

Breast Cancer Detection and Classification using Artificial Neural Network with Particle Swarm Optimization

Mrs.K.Preetha^{#1}, Dr.S.K.Jayanthi^{*2}

*Department of Computer Science, Vellalar College for Women,
Erode, Tamil Nadu, India*

¹kpreetha.msc@gmail.com

²jayanthiskp@gmail.com

Abstract— Breast cancer is among the most common causes of deaths today, coming fifth after lung, stomach, liver and colon cancers. Primary prevention in the early stages of cancer becomes more complex. The early diagnosis of breast cancer is accomplished by X-ray mammography. It is the main test used for screening and its analysis and processing are the keys to improve breast cancer prognosis. It detects around 80% to 90% of breast cancer. As a result, a large number of images need to be examined by limited number of radiologists, resulting in misdiagnosis due to human errors by visual fatigue. This paper focuses on detection of breast cancer classification using Artificial Neural Network whose weights are optimized by Particle Swarm Optimization. Segmentation is used to identify the suspicious region from the mammogram image. Intensity, Texture and Shape features are extracted from the segmented mammogram image. The feature extracted from the known segmented images are used to train using ANN-PSO and tested for the detection of breast cancer, classifies as normal or abnormal. This helps the radiologists to increase the accuracy of mammogram examination to diagnose the cancer.

Keywords— Breast Cancer, Mammogram, Neural Network, Particle Swarm Optimization.

I. INTRODUCTION

Cancer is the type of diseases that causes the cells of the body to change its characteristics and cause abnormal growth of cells. Most types of cancer cells eventually become a mass called tumor. Breast cancer is a major cause of death in women. Most breast cancer cases occur in women aged 40 and above but certain women with high-risk characteristics, often hereditary, may develop breast cancer at a younger age. Several imaging techniques are available for detecting breast cancer such as ultrasound imaging, MRI imaging and digital Mammography. Breast tumors and masses usually appear in the form of dense regions in mammograms. A typical benign mass has a round, smooth and well circumscribed boundary; on the other hand, a malignant tumor usually has a speculated, rough, and blurry boundary. If the cancer can be detected early, the options of treatment and the chances of total recovery will increase. The objective of this paper is to increase the effectiveness and efficiency of the classification process in order to reduce the number of false-positive of malignancies.

This paper is organized as follows: Section 2 describes the literature review. Section 3 describes the

proposed methodology. Section 4 describes the experimental result analysis and discussion and section 5 describes the Performance Evaluation. Finally, Section 6 describes conclusion with future work.

II. LITERATURE REVIEW

Currently, breast cancer detection is a challenging issue for women. Breast cancer is curable, if it is detected in an initial stage. Number of researchers has tried to arrive an exact solution for this work by proposing different classification techniques.

Abdulla and Zaki [8] proposed a method for detection of masses in digital mammogram using ANN and GLCM features extraction, and achieved 91% sensitivity and 84% specificity for classifying 90 mammogram images randomly selected from the Mini-Mias database. Islam et al. [9] also proposed a classification method using ANN and GLCM features to classify benign-malignant classes of mammogram images which achieved 90% sensitivity and 84% specificity.

A comparative study on GLCM feature extraction for breast cancer classification by R.Nithya and B.Santhi [2], the study used a sample of 50 mammogram images from the Digital Database for Screening Mammography (DDSM) Database shows an excellent result. R.Nithya and B.Santhi [10] proposed a method for Classification of normal and abnormal patterns in Digital mammograms for the breast cancer diagnosis using ANN and GLCM features. The work shows that the sensitivity and specificity more than 90% for a sample set of 50 digital mammogram images from the DDSM Database.

The overall literature survey says that various methods and classification techniques are applied for classifying the images into normal or abnormal. The existing methods are tested with only limited number of mammogram images from the MIAS database.

