense tissue that they are nearly undetectable in the images. Mammogram samples with marked malignant tumor as shown in figure1.Digital mammography is proven as efficient tool to detect breast cancer before clinical symptoms appear [11]. Digital mammography is currently considered as standard procedure for breast cancer diagnosis, various artificial intelligence techniques are used for classification problems in the area of medical diagnosis. Feature extraction of image is important step in mammogram classification. These features are extracted using image processing techniques. Several types of feature extraction from digital mammograms including position feature, shape feature and texture feature etc. Textures are one of the important features used for many applications. Texture features have been widely used in mammogram classification. The texture features are ability to distinguish between abnormal and normal cases. Texture can be characterized as the space distribution of gray levels in a neighbourhood [1, 2]. Texture feature have been proven to be useful in differentiating normal and abnormal pattern. Extracted texture features provide information about textural characteristics of the image. Different classifiers are used for medical imaging application including artificial intelligence, wavelet etc. Texture measures are two types, first order and second order. In the first order, texture measure are statistics calculated from an individual pixel and do not consider pixel neighbour relationships. Intensity feature are first order texture calculation. In the second order, measures consider the relationship between neighbour pixels GLCM is a second order texture calculation [3, 4]. Texture features has been extracted and used as parameter to enhance the classification result.

2. Data Acquisition:

Images of Mammogram are available in the Department of Electrical Engineering, MITS Gwalior. These images are available with the same specification (3000x4500 pixels with 16-bit pixel depth) and the total numbers of images are 68 in which 32 are healthy cases and remaining are diseased cases.

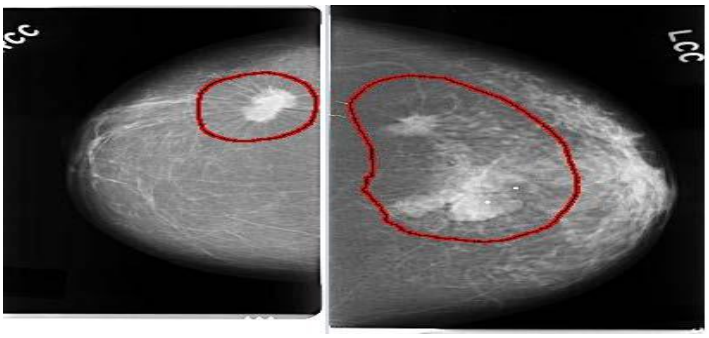

Figure: 1-Mammogram samples with marked malignant tumor

3. Methods and Material:

The use of image processing is a physical process used to convert an image signal of the breast into a physical image. This image signal is a digital signal or analog signal, and the actual output can be either an actual physical image or the characteristics of this image. Most of the methods and techniques concerned the breast cancer diagnosis. Firstly, we find the area of interest which performs the breast in the mammogram image using step by step sequences of stages and second step to extract the features of the mammogram and classify the breast tumor stage by ANN and detect weather is malignant or benign after this work apply confusion matrix and differentiate which type of texture features gives accurate result [6].

4. Feature Extractions:

Texture analysis of mammograms- The texture of a mammographic image is analyzed based on the difference between high and low gray levels in it. Various texture-related parameters of mammographic images help us to determine them as normal or abnormal [7, 8, and 12]. In this work we have used two types of texture measure, first order and second order. As shown in Figure 2.

4.1 Intensity Based Features:

Intensity based features are first order statistics depends only on individual pixel values. The pixel intensities are Figure 2 Mammogram proposed method simplest available feature useful for pattern recognition. The intensity and its variation inside the mammograms can be measured by features like mean and standard deviation using 68 samples of mammograms. Experimental features are shown in Table 1.

4.4.1Mean Value:

The mean gives the average intensity value of an image. Mammographic images that contain micro calcifications have a higher mean than those of normal images. Mean calculated from the image as per the following equation.

$$\mu = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} P(i,j)$$

Where '*i*' *i*ndicates the rows of the image, 'j' indicates the columns of the image and P (i, j) is the cell denoted by the row and the column of the image.

