International Journal of Scientific & Engineering Research, Volume 7, Issue 5, May-2016
ISSN 2229-5518



# Mammogram mass detection based on complex feature analysis

Reeba Mariyam Cherian, Dr. Shyama Das, Renju Rachel Varghese

**Abstract**— As breast cancer accounts for the greatest threat of cancerous form diagnosed in women. So it is important to classify the malignant and benign mass growths in mammogram, which stands as one of the reasons for breast cancer. This paper introduces CAD system which acts as a object machine opinion to assist radiologist and to increase the accuracy. This system initially pre-process the mammogram followed by finding out the suspicious mass regions. Where the features of suspected regions are extracted which are made use for the classification. The system considers both complex features by analyzing the values of co-occurrence matrix and by considering optical density transformation. Finally, this paper uses SVM for classification, achieving satisfactory detection performances..

**Index Terms**—Computer aided detection (CAD), Gray level co-occurrence matrix (GLCM), feature extraction, mammography mass detection, Optical density, Sech template, Support vector machine(SVM).

## 1 INTRODUCTION

Breast cancer accounts for the prevailing form of cancer that spreads very quickly and harmfully in females worldwide. 18.2%of all cancer deaths including both males and females are from breast cancer. Breast cancer treatments involve a combination of surgery, therapies. Clinical data show that women diagnosed with early stage breast cancer are less likely to affect badly of the disease than those diagnosed with more advanced stages.

Computer aided diagnosis is an important tool used by radiologists for interpreting medical images. Image processing techniques can be employed on the mammograms for the detection of breast cancer at an early stage. CAD, defined as a radiologist diagnosis where making use of a computerized analysis of medical images as a second opinion in lesion detection, assessing the extent of disease and making diagnostic decisions and is expected to increase the efficiency of interpretation component of medical imaging. With CAD the final diagnosis is made by the radiologist. Computerized image analysis has been applied mainly to medical imaging techniques such as X-ray, sonography and Magnetic Resonance Imaging.

This paper proposes an automatic CAD which helps radiologists to increase the accuracy of diagnosis. The abnormalities in mammograms can be classified into two: calcifications and masses. Calcification is calcium mineral deposits in breast tissues. They are seen as small white spots. Calcifications are of two types: micro calcifications and macro calcifications where micro calcifications are more cancerous than macro calcifications. Where the masses are localized collections of tissues. A mass is a space occupying affected tissue collection that can be seen in two different projections of X-ray view. When the projection is seen in one view, it is referred to as mammographic density. When the density is seen only in one view other views are required to confirm the presence of a mass. Shape, size, orientation and density of a mass stand out as an important factors in detecting whether a mass is suspicious or not. Mass detection is still challenging as it is surrounded by overlapping fat tissues.

There are different researchers based on mammographic mass detection. There are mainly three methods for the analysis of mass detection, by means of filtering the digital image or by template matching or with the help of gradient analysis orientation. All the three methods of mass detection are pixel based where the local image properties are computed at each pixel. In this paper, we follow template matching for the mass detection where this can be implemented by shifting a window across the mammogram while locally computing the correlation measure between the overlapping region and the assumed model.

