IJCTA, 8(3), 2015, pp. 899-906

Detection of Breast Masses in Digital Mammograms using SVM

S.P. Meharunnisa, Dr. K. Suresh, Dr. Ravishankar M. and Amith Bhaskar

Abstract: Breast Cancer stands to be the most deadly disease among women caused due to hormonal imbalances, varying stress levels and many other unknown reasons. Though the technology required to prevent this type of cancer remains yet to be discovered, we can devise ways to detect & diagnose cancer at its early stages and treating it clinically before proceeding to biopsy. The advancement of CAD (Computer Aided Design) has enabled us to indigenously design a novel system to recognize and differentiate probable signs of cancer from normal and abnormal patients. In this paper, we have proposed a method to investigate the masses in breast cancer patients by extracting texture features from digital mammograms and predict the condition of diagnosis in these patients. We have used image processing techniques on the publicly available mini-MIAS mammogram database such as median filtering, CLAHE, GLCM and SVM to obtain qualitative features of the image and classify it. This type of CAD method helps in greatly reducing the burden of radiologists and acting as a second opinion to doctors in the process of reducing deaths due to breast cancer. We have obtained a CAD model with 87.5% of Sensitivity and 100% Specificity with 95% Overall Accuracy.

Key words*:* SVM; GLCM; K-means clustering; Breast Cancer.

1. INTRODUCTION

Breast cancer has become a significant issue worldwide. In order to prevent the increase of deaths caused by breast cancer, early diagnosis of this disease is important. Detected early, breast cancer is much easier to treat, with fewer risks and reduces mortality by 25%. Prevention of breast cancer is only possible if probable patients go through a regular screening process, as it takes more than five years for any breast mass to grow 1 mm, another 2 yrs to reach half of it and within 1yr grows to 2cm, visible enough to be extracted. The suspected symptoms causing breast cancer are post menopause, stress, hereditary, alcoholism, obesity, hormonal imbalances and genetically mutated abnormalities. Even though there are many advancements in image modalities, such as MRI & sonography, full field digital mammogramsare the most reliable and inexpensive. It is well known that there is no technology at present, which is capable of curing cancer, but detected early, it can aid in recovery and prolong patient life.

A large group of errors remains due to the process followed by manual inspection by radiologists. Due to this visual inspection process of countless mammograms, radiologists easily misinterpret cancers, also miss out vital points while investigating these scans. Mammography being the best available inspection technique, and relatively less expensive, to detect the symptoms of breast cancer at the early stage, can disclose many information about these abnormalities like masses, microcalcifications, architectural distortion and bilateral asymmetry.

For the hundreds of mammographic images scanned by a radiologist, only a few are cancerous. While detecting abnormalities, some of them may be missed as the investigation of suspicious regions is a

DSCE, SDMCE, JCE & DSCE
* e-mail: lipy2@163.com

recurrent mission that causes fatigue and eyestrain. Mammography is the breast image taken with a special low-dose of X-ray, bright contrast and mega resolution films. Using enhanced images & segmentation of suspicious areas, extraction of features, we can select more accurate features and further classify them into appropriate category of Breast Cancers. These are the mandatory stages that all CAD are required to follow. Sometimes less contrast exposure and stronger noisecauses digitized mammography outputs to become unreadable. As any cancerous abnormalities appear opaque and denser tissue transparent they appear quite easy to detect using image processing. The difficulty in going through laborious mammograms can be overcome by CAD and eliminate the drawbacks in the existing method of cancer detection, whereas an improvement in local information and noise removal from thesemammograms are required.

2. LITERATURE SURVEY

A complete review of previous studies provides enormous insight before carrying out any proposed work and the literatures mentioned here are the motivations behind this work. The main focus of the study was to explore the existing technologies and identify methods employed to implement algorithms for identification of early signs of breast cancer.

Bommeswari Barathi [1] proposed that Mean filter and Median filter where Mean filter is also called as average filter which helps in smoothing the image and this filter performed as low-pass one. Each pixel is replaced by the mathematical median of gray levels in its neighborhood. In mean filter, every part of the pixel which falls below the mask are averaged to form single pixel. Advantage in this method is that it reduces the variance, easy to implement and disadvantage is that poor in edge preserving. Mathematical median value tends to preserve the image sharpness and its object borders. Every pixel of filtered image is considered as local brightness value of its corresponding neighborhood in original image. Merits in median filter is to reduce the errors and demerits is it removes some fine detail and noise which is acceptable.

In this study [2],we discuss the development of a technology which is able to mark the positions of possible masses, allowing further assessment by radiologists and effectively increasing the rate of correct diagnosis of breast cancer. Because masses in mammograms present themselves as low frequency signals, we have established the following steps for detecting them: Firstly, the original image undergoes wavelet transformation and enhances the mass signals before being inverse-transformed backward to an image; an image with enhanced processes would make masses easier to discern. Second, possible masses are identified and positioned using particle swarm optimization, PSO. Experimental results show that detection rate of 94.44% or higher can be achieved using this method, hence improved accuracy in breast cancer lesion detection.