III. PROPOSED METHODOLOGY

The proposed system consists of four phases for breast cancer detection from the mammogram image. The anticipated research uses the data set obtained from Mammographic Image Analysis Society (MIAS) [7]. The set consists of 322 images that fall into one of the following classes: 67 benign, 54 malignant and 201 Normal images. The overview of proposed methodology is depicted in Fig.1.

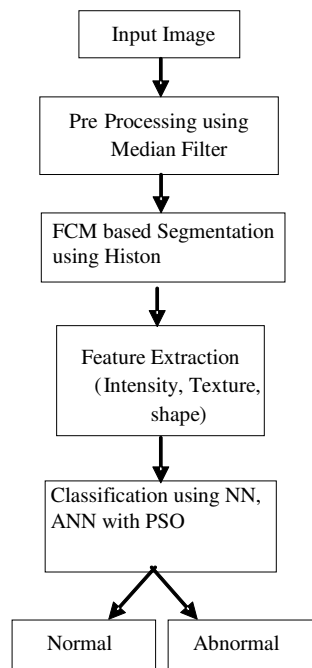


Fig. 1 Diagrammatic representation of proposed method

A. Pre-Processing

The input image which is obtained is preprocessed to remove the noise and to enhance the quality of the image. Before any image-processing algorithm applied on mammogram, preprocessing steps are very important in order to limit the search for abnormalities without undue influence from background of the mammogram. Digital mammograms are medical images that are difficult to be interpreted, thus a Pre-Processing phase is needed in order to improve the image quality and make the segmentation results more accurate. The mammogram images are preprocessed by using median filter. Median filtering is useful for reducing speckle noise that can do a better job of preserving edges than simple smoothing filters.

B. Segmentation

From the enhanced image, the suspicious region is identified using Histon based Fuzzy c-means segmentation algorithm. Segmentation is an important step in many medical imaging applications and a variety of image segmentation techniques already exist. Segmentation is the process of partitioning an image into multiple segments, so as to change the representation of an image into something that is more meaningful and easier to analyze. *Mohabey and Ray[11]* introduced the concept of histon means for the visualization of color information for the evaluation of similar color regions in an image. Histon is the segregation of the elements at the boundary, which can be applied in the process of the image segmentation. Histon based Fuzzy c-means segmentation find the cancer detected regions in the image and achieve the better segmentation results.

1). *Fuzzy C-Means Algorithm*: FCM is a method of clustering which allows one piece of data to belong to two or

more clusters (i.e.) it allows the pixels to belong to multiple classes with varying degrees of membership. It is based on minimization of the objective function defined as follows.

$$J = \sum_{j=1}^N \sum_{i=1}^c U_{ij}^m \|x_j - v_i\|^2 \quad (2)$$

where U_{ij} represents the membership of pixel x_j in the i th cluster, v_i is the i th cluster center, and m is a constant. The parameter m controls the fuzziness of the resulting partition, and $m=2$ are used in this study.

The FCM algorithm assigns pixels to each category by using fuzzy memberships. Let $X = \{x_1, x_2, x_3, \dots, x_n\}$ which denotes an image with N pixels (set of data points) to be partitioned into c cluster centers or centroids and $V = \{v_1, v_2, v_3, \dots, v_c\}$ be the set of centers.

Step 1: Randomly select 'c' cluster centers.

Step 2: Calculate the fuzzy membership 'U_{ij}'. The cost function is minimized when pixels close to the centroid of their clusters are assigned high membership values, and low membership values are assigned to pixels with data far from the centroid. The membership function represents the probability that a pixel belongs to a specific cluster. In the FCM algorithm, the probability is dependent solely on the distance between the pixel and each individual cluster center in the feature domain. The membership functions and cluster centers are updated by the following:

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{\|x_j - v_i\|}{\|x_j - v_k\|} \right)^{\frac{2}{m-1}}} \quad (3)$$

Step 3: Compute the cluster centers 'v_j' using:

$$v_i = \frac{\sum_{j=1}^N u_{ij}^m x_j}{\sum_{j=1}^N u_{ij}^m} \quad (4)$$

Step 4: Repeat step (2) & (3) until the minimum 'J' value is achieved.