## RELATED WORKS

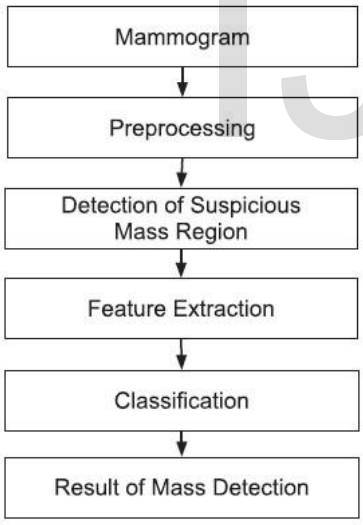
Xinbo Gao et al.[2] recently proposed a scheme to examine mammograms based on morphological component analysis (MCA), where the analysis of the extracted morphological features are done based on the piecewise smooth component for detecting ROIs. Finally ROIs, which satisfy the new concentric layer rules are extracted as the suspicious mass regions. Sameti et al. proposed a paper to examine mammograms for signs of tumor development without using previous mammograms as reference images. The paper is based on the fact that there exist differences between the region that subsequently becomes a malignant mass, and other normal areas of the mammography images taken in the last screening examination prior to the detection of a mass. Another technique proposed by Nevine H. Eltonsy et al. [4] based on the presence of concentric layers surrounding a focal area with suspicious morphological characteristics and low relative incidence in the breast region. A technique based on filtering is proposed by Nicholas Petrick et al. [5] introduces an algorithm using DWCE filtering with Laplacian Gaussian edge detection for segmentation of low contrast objects in digital mammograms. This algorithm is applied to enhance objects so that a simple edge detector can determine the object boundaries. Once they are known features can be extracted which can be used for classification.

In the paper, we propose a CAD system where we preprocess the mammogram were features are extracted from the suspicious regions followed by classifying them either as cancerous or not.

## 2 METHODS

In short the proposed scheme first provides a preprocessing step to preserve the breast area and eliminate the structural noise in the mammograms. Then, suspicious regions, select from the breast area using the Sech template matching method, and adaptive square ROIs are segmented from the original mammogram corresponding to suspected regions. Further, the characteristics of each square ROI are extracted based on the gray level image and the optical density image, respectively, for comparing predictive capabilities. Finally, SVM is used as the classifier which outputs with appreciable performance.

Thus the CADe system for mammographic mass detection comprises four major stages: preprocessing, detection of suspicious mass region, feature extraction and classification.

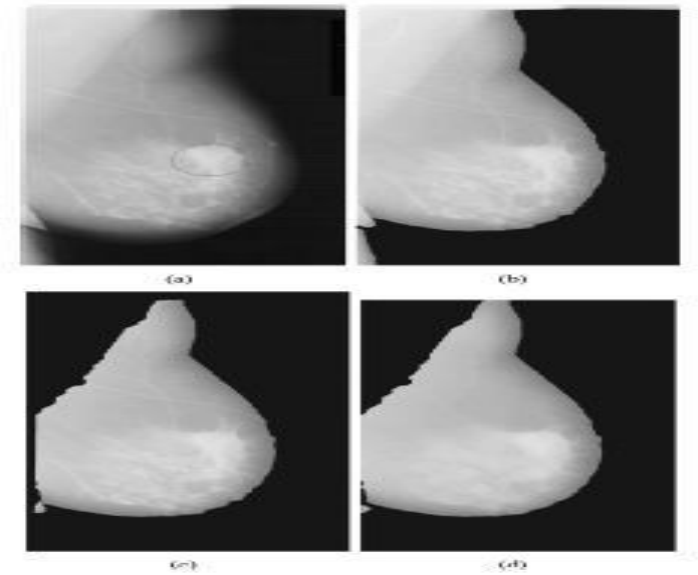



. Block diagram of the proposed mammographic mass detection scheme.

### 2.1 Image Preprocessing

Preprocessing is an important issue in low-level image processing. The underlying principle of preprocessing is to enlarge the intensity difference between objects and background and to produce reliable representation of breast tissue structures. An effective method for mammogram enhancement must aim to enhance the texture and features of masses. The reasons are: 1. Low-contrast of mammographic images, 2.hard to read masses in mammogram 3.generally, variation of the intensities of the masses such that radiopaque mass with high density and radiolucent mass with low-density in comparison with the background.

The breast region must be initially segmented from the image as shown in the image. Where in the image it is processed



To find the foreground of concern in digital mammogram Otsu thresholding is applied, where the foreground consist of a breast region and a pectoral muscle region in most mediolateral oblique (MLO) views of mammograms. As pectoral muscles are much brighter than the masses it may affect the detection result. So it is required to eliminate the pectoral muscle region Hence the whole foreground is transformed by gamma correction with a decoding gamma to preserve the brighter luminance and suppress the darker luminance. Thus gamma expansion enhances the pectoral muscle.