4.1.2 Standard Deviation:

The standard deviation is a parameter closely associated with the mean. It refers to the dispersion of values in a mammographic image around the mean. Standard deviation is given as:

$$SD=\sqrt{(mean)^2}$$

4.2 GLCM Features:

The Gray Level Co-occurrence Matrix (GLCM) texture Measurement is a method to analyze image texture [5, 6]. It is a robust method that has been developed for calculating first and second order texture features from image. The GLCM matrix is a tabulation of how often different combinations of gray levels occur in an image. It considers the relationship between two pixels at a time, the reference and neighbour pixel. Usually, the Neighbour pixel is to the right of the reference pixel. Each cell of the matrix is normalized i.e., it Contains the probability of occurrence of a pixel pair and not just the count. The texture Features are those that are calculated from the original image and do not consider relationships between neighbouring pixels. Second order features consider the relationship between groups of two pixels in an image. GLCM indicate the possible pixel values of the image matrix [9]. A particular cell of the GLCM matrix refers to the number of occurrences of a pixel pair with the values specified by the corresponding row (i) and column (j). For example, the pixel pair

(1,1) occurs once in the image. Correspondingly, the first cell of the GLCM matrix holds the value 1.Similarly, the pixel pair (1,2) occurs twice and hence the second cell holds the value 2.Each of these cells is then normalized to obtain the probability of the occurrence of a pixel pair. Once the GLCM matrix has been formed, a set of formulae based on it helps us to calculate texture features [3]. A list of the set of features calculated is explained here, AB= features of abnormal mammogram and NB= features of normal mammogram. As shown in Table 1.

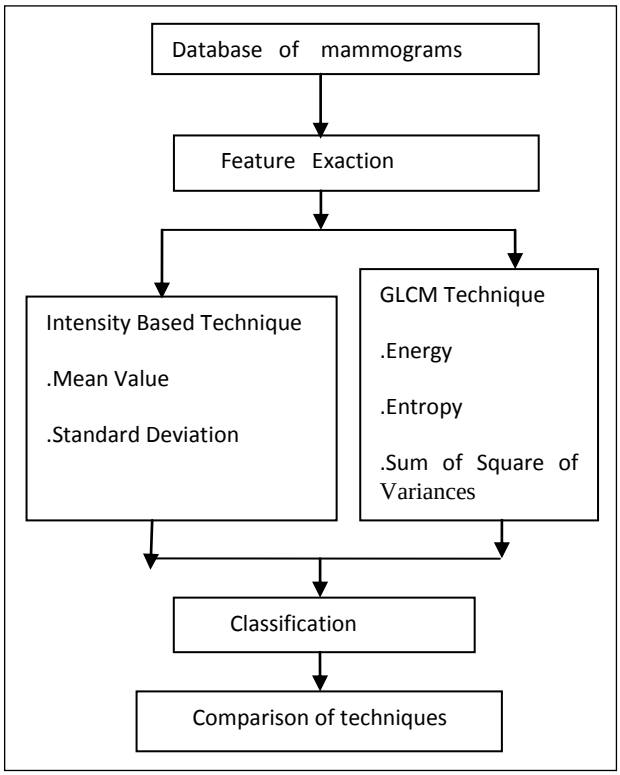

Figure: 2-Mammogram proposed method

4.2.1 Energy:

Energy represents the orderliness of a mammographic image. Energy is generally given by the mean squared value of a signal. Energy calculated from the image as per the following equation.

$$E=\sum_{i,j=0}^{n-1} P(i,j)^2$$

Where '*i*' *indicates* the rows of the GLCM matrix, indicates the columns of the GLCM matrix and P (i, j) is the cell denoted by the row and the column of the GLCM matrix.

4.2.2Entropy:

Entropy measures the amount of disorder in a mammographic image. In the case of micro calcifications, the entropy value is high. This is because the variation in intensity values in the image is high due to the presence of white calcification spots. Entropy calculated from the image as per the following equation.

$$H= -\sum_{i=0}^{n} P(i,j) \log P(i,j)$$

Where, P (i, j) is the cell denoted by the row and column of the image, entropy is also known as measure of randomness.