The cluster center or centroids is calculated by using K-Means clustering algorithm, the FCM converges to a solution for v_i representing the local minimum or a saddle point of the cost function. Convergence can be detected by comparing the changes in the membership function or the cluster center at two successive iteration steps. The suspicious regions will be identified from this method.

2). Histon Process

Histon is basically a contour plotted on the top of the histograms by considering a similar color sphere of predefined radius around a pixel. The base histogram is considered to be the lower approximation and the histon as upper approximation. The upper approximation is a collection of all points, which may or may not belong to one segment but certainly share a unique property that the elements have similar colors. For segmentation, only the upper approximation is considered and the histogram based technique is applied on the histon to find the regions in the image. For every intensity value in histogram, the number of pixels encapsulated in the similar intensity sphere is evaluated. This count is then added to the value of the histogram at that

particular intensity value. This computation is carried out for all the intensity values that lead to the formation of histogram.

The steps of Histogram process are given by:

- Using the difference between the nearest neighbors, calculate the pixel value.
- The formula to find the difference between the nearest neighbors for all the pixels values as follows

Difference of a particular pixel = Nearest neighbor 1 - Nearest neighbor 2

- After finding the pixels values update the pixels values update the pixels values by setting a threshold. Here we used the threshold value as greater than 1 or less than 1.
- Then to find the intensity values of a pixel, we find the difference between the nearest neighbors for the particular pixel values. Using the above formula we find the intensity values of a pixel.
- After finding the intensity values of a pixel, we have to update the values in the image by setting the threshold. The threshold value is greater than or less than 1
- Update the count values in the particular intensity value of a pixel and we check one by one via histogram and plot the values.
- Here, with the difference between the neighbors the intensity values of a pixel is calculated. By keeping threshold, the intensity value of a pixel is calculated. (i.e) if the calculated difference between the neighbors is greater than one means replace the pixel value with 2 and if the calculated difference between the neighbors is less than one means replace the pixel value with 0

The above process repeats until the eligibility criterion occurs.

By using the above process, we compute and rearrange the pixel values to find the roughness areas without affecting the lower dimensional areas. Finally we consider upper approximation for further processing. These values are given to k-means for clustering the medical images.

3). Centroid Selection using k-means algorithm

K-means clustering is to minimize the sum of squared distances between data and the corresponding centroid of the cluster. Here, K-means clustering groups the pixels into two distinct clusters (k=2) based on the values of Histogram. Centroid is calculated in this approach using k-mean clustering algorithm. The step-by-step of K-means clustering algorithm is presented as follows:

Let $X = \{x_1, x_2, x_3 \dots x_n\}$ be the set of data points and $V = \{v_1, v_2, v_3 \dots v_c\}$ be the set of centers.

Step 1: Initialize the cluster centers by Histogram process.

Step 2: Calculate the distance between each data point and cluster centers.

Step 3: Assign the data point to the cluster centre whose distance from the cluster centre is minimum of all the cluster centres.

Step 4: Recalculate the new cluster centre using

$$v_i = \left(\frac{1}{c_i}\right) \sum_{j=1}^{c_i} x_j \quad (1)$$

Where 'Ci' represents the number of data points in ith cluster.

Step 5: Recalculate the distance between each data point and newly obtained cluster centres.

Step 6: If no data point was reassigned then stop, otherwise repeat from step 3.

After calculating the centroid values, it will assign to the FCM for better clustering on medical images.

4). FCM based segmentation using Histogram

The step by step procedures are given by:

Step 1: Roughness areas are computed and initialize the cluster centers using Histogram process.

Step 2: Centroid selection using k-means algorithm.

Step 3: FCM assigns the cluster centers from the above step for better clustering on medical images.