The maximum connected component finds the position of the pectoral region, and the border of the pectoral muscle is mended by erosion filter and dilation filter which are morphological filters. Consequently, the breast region is obtained by removing the pectoral muscle from the foreground. Mammographic preprocessing can also reduce the effects of image noises, blood vessels and glandular tissues, which lead to many FPs in the suspicious region detection stage both filters break narrow connections and eliminate minor protrusions. This slightly enhances the ambiguous margin of the mass reduces structural noises.

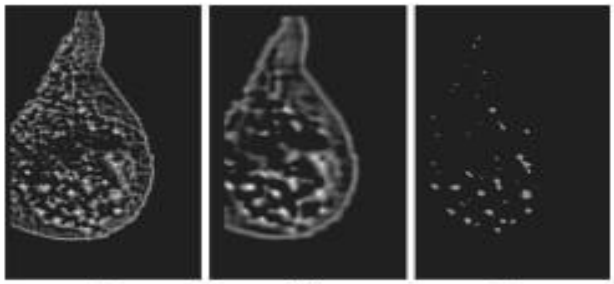
### 2.2 Detection of Suspicious Regions

Template matching is one of the most common approaches for medical image segmentation. This method uses the prior information of mammograms, and segments possible masses from the background using the prototypes. The prototype of possible masses is created based on the characteristics or physical features of the targeted masses, or based on the two dimensional search function. When the priori information about the size of the masses is not available, a range of sizes for the template is employed.

This paper makes use of two templates of Sech template to detect the suspicious breast masses on mammograms. Where the average of the two templates is accounted for the analysis.

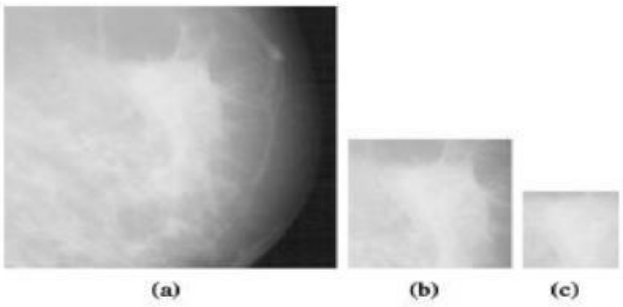
The Sech template used in this paper is defined as $S(x,y)=2/\exp((B*\text{sqrt}(x^2+y^2)+\exp(-B*\text{sqrt}(x^2+y^2))$.

Where x, y represents the coordinates of the template and various templates that can be obtained by different values of beta. Part based template (33*33 pixels) and complete based template (65*65 pixels) are the two templates used by Sech




template to measure the similarity between breast region and the template. Two correlation maps were obtained by the correlation measurement shown in following figure. Where (a) shows complete Sech template, (b) shows the part-based template and (c) shows the average correlation map.

Out of the two correlation maps an average correlation map is obtained is obtained to segment suspicious regions. By providing an appropriate threshold to the average correlation map which represents all the suspicious regions. As mass varies in size significantly an adaptive square ROI in the original mammogram was determined to obtain an object region based on the size of the related suspicious region. That means choosing an appropriate square ROI plays an important task.




An apt threshold value will fit the suspicious mass area completely in the region of interest. Several threshold values have been tested to obtain an optimal result.

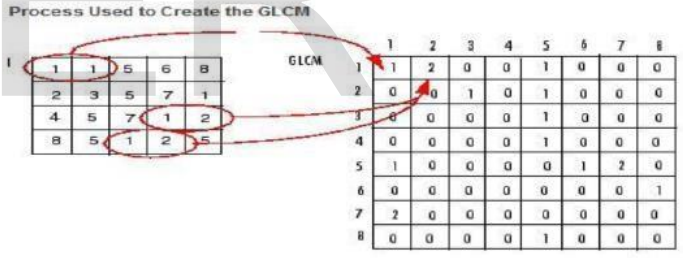
The above figure shows different mass segmented results of object region from ROI with different thresholds and proves that the optimal threshold value is 0.4 for this system. As in (b)as (a) includes too much normal tissue and (c) cannot cover the entire mass area. To extract the background information Of each object region a whole ROI is expanded proportionally.