4.2.3 Contrast:

Contrast is a measure of the extent to which an object is distinguishable from its background. It represents the local variations present in an image, and calculates the intensity contrast between a pixel and its neighbour Contrast calculated from the image as per the following equation.

$$C=\sum_{i,j=0}^{n-1}(i.j)^2 P(i,j)$$

Where, n denotes the number of pixels in the image and P (i, j) is the cell denoted by the row and column of the image.

4.2.4 Cluster Shade:

Cluster shade measures the uniformity of a mammographic image. When a large number of pixels of similar intensity exist together, homogeneity is high and is dependent on the gray scale value of the pixels that have similar intensities. Cluster shade is calculated from the image as per the following equation.

$$C_s= -\sum_{i,j=0}^{n-1} P_{(i,j)}(log_2\ P_{i,j})$$

Where, P (i, j) is the cell denoted by the row and column of the image.

4.2.5Sum of Square Variances:

Sum of square variances is depends on the difference in gray level between adjacent pixels.

This feature puts relatively high weights on the elements that differ from the average value of P(i,j). Sum of square variances is calculated from the image as per the following equation.

$$S_v = \sum_{i,j=0}^{n-1} P_{(i,j)}(i-j)^2$$

5. Classification:

The algorithm uses a feed-forward back propagation network The schematic representation of neural network with 'n' inputs, 'm'hidden units and one output unit. The extracted features are considered as input to the neural classifier. A neural network is a set of connected input/output units in which each connection has a weight associated with it. The neural network trained by adjusting the weights so as to be able to predict the correct class. The desired output was specified as 0 for non-cancerous and 1 for cancerous. The input features are normalized between 0 and 1. The classification process is divided into the training phase and the testing phase. During training, the features are extracted from the images in which the diagnosis is known. After training is over, the trained networks are stored to be used in the algorithm. Whenever an image is taken as input in the algorithm, it is simulated with the trained net-works and goes for testing the data. The accuracy, sensitivity, specificity of the classification is depends on the efficiency of the training [10, 13].

Matlab is a good programming toolbox package of version 7.8, provides functional software environment for creating neural network. The main goal of this package is to provide users with a set of integrated tools neural networks to create models of biological and simulate them easily, without the need of extensive coding.

5.1Creation network:

The function and its own parameters below are used to create and define our neural network:

Net = newfit(inputs, targets);

The argument of this function is arranged as follows:

1. The inputs of neural network where it contains two important features based (intensity based, GLCM) extracted from the image.
2. Stated target for each stage performed by one dimensional binary array just one element of this array has value '1' and other elements are assigned to '0' that's to separate the desired target from other ones.

5.2Training and Testing Stages:

The function and its own parameters below are used to train our neural network: net = train (net, inputs, targets); the function parameters are

1. Net: the neural network which created previously.
2. Inputs: inputs of the created neural network as defined before.
3. Targets: stated target the neural network.

Table: 1- Features of GLCM and Intensity Based

FEATURE TYPE	ABI	NB1	AB2	NB2	AB3	NB3
ENTROPY	6.7281	6.1934	6.4167	6.9864	6.0263	6.5535
CONTRAST	4.79E+03	8.84E-01	1.00E+04	9.17E+04	1.74E+03	8.70E+03
CLUSTUR SHADE	0.4299	-0.0314	0.2847	-0.0394	0.1475	-0.0504
ENERGY	2.24E-05	1.51E-01	3.33E-05	1.65E-05	3.06E-05	1.75E-05
SUM OF SQUARE OF VARIENCE	0.0499	9.4709	0.0305	8.0352	0.032	6.0386
STANDARD DEVIATION	49.2184	13.361	71.4467	59.9481	60.4054	40.1557
MEAN	71.4791	56.5381	65.2625	125.5012	60.2067	103.8827

6. Measures for Performance:

A number of different measures are commonly used to evaluate the performance of the proposed method. These measures including classification, sensitivity, specificity, methews correlation coefficient (MCC) are calculated from confusion matrix using the equation 1,2,3&4. It returns a value from -1(inverse prediction) to +1(perfect prediction). The confusion matrix describes actual and predicted classes of the proposed method. Table 2 shows that actual and predicted classes of the proposed method. And Figure 3 shows that output and target classes in confusion matrix.