The suspicious regions will be identified from this method.

C. Feature Extraction

After the segmentation is performed, the segmented breast image is used for feature extraction. A feature is significant piece of information extracted from an image which provides the more detailed understanding of an image. Several types of features like Intensity, Texture and shape features are extracted. Shape measurements are physical dimensional measures that characterize the appearance of an object.

1). Intensity Feature

Mean reveals the general brightness of an image. Bright image should have high mean while dark image should have low mean, and also mean values characterize individual calcifications. Standard deviation or variance reveals the contrast of an image. Image with good contrast should have high variance. Standard Deviations (SD) also characterize the distribution of the gray level. Image with bimodal histogram distribution (object in contrast background) should have high variance but low skew distribution (one peak at each side of mean). Energy measurement is closely related to skew. Highly skew distribution usually gives high-energy measurement. Entropy measures the average number of bits to code each gray level. It has inverse relationship with skew and energy measurement. Highly skew distribution tends to yield low Entropy. These are summarized in Table 1. Then features are calculated for classification.

2). Texture LBP

Texture features is useful in differentiating normal and abnormal cells. The texture feature extraction is calculated by using Local Binary Pattern Operator. LBP operator combines the characteristics of statistical and structural texture analysis.

TABLE I. INTENSITY FEATURE

Intensity	Formula
Mean	$\mu = \sum_{x=1}^{N_g} x \cdot p(x)$
Variance	$\sigma^2 = \sum_{x=1}^{N_g} (x - \mu)^2 \cdot p(x)$
Skewness	$\mu_3 = \sigma^{-3} \sum_{x=1}^{N_g} (x - \mu)^3 \cdot p(x)$
Kurtosis	$\mu_4 = \sigma^{-4} \sum_{x=1}^{N_g} (x - \mu)^4 \cdot p(x) - 3$
Entropy	$f_1 = \sum_{x=2}^{2N_g} p_{i+j}(x) \log\{p_{i+j}(x)\}$

The LBP operator is used to perform gray scale invariant two-dimensional texture analysis. The LBP operator labels the pixel of an image by thresholding the neighborhood (i.e. 3×3) of each pixel with the center value and considering the result of this thresholding as a binary number. When all the pixels have been labeled with the corresponding LBP codes, histogram of the labels are computed and used as a texture descriptor.

Algorithm

Step 1: Given a pixel in the image LBP can be computed by comparing it with its neighbors

$$LBP(x_c, y_c) = \sum_{n=0}^p 2^n C(I_n, I_c) \tag{5}$$

Where, $C(I_n, I_c) = \begin{cases} 1; & \text{if } I_n > I_c \\ 0; & \text{else} \end{cases}$. I_n is the gray level

value of the central pixel, I_c is the value of its neighbors. P is the number of involved neighbors.

Step 2: After LBP pattern of each pixel is identified, a histogram is used to represent the texture image

$$H_i = \sum_{x,y} I \{ f_i(x,y) = i \}, i = 0, \dots, n - 1, \tag{6}$$

Where n is the number of different labels produced by the LBP operator, and $I\{A\}$ is 1 if A is true and 0 if A is false.

Step 3: When the image patches whose histograms are to be compared have different sizes, the histograms must be normalized to get a coherent description

$$N_i = \frac{H_i}{\sum_{j=0}^{n-1} H_j} \tag{7}$$

3). Shape Features

Shape values can be used to distinguish between benign and malignant tumors. Benign lesions usually have smooth shapes and so they produce a regular shape, whereas malignant lesions present irregular shapes. The features are Eccentricity, Orientation, Solidity and Extent. The eccentricity of the image gives a measure of just how 'squashed' it is. The eccentricity is obtained using the ellipse that has the same second-moments as a region. It is the ration of the distance between the foci of the ellipse and its major axis length. The value is between 0 and 1. Solidity is the measurement of the overall concavity of a particle. Solidity is defined as the image area divided by the convex hull area. Thus a particle becomes more solid. Solidity of a convex shape is always one and it is lower for an object with rough perimeter or having holes in it. Orientation is used to find an object in an image. Extent of an image shows that how far the cancer region may have spread. It is the proportion of the pixels in the bounding box (the smallest rectangle containing the region) that are also in the region. These are summarized in Table II