## 2.3 Feature Extraction

The third stage of mass detection by CAD schemes in the feature extraction and selection. The features can be calculated from the ROI characterstics such as shape, size, density and smoothness of borders, etc,.The feature space is very large and complex due to the wide diversity of the normal tissues and the variety of the abnormalities. Only some of them are significant. Using excessive features may degrade the performance of the algorithm and increase the complexity of the classifier. To improve the efficiency of the classifier some redundant features must be avoided.

Followed by ROI segmentation some features are needed to be extracted to express the characterstics of the suspicious object region. Generally speaking, the intensity distribution of masses is an important characterstic for mass detection. Therefore, in our paper we make use of a pattern recognition method GLCM (gray level coocurrence matrix) to extract characterstics.

The idea behind GLCM is to describe features by a matrix of pair gray level appearing probabilities. Further, fourteen statistics that can be calculated from a coocurrence matrix with the intent of describing the texture of the image. The 14 texture features are defined as: Entropy, Energy, Local homogeneous, Contrast, Intensity, Correlation, Inverse difference moment, Sum average, Sum of squares variance, Sum entropy, Difference entropy, Inertia, Cluster shade, and Cluster prominence.




This paper proposes two complex feature extraction methods to achieve a complete description of quantitative characterstics. The first feature extraction module adopts GLCM and optical density features. This is a type of complex texture feature extraction method that extracts the information of local intensity relation and discrete photometric distribution.

The proposed scheme computes four co-occurrence matrices with one pixel distance in four directions: left diagonal, right diagonal, vertical and horizontal.

GLCM function characterizes the texture by calculating how often pairs of pixel with specific values and in a specified spatial relationship occur in an image. Where homogeneity is calculated by measuring the closeness of elements in the GLCM to the GLCM diagonal. Similarly, energy is measured by the sum of squared elements in the GLCM. Contrast is measured by the local variations in the GLCM.

Another complex feature extraction method is also constructed that is similar to the proposed complex module, but translates the GLCM into the optical density co-occurrence matrix (ODCM) to characterize the photometric textures.

The ODCM is a co-occurrence matrix of the optical density image. An optical density image can be obtained by changing the intensity of the gray level image into optical density and linearly mapping each optical density value to an image with 8-bit depth data. The minimum optical density vale was mapped to 0, and the maximum optical density value was mapped to 255.


The above figure shows a stimulated ROI of a mammographic mass in a breast with high density score.

The intensities were high and close and making it difficult for human eyes to determine the difference in a dense density mammogram. There are three manual gray level values in (a) 245, 250 and 255 from peripheral to the center of the ROI. (b) shows the optical density image of a. The gray level values in fig b are 0, 129 and 255, which correspond to manual values in (a) respectively. After transforming the gray level image into the optical density image, the differences between gray level

values are enlarged, enhancing the simulated mass region. As the background represents the surrounding normal tissues in an ROI with appropriate thresholding, an optical density image can serve as a graph that represents the degree of abnormal tissue based on the intensity (the lighter area represents a greater possibility of diseased tissue). Eventually the two proposed methods combining texture features and optical density features use 70 statistics to achieve a perfect description of features.