True Positive (TP) – counts of all samples which are correctly called by the algorithm as being cancer.

False Positive (FP) – counts of all samples which are incorrectly called by the algorithm as being cancer while they are normal.

True Negative (TN) – counts of all samples which are correctly called by the algorithm as being normal.

False Negative (FN) – count of all samples which are incorrectly called by the algorithm as being normal while they are cancer.

The performance of the classification algorithms was evaluated by computing the percentages of Sensitivity (SE), Specificity (SP), Accuracy (AC) and Mathews Correlation Coefficient (MCC), the respective definitions are as follows:

$$SE = TP/(TP+FN)*100 \quad (1)$$

$$SP = TN/(TN+TP)*100 \quad (2)$$

$$AC = (TP+TN)/(TN+TP+FN+FP)*100 \quad (3)$$

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4)$$

7. Results and Discussions:

The effectiveness of the two texture feature extraction methods are trained and testing is done using neural classifier. The dataset used for this work is composed of each method, the architecture of the neural network; training and testing samples are same. In the experiment 1, Intensity based features are extracted and its classification is done using neural classifier. In the experiment 2, GLCM features are extracted and its classification is obtained. The result shows that intensity based neural network is giving 94.11% classification rate and GLCM based neural network is giving 100% classification rate. The confusion matrix for two different feature extraction method presented in Table 3 to Table 5. The performance measures are calculated individually for the two different feature extraction methods are shown in figure 5. The proposed network was trained with all 68 tumor data set. These 68 cases are fed to the FNN with 7 input neurons, one hidden layer of 20 neurons and 4 output neurons. A Matlab software package version 7.8 is used to implement the software in the current work. When the training process is completed for the training set (68 cases), the last weight of the network were saved to be ready for the testing procedure. The testing process is done for 68 cases. These 68 cases are fed to proposed network and their output is recorded for calculation of the sensitivity, specificity, accuracy and Mathews Correlation Coefficient of prediction. The two different features set has taken one is intensity based and another is GLCM based. In the intensity based features the 68 data set which have 34 benign and 34 malignant classified by the EBP-NN with the 2 neurons. The number of data set for training, testing and validation was distributed in 40, 14 and 14 respectively. We have applied a confusion matrix on the result which have found from the EBP-NN and predict the cancerous and non-cancerous from actual data set. The confusion matrix have four categories output which are true positive (TP), false positive (FP), true negative (TN) and false negative (FN). From the confusion matrix the number of output of data sets of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) is 32,0,32 and 4 respectively. The Sensitivity, Specificity, Accuracy and MCC (Mathew's correlation co-efficient) indexes were equal to 88.8%, 100%, 94.11% and 88.8%. The other features set which is GLCM which have 34 benign and 34 malignant classified by the EBP-NN with the 5 neurons. The training, testing and validation are similar like intensity based feature. Now we have apply the confusion matrix which predicts cancerous and non- cancerous from actual number of dataset. The confusion matrix have four categories output which are true positive (TP), false positive (FP), true negative (TN) and false negative (FN). From the confusion matrix the number of output of data sets of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) is 36,0,32 and 0 respectively. The Sensitivity, Specificity, Accuracy and MCC (Mathew's correlation co-efficient) indexes were equal to 100% each respectively. After this differentiating process the accuracy of GLCM feature is best for differentiate the cancerous and non-cancerous patterns. The results are shown in Figure 4(1) this Figure shows the best validation performance at epoch 3 & in Figure4 (2) shows that best gradient and mutation at epoch 3.Table 3, Table4 and Table 5 shows that Intensity based confusion matrix, GLCM confusion matrix and Evaluation result. In Figure 5 there is comparison between both the features and plot their accuracy, sensitivity, specificity and Mathew's correlation coefficients.