TABLE II. SHAPE MEASUREMENTS

Shape Measurements	Formula
Eccentricity	$\frac{c}{a}$ Where ,c is the distance from the center to a focus. a is the distance from that focus to a vertex
Solidity	$\frac{Total Area}{Convex Area}$
Orientation	Scalar; the angle (in degrees ranging from -90 to 90 degrees) between the x-axis and the major axis of the ellipse that has the same second-moments as the region.
Extent	$\frac{Total Area}{Area of the bounding box}$

D. Classification

Classification is the final step in mammogram abnormality detection. The extracted features are considered as input to the classifier to classify the detected suspicious areas into normal, benign or malignant. Classifier such as artificial neural network (ANN) has performed well. The classification of breast cancer detection is divided into the training phase and the testing phase. During training, the features are extracted from the segmented images are input to ANN whose weights are optimized by particle swarm optimization in which the diagnosis is known. Whenever an image is taken as input to the algorithm, it is simulated with the trained networks and goes for testing the image.

1). Artificial Neural Network

An Artificial Neural Network (ANN) is inspired by the way biological nervous systems process information. An ANN consists of a collection of processing elements that are highly

interconnected and transform a set of inputs to a set of desired outputs in which each connection has a weight associated with it. The advantage of ANN is their capability of self-learning, and often suitable to solve the problems that are too complex to use the conventional techniques, or hard to find algorithmic solutions. The neural network trained by adjusting the weights so as to be able to predict the correct class. In this work, a method that combines ANN with PSO algorithm was proposed to optimize the weights of target value.

Algorithm

- Step 1: Extract features from mammogram images.
- Step 2: Create input and target for normal and abnormal class.
- Step 3: The initial weights are chosen randomly.
- Step 4: Calculate the predicted output.

An issue in neural network is difficult to train: the training outcome can be nondeterministic and depend crucially on the choice of initial parameters. To address the issue hybrid approach of neural network with particle swarm optimization learning algorithm for cancer prediction.

2). Neural Network with Particle Swarm Optimization

Algorithm

- Step 1: Extract features from mammogram images.
- Step 2: Create input and target for normal and abnormal class.
- Step 3: The weights are optimized by Particle Swarm Optimization algorithm.
- Step 4: Output node predicts the correct class (i.e.) Normal or abnormal

Using the above algorithm the weightage can be optimized on target values for the efficient breast cancer detection using PSO. The result shows that one can easily and effectively detect breast cancer at an early stage.

IV. EXPERIMENTAL RESULT ANALYSIS AND DISCUSSION

The proposed system for medical image segmentation for breast cancer detection was evaluated. It is implemented in the working platform of MATLAB. The input image is pre-processed to remove the noise and to enhance the quality of the image using Median filter. From the enhanced image, the suspicious region is identified using FCM segmentation. Segmented image is used for feature extraction. The features are Intensity feature like Mean, Variance, Skewness, Kurtosis and Entropy, The texture feature extraction using local binary pattern and shape measurements like Ecc (Eccentricity), Ori (Orientation), Sol (solidity) and Ext (Extent). The extracted features values are passed to train the Neural Network to classify whether region is normal or abnormal using PSO algorithm. Thus one can able to find out whether it is affected by cancer or not.