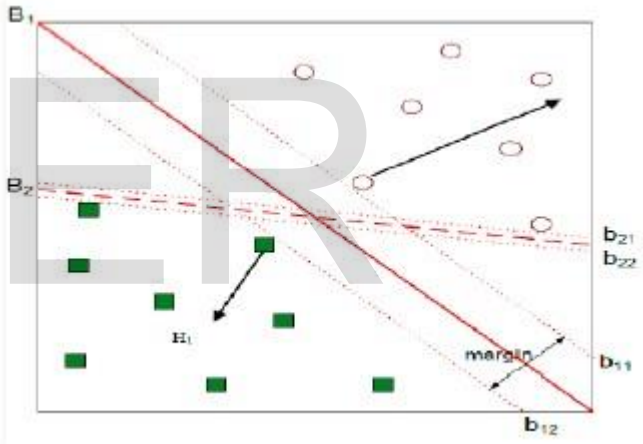
Features extracted from the gray level characteristics, pattern and texture of the lesion and the surrounding tissue can usually be shown as a mathematical description, and are helpful for a classifier to distinguish masses as malignant or benign. Merely it is really hard to predict which feature or feature combinations will achieve a better classification rate. More often than not, different feature combinations will result in different performance. In addition relatively few features used in a classifier can keep the classification performance robust.

### 2.4 Classification

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Support Vector Machines (SVMs) are a relatively new supervised classification technique to mammography field. They hold their roots in Statiscal Learning Theory and have acquired prominence because they are robust, accurate and are efficient even when applying a small training sample. By their nature SVMs are essentially binary classifiers, all the same, they can be taken to cover the multiple classification tasks.

Passed a lot of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one class or the other, causing it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are split by a clear gap that is as broad as possible. New cases are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

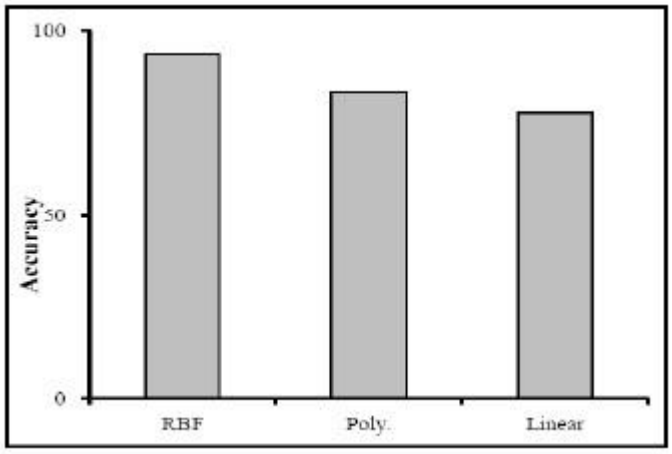
The following figure shows the basics of SVM.




When data are not labeled, a supervised learning is not possible, and an an unsupervised learning is required, that would find natural clustering of the data to groups, and map new data to these formed groups. The clustering algorithm which provides an improvement to the support vector machines is called support vector clustering and is frequently applied in industrial applications either when data is not labeled or when only some data is labeled as a preprocessing for a classification pass.

In addition to performing linear classification, SVMs can efficiently perform a nonlinear classification using what is called the kernel trick, implicit mapping their inputs into high dimensional feature spaces.

SVM solves the problem by using kernel functions. Kernel functions effectively map the original feature vectors into higher dimensional space without explicit calculation. There exist different types of predetermined kernel: Linear, Polynomial, Quadratic, Radial basis function. Comparison of different kernel functions is shown below:




In this paper, we make use of kernel function radial basis function.

## 4 Experimental results

### 4.1 Dataset

This paper has used data form MIAS (Mammographic Image Analysis Society).Where MIAS is a UK based organization specialized for research interested in the understanding of

mammograms and have generated a database of digital mammograms. Its purpose is to fend for the CAD system by offering a packet of information including images, patient information, lesion information, pathology. The database scheme is able to store and recover images in time for research purposes and for radiologists Image films have been withdrawn from the UK National Breast Screening Program which have been digitized to a 50 micron pixel edge with a Joce-Loebl scanning microdensitometer in which each pixel is represented with an 8-bit word. The database consists of 322 digitized films. It also consists of comments of radiologists regarding the abnormalities, which are used for comparing the results of classifiers. The system evaluates 322 mammograms from MIAS which is divided into a training set and test set.