Table: 2- Confusion Matrix

	Predicted	
Actual	Cancerous	Non-cancerous
Malignant	TP	FP
Benign	FN	TN

Table: 3- Intensity Based Confusion Matrix

	Predicted	
Actual	Cancerous	Non-cancerous
Malignant	32(TP)	0(FP)
Benign	4(FN)	32(TN)

Table: 4- GLCM Confusion matrix

Actual	Predicted	
	Cancerous	Non-cancerous
Malignant	36(TP)	0(FP)
Benign	0(FN)	32(TN)

Table: 5- Evaluation Results

Feature Selection	No.of cases	S E	S P	AC (%)	MCC
Intensity Based	68	0.88	01	94.11%	0.88
GLCM	68	01	01	100%	01

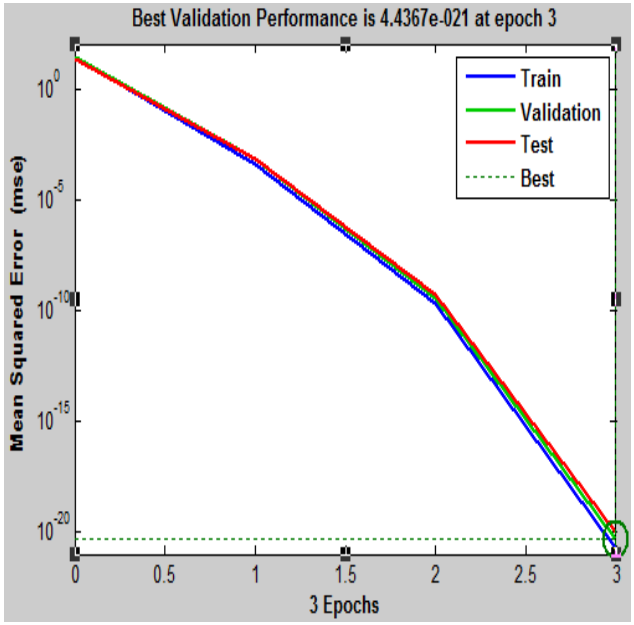

Figure: 4(1)-Results from ANN

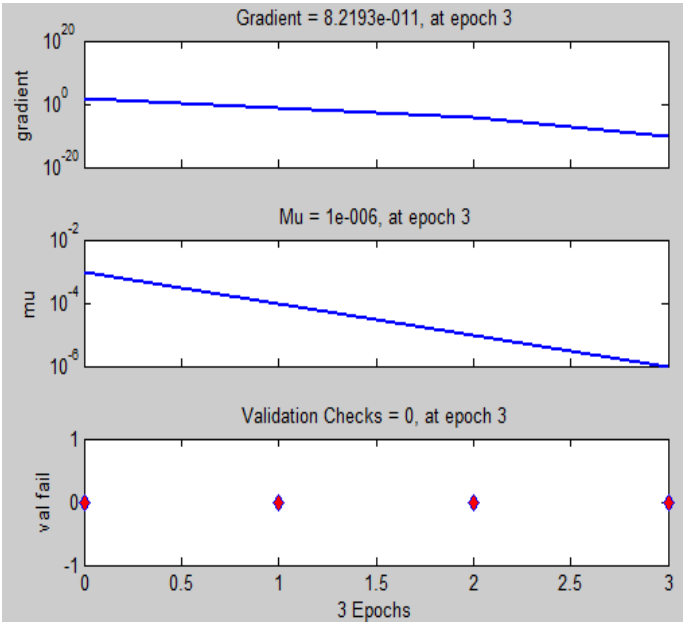

Figure: 4(2)-Results from ANN

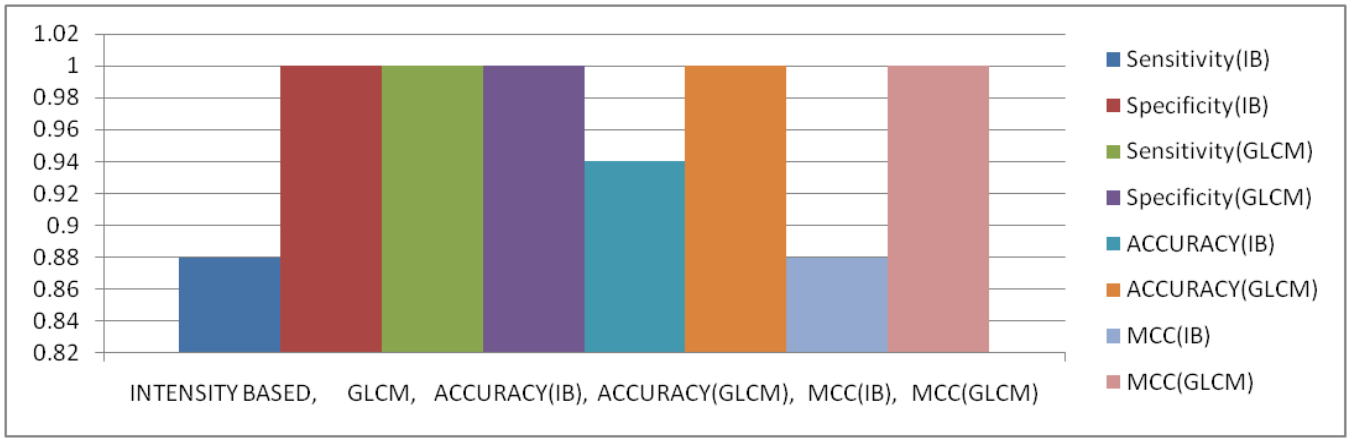

Figure: 5- Performance measure comparisons

8. Conclusion:

We can draw the following conclusions based on the above experimental results. Breast cancer is one of the major causes of death among women. So early diagnosis through regular screening and timely treatment has been shown to prevent cancer. In this paper we have presented a novel approach to identify the presence of breast cancer mass and calcification in mammograms and extracted clinically features and after this experiments we use ANN soft computing method for detect the cancer and easily differentiate the benign and malignant. This will help doctor to take or analyze in which stage of cancer the patient have and according to which he/she can take necessary and appropriate treatment steps. This proposed method is low cost as it can be implemented in general computer .This paper is based on visual detection method of the processed mammogram images. In future aspect a real-time system can be implemented using suitable data acquisition software and hardware interface with digital mammography systems.