Jawad Nagi, Sameen Abdul Kareem, FarruknNagi, Syed Khaleel Ahmed [3] proposed the breast profile segmentation by utilizinga Seed region growth segmentation technique in which a cluster of gray levels with average intensities are selected and the region grows to form a segmented output. Seeded region growing performs a segmentation of an image with respect to set of points known as seed. In this the process evolved from the initial state of seeds set, S1,S2,….., Sn. In this algorithm, each step involvesthe summation of each pixel to the connected pixel group. It provides clear visual outputs having proper edges Disadvantageof this technique is that we obtain a huge quantity of segmentationoutputs in the image having local minima in each mammogram.

R. Ramani, Dr. S. Suthanthiravanitha, S. Valarmathy [4] described about the fuzzy C-mean clustering which is similar to k-mean but in fuzzy, each point has a average weight associated with a cluster. Drawback is that fuzzy c-mean suffer from the presence of outliers and noises. Hence it is not easy to identify the initial partition. Advantage is it gives better result than k-mean algorithm in performance.

D. Lavanya and Dr. K. Usha Rani [5] proposed that decision tree is one of the classification methods, which classify the labeled trained data into a tree or rules. Once the tree or rules are derived in learning phase to test the accuracy of a classifier test data is taken randomly from training data. After verification of accuracy, unlabeled data is classified using the tree or rules obtained in learning phase. The decision tree structure represents a tree with roots, branches to the left and right. The leaf nodes in a tree represent a class label. The arcs from one node to another node denote the conditions on the attributes. The Tree can be built as, the selection of attribute as a root node is done based on attribute splits, the decisions about the node to represent as terminal node or to continue for splitting the node and the assignment of terminal node to a class.

Sheng-Wen Zheng, Jui Liu, Chen-Chung Liu [6] proposed algorithm adapts the "modified gradient vector flow (MGVF) snake" method to determine the breast region from a mammogram image, and uses Otsu's threshold technique and recursive regression analysis to delete the pectoral muscle from the breast region. It further utilizes upper outlier detection and texture complexity analysis to segment the initial breast tumor regions, and finally, segments the final breast tumor image from the initial breast tumor regions by using random walk scheme.

Based on the survey, most of the works mainly concentrate on identification of shape and region characteristics present in the mammograms. Additional processing steps are combined with existing algorithms to improve detection rate by further contrast enhancement in the case of identified masses. The mammograms are pre-processed before actually segmentingthe breast masses which helped to achieve higher accuracy.

3. PROPOSED WORK

Masses and micro calcifications are the main indications of abnormalities in many mammograms. Numerous literatures report that determining mass ismuch difficult to recognize due to its uneven appearance with shallow borders than microcalcifications, therefore making detection of masses a challenging task. A Breast Mass is characterized as spaceoccupiedtumordetermined by its shape and marginal property. We propose a method in which we preprocess the mammogram image using a 2D Median Filter, segment the suspicious mass region using K-means Clustering, and extract features before classifying it with Support vector machine (SVM) into normal or abnormal mammograms.

3.1 Materials and Methods

The Mammographic Image Analysis Society (MIAS) is a self-established researcher group in UK indulged in collecting and publishing digital mammograms over the internet. Images acquired consist of fatty-glandular, fatty and dense-glandular mammograms of breast cancer patients. The images, digitized at 200 micron pixel edge and padded in order to obtain a uniform image size of 1024 × 1024 pixels. MIAS database consist of 322 Images which is classified as Normal (209), Benign (62) and Malignant (52). Images format is in PGM (Portable Gray Map) which is a type of lossless image format and the details of the image will not lost at the time of data compression.

3.2 Preprocessing and Segmentation

The median filter is utilized to average the noise artifacts present in the image. The technique is implemented by considering the image as a 2D matrix. The functioning of this filter is realized by convolving the sum of products of the two elements from the 2D matrix one over the other. This process can be made quick enough by using smaller filters such as 5x5 or 3x3 dimensions.The contrast of mammogram image is increased by using CLAHE (Contrast Limited Adaptive Histogram Equalization).

Contrast Limited AHE (CLAHE) is differentwhen compared to regular Adaptive Histogram Equalization techniques due to its limiting in contrast. This modification when applied globally, the histogram becomes equalized, and gives rise to Contrast Limited Adaptive Histogram Equalization.CLAHE preventsover amplification of noise that is seen in regular adaptive histogram equalization techniques. A transformation function is implemented transforming each neighborhood of gray level intensities.

Image segmentation is the procedure of separating an image into significantareas based on similarity or heterogeneity measures. Using K-