### 4.2 Feature Analysis

After the training phase the discriminant functions which are the combinations of GLCM features and optical density features. 14 statistical features of mammogram and optical density image are extracted using GLCM. The combination of 14 features will result in 70 statistics to achieve a perfect description of the mammogram.

The basic 14 features analyses include significant features which are as follows: Entropy, Energy, Local homogeneous, Contrast, Intensity, Correlation, Inverse difference moment, Sum average, Sum of square variance, Sum entropy, Difference entropy, Inertia, Cluster shade, Cluster prominence.

The Intensity values reveal the average brightness of the whole object of concern. Larger values of brightness exposes a greater possibility of a mass. That does not mean smaller intensities does not take in the bearing of a mass. Similarly to detect the disorders in textures can be ground out by evaluating the texture uniformity which can be reached by local homogenous feature. To find the amount of edged information we need to cognize about the complexity of the sum of pixel pair which is broken by sum entropy. Similarly, each feature serves an important part in the image data extraction.

### 4.3 Comparison with basic classifiers

Classifiers are used for assigning objects into related classes by means of data mining algorithms. The comparison of classifiers and selecting the most appropriate classifier is very important. Each classification method shows different efficiency and accuracy based on the kinds of dataset. In this paper, we make use of a comparison of four different classification methods.

**1. Naive Bayes classifier**

Bayesian classifiers are statistical classifiers. They are used to predict class membership probabilities. These probabilistic classifiers are commonly used in machine learning.

Naïve Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. This considers each of these characteristics to contribute independently to the probability regardless of any possible correlations.

**2. Linear Discriminant Analysis classifier**

While talking about linear discriminant analysis using the step wise feature selection method it selects the first feature based on each features individual performance. Subsequent features are selected based on the combined improvement inn classification performance. The advantages of this method are that it implicitly accounts for the correlation of features and selects the number of features to be used as inputs.

A common method to reduce the number of the features and to obtain the best features it is known as feature selection using step wise linear discriminant analysis. Stepwise feature selection is a heuristic procedure using statistical techniques based on Fishers linear discriminant. T the beginning, the selected feature pool is empty. At each step followed one available feature input into or removed from the selected feature pool by analyzing its effect on a selection criteria.

**3. Decision Tree classifier**

Decision trees are used to support decision making in an uncertain environment. Decision trees are made for a type of analysis which are relatively easy to apply.

A decision tree has three types of nodes:

1. Decision node.
2. Chance node.
3. Leaf node.branches.

The branches originating from a decision node represent options available. At each chance node each branch is assigned a conditional probability equal to the probability of the event presented by the branch, conditioned upon the knowledge available at the node. Leaf nodes represent the possible endpoints. That is, the results of the decisions and chance outcomes associated with the tree path from the start of the tree.

**4. Support Vector Machine classifier**

A support vector machine (SVM) is a classification algorithm that makes use of a non linear mapping to transform the original training data into a higher dimension. This will result in a new linear optimal separating hyperplane. This hyperplane is a decision boundary that separates the tuples of one class from another. Support vectors (training set) and the margins (defined by the support vectors) finds the hyperplane for the SVM.

The basic idea behind SVM is that there exists a linear boundary that separates the dataset into two. Then we need to see out the linear hyperplane that separates the peaks into two different categories. In principle there exist infinitely number of hyperplanes that can separate the training data. So by SVM, we need to find the optimal hyperplane.

Although the training time of SVM can be slow they are highly accurate when compared with the other classifiers.They are much less prone to overfitting than other methods.

## 6 CONCLUSION

This paper presents a CAD system that helps in the detection of masses in mammograms, The CAD system helps radiologists in the diagnosis of abnormalities quicker than earlier procedures. This report introduces an automated mammogram mass classification method based on complex texture features. Extracted features classified with SVM which resulted in appreciable accuracy. Dataset considered in this report is used up from mini MIAS digital datasets of mammograms.