9. Acknowledgement:

Authors are thankful to UGC NEW DEHI for providing financial assistance to this research paper under UGC major research project. Research grant sanctioned from UGC vide F-NO.37-505/2009(SR).dated 21.12.2009.

References:

[1] Pradeep N, Girlish H & et.al./"Feature extraction of mammogram", International Journal of Bioinformatics research 2012, vol.4 pp-214- 244.

[2] Shekhar Singh, Dr P R Gupta/"breast cancer detection & Classification by NN"vol no.6.

[3] R.Nithya,B.Santhi/"comparative study of feature extraction, method for breast cancer classification"/Journal of theoretical and applied Information Technology, 30.11.2011, vol.33 no.2.pp 220-224.

[4] Tabar L. and Dean P.B.(2003)/"Composition &methods for detection and treatment of breast cancer".GynaecolObstet,82, pp319- 326.

[5] M.Vasantha et.al./"Medical Image feature extraction selection and classification"/International Journal of engineering science and technology vol.2 (6), 2010,pp. 2071-2076.

[6] A.MohdKhuzi, R.Besar & et.al.\"Identification of masses in digital mammogram using gray level co-occurrences matrices"/Biomedical Imaging and Intervention Journal, 2009.

[7] Kulkarni D, Bhagyashree S and Udupi G(2010) Texture Analysis of mammographic images.int.j.com.appl.5,12_17.

[8] Dubey R, Hanumandla M & Gupta S(2010) level set detected masses in digital ammograms. Indian.j.sci.Technol.3(1):9-13.