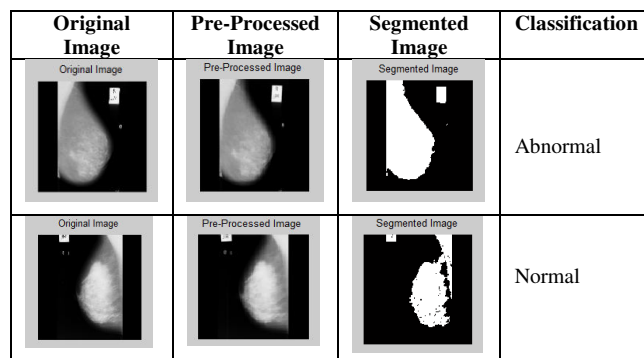


Fig 2. Experimental Result

V. PERFORMANCE EVALUATION

The various assessment metrics are used to calculate and analyze that our proposed technique is the efficient mammogram image segmentation and classification of breast cancer. The metric values like Sensitivity (SE), Specificity (SP) and Accuracy (AC) are used to evaluate the performance of the classifier. The formulas are given in Table II. Sensitivity is a proportion of positive cases that are well detected by the test and the specificity is proportion of negative cases that are well detected by the test. Classification accuracy depends on the number of samples correctly classified.

TABLE III. FORMULA FOR MEASURES

Measures	Formula
Sensitivity	$SE = TP / (TP + FN)$
Specificity	$SP = TN / (TN + FP)$
Accuracy	$AC = (TP + TN) / (TP + FP + TN + FN)$

Where, TP is the number of true positives; FP is the number of false positives; TN is the number of true negatives; FN is the number of false negatives. Confusion matrix is shown in Table IV.

TABLE IV. CONFUSION MATRIX

Actual	Predicted	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

- TP - Predicts abnormal as abnormal.
- FP - Predicts abnormal as normal.
- TN - Predicts normal as normal.
- FN - Predicts normal as abnormal.

TABLE V. CONFUSION MATRIX FOR TESTING ANN-PSO

Method	Actual	Predicted	
		Cancer(Positive)	Normal(Negative)
ANN-PSO	Cancer(Positive)	41(TP)	7(FP)
	Normal(Negative)	6(FN)	246(TN)

TABLE VI. CONFUSION MATRIX FOR TESTING ANN

Method	Actual	Predicted	
		Cancer(Positive)	Normal(Negative)
ANN	Cancer(Positive)	32(TP)	16(FP)
	Normal(Negative)	13(FN)	239(TN)

TABLE VII. PERFORMANCE COMPARISON BETWEEN ANN AND ANN_PSO

Test Image	Methods	Sensitivity	Specificity	Accuracy
48 Malign 252 Normal	ANN-PSO	87.23%	97.23%	95.66%
	ANN	72.72%	93.6%	88.66%

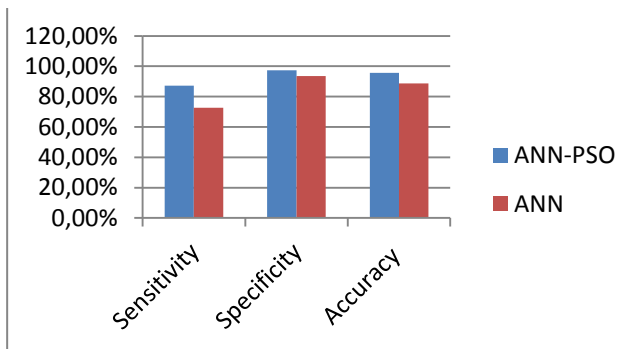


Fig. 1. Comparison between ANN and ANN with PSO

To evaluate this work, the proposed method trained with 322 mammogram images and tested with 300 mammograms (48 malignant and 252 normal) images. Confusion Matrix for testing ANN-PSO and ANN is shown in Table IV and Table V. The result in Table VII shows the classification effectiveness of normal and abnormal mammogram images with the help of extracted features. Figure 3 show the computed sensitivity, Specificity and accuracy for testing data of proposed and existing method. ANN-PSO reveals a better classification rate in sensitivity, specificity and accuracy.