Presence of localized collections of tissues represents a mass, where it is one among the abnormalities seen on mammograms. Earlier detection of abnormalities helps in getting rid of the necessity for invasive surgical operations and handling.

This paper makes use of template matching method for detection. For acquiring better results we demand to consider only mass regions and demand to get rid of all other regions. For the preprocessing this paper makes use of Otsu thresholding followed by gamma correction, which enhances the pectoral muscle region, which takes in the mammogram other than the part of interest. So these pectoral muscles needed to be eliminated by morphological filters. Thus, as a result of preprocessing the mass regions are highlighted.

Thus the suspicious mass regions are detected by a template matching method. For that Sech template is made used, for which two types of Sech template i.e., complete template and part-based template are made used. Out of which a correlation average map is considered highlighting the suspicious mass regions.

From the suspicious mass regions features are extracted by means of GLCM. GLCM function identifies the texture by calculating how often pairs of pixel with specific values and in specified relationship occur in an image. GLCM stands as the factor for extracting features from the original image and from the optical density image. 14 texture features can be extracted by means of GLCM. This combination of features results in the study of 70 statistical texture features.

Finally Support Vector Machine is employed for classification which is a linear classifier resulting in accurate, robust, and effective outputs. By SVM an optical hyperplane is designed grouping the set into two. Out of many kernel functions in SVM this paper makes use of Radial Basis Function as the core component.

The experiments have shown that the method pursued in this paper achieves appreciable detection results.



## REFERENCES

[1] Shen-Chuan Tai, ZihSiou Chen, Wei Ting Tsai IEEE, An automatic mass detection system in mammograms based on complex texture features, IEEE journal of Biomedical and Health Informatics, VOL,. 18, NO.2, pp. 618-627, Mar.,2014.

[2] X. Gao , Y. Wang, X. Li and D,Tao, On combining morphological component analysis and conventric morphology model for mammographic mass detection, IEEE Trans. Inf. Technol. Biomed., vol.14, no. 2, pp. 266-273, Mar.2010.

[3] M. Sameti, R. Ward, J. Morgan-Parkes, nad B, Placic, Image feature extracton in the last screening mammograms prior to detection of breast cancer, IEEE J. Sel. Topics Signal Process, vol. 3, no 1,pp. 46-52, Feb. 2009

[4] N. Eltonsy, G.Tourassi, and A. Elmaghraby, A concentric morphology model for the detection of masses in mammography, IEEE Trans. Med. Ima., vol. 26, pp. 880-889, Jun. 2007.

[5] Niholas Petrick, Member, IEEE, heang-Ping Chan, Berkman Sahiner, Mmeber, IEEE, and Datong Wei, An adaptive density weighted contrast enhancement fiter for mammographic breast mass detection, IEEE Trans. Med. Imag. Vol 15, no. 59-67 Feb. 1996.

[6] Fritz ALbregtsen. Image processing laboratoru, Department of Informatics University of Oslo, Statistical Texture Measures Computed from Gray Level Coocurrence Matrices, pp. 1-14, Nov 5, 2008.

[7] P. Mohanaih, P. Sathyanarayana, L. Gurukumar, Image Texture Feature Extraction Using GLCM A approach, International Journal of Scientific and Research Publications, Vol 3, issue 5, May 2013.

[8] Fractal Based Techniques for Classification of Mammograms and Identification of Microcalcifications, Thesis submitted to Cochin University of Science and Technology by Deepa Sankar.

[9] Guido M. te Brake and Nico Karssemeijer, Single and multiscale detection of masses in digital mammograms. IEEE Trans. Med. Imag, vol 18, no.7 July 1999.

[10] Fraschini, Mammographic mass classification: Novel and simple signal analysis method, Electron Lett., vol. 47, pp 14-15,2011.