[9] Singh N and Mohapatra A (2011)Breast cancer mass detection inmammograms using K-means clustering .int.j.com.appl.22(2),15-21.

[10] Usha Rani K(2010) parallel approach for diagnosis of breast cancer using NN, int.j.com.appl.(09758887),10(3), 1-5.

[11] Ms Tripty Singh, Dr. Sarita Singh Bhadauoria, Dr A.K Wadwani and Dr S.Wadhwani, "Contrast Enhancement of Clusters In Images Using Fuzzy-Rule Based Algorithm", International Journal of Advances in Scienceand Technology, Feb Issue Vol. 2, No. 2, pp. 18-28, 2011.

[12] Ms Tripty Singh, Dr. Sarita Singh Bhadauoria, Dr. A.K Wadwani and Dr. S.Wadhwani, "Wavelet Based Contrast Enhancement and Adaptive Noise Cancellation in Mammograms", International Journal of Science and Advances Technology, Vol. 1, Issue. 9, ,pp. 256-263, Nov 2011.

[13] Ms Tripty Singh, Dr. Sarita Singh Bhadauoria, Dr. A.K Wadwani and Dr. S.Wadhwani, "Computer Aided Diagnosis System for Breast Cancer", International Journal of Scientific and Research Publications, Vol. 1, Issue. 2, , pp. 1-5 January-2012.

Dr. Sulochana Wadhwani is a Associate Professor at Madhav Institute of Technology & Science Gwalior. Dr. Wadhwani has obtained her Ph.D. in Application of DSP in condition monitoring of Electrical Machines from IIT Roorkee in 2007, M.E Electrical (Control system) from GEC Jabalpur in 1994 and B.E Electrical from GEC Jabalpur in 1991. She has total 18 years of teaching & research experience. she has published more than 80 research papers in various International / National Journals/conferences. 03 Ph.D. thesis are submitted under her supervision and 8 are under progressed. Wadhwani has guided number of M.E dissertations and UG projects. Her research areas of interest are Application of Artificial Neural Network, Fuzzy Logic and Wavelets in field of Biomedical Engineering, Condition Monitoring of Electrical Machines & signal processing. Dr.Wadhwani is supervising number of research projects sponsored by various funding agencies such as AICTE, UGC etc. She has organized 04 national conferences and 02 staff development programmes sponsored by AICTE. She is a life member professional bodies like ISTE, IETE and IE (I). Dr.Wadhwani is well known for her contributions in Bio-medical Engineering, condition monitoring of electrical machines, Measurement & Instrumentation and signal processing.

Dr. Arun Kumar Wadhwani is a Professor at Madhav Institute of Technology & Science Gwalior. Dr. Wadhwani has obtained his Ph.D. in Application of DSP in Biomedical Engineering from IIT Roorkee in 2003, M.E Electrical (Measurement & Instrumentation) from University of Roorkee in 1993 and B.E Electrical from Bhopal University in 1987. He has total 25 years of teaching & research experience. He has published more than 125 research papers in various International / National Journals/conferences. 04 Ph.D. thesis are awarded under his supervision and 11 are under progressed. Wadhwani has guided number of M.E dissertations and UG projects. His research areas of interest are Application of Artificial Neural Network, Fuzzy Logic and Wavelets in field of Biomedical Engineering.Dr.Wadhwani is supervising number of research projects sponsored by various funding agencies such as AICTE, UGC etc. He has organized 04 national conferences and 02 staff development programmes sponsored by AICTE.He is a life member professional bodies like ISTE, IETE and IE(I). Dr. Wadhwani is well known for his contributions in Bio-medical Engineering, Measurement & Instrumentation and signal processing

Monika Saraswat student of Electrical Engineering Dept of MITS Gwalior. She is persuing her M.E in Measurement and Control from RGTU university Bhopal .Her B.Tech is in Electrical Engineering from MPCT, Gwalior. Her area of interest are Artificial Intelligence & Image Processing, She has 1years of teaching experience. She has done one International conference and two National conferences.