VI. CONCLUSION AND FUTURE SCOPE

Breast cancer is one of the major causes of death among women. Breast cancer is curable when detected in early stages. The classification of mammogram images is emphasized in this paper for cancer diagnosis using Artificial Neural Network with PSO algorithm. It provides a faster diagnosis of breast cancer into normal or abnormal with accurate results. The result indicates that this system can facilitate the doctor to detect the breast cancer in the early stage of diagnosis as well as identify the suspicious region. The future work is to stage

the breast cancer in the abnormal mammogram images. It will help the doctor and radiologists to analyze the stage of cancer, the patient is in and according to which he/she can take necessary and appropriate treatment steps.

REFERENCES

- [1] A. Velayudham, R. Kanthavel, K. Madhan Kumar, "A Hybrid and Novel Optimization Framework for Denoising and Classification of Medical Images using DTCWP and Neuro-Fuzzy Classifiers", *Journal of Theoretical and Applied Information Technology*, Vol. 65, No.1, 10th July 2014.
- [2] R. Nithya, B. Santhi, "Comparitive Study on Feature Extraction Method for Breast Cancer Classification", *Journal of Theoretical and Applied Information Technology*, Vol. 33, No.2, 30th November 2011.
- [3] Anupa Maria Sabu, D. Narain Ponraj, Dr. Poongodi, "Textural Features Based Breast Cancer Detection: A Survey", *Journal of Emerging Trends in Computing and Information Sciences*, Vol. 3, No.9, September 2012.
- [4] Naveen Chandra Yadav and Prafull Gajbhiye, "Diagnosis of Breast Cancer using Neural Network Approach", *BMR Journals*, Vol.1, Issue. 1, pp. 1-7, 2014.
- [5] Seema Singh, Sushmitha H, Harini J and Surabhi B.R, "An Efficient Neural Network Based System for Diagnosis of Breast Cancer", *International Journal of Computer Science and Information Technologies*, Vol. 5(3), 2014.
- [6] Nalini Singh, Ambarish G Mohapatra, Biranchi Narayan Rath, and Guru Kalyan Kanungo, "GUI Based Automatic Breast Cancer Mass and Calcification Detection in Mammogram Images using K-means and Fuzzy C-means Methods", *International Journal of Machine Learning and Computing*, Vol. 2, No. 1, February 2012.
- [7] Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I., Boggis, C., et al., "The mammographic images analysis society digital mammogram database", *Excerpta Medical International Congress Series*, vol. 1069, 1994, pp. 375-378. URL: <http://peipa.essex.ac.uk/ipa/pix/mias/>
- [8] Al Mutaz M. Abdalla, Safaai Dress, Nazar Zaki, "Detection of masses in Digital Mammogram using Second Order Statistics and Artificial Neural Network", *International Journal of Computer Science and Technology(IJCSIT)*, Vol. 3 No.3 June 2011.
- [9] Mohammed J.Islam, Majid Ahmadi and Maher A. Sid-Ahmed "An Efficient Automatic Mass Classification Method in Digitized Mammograms using Artificial Neural Network", *International Journal of Artificial Intelligence & Applications (IJAA)*, Vol.1, No.3, July 2010.
- [10] R.Nithya and B.Santhi, "Classification of normal and abnormal patterns in digital mammograms for diagnosis of breast cancer", *International Journal of computer Applications*, vol 28 No.6 August 2011.
- [11] M.Y.Bhanumurthy and Koteswararao Anne "Multiple Sclerosis segmentation in Brain MR Images using Modified Histon based Fast Fuzzy C-Means" *International Journal of Applied Engineering Research ISSN 0973-4562 Vol.10 No.16 September 2015*.
- [12] Mr. Athar Hussain and Prof. Anjali B.Raut " A Study Approach Through Segmentation by using Riughness Index and MRI" *Journal of Computer Science and Engineering ISSN-3785-0855 Vol.2 Issue 4 April 2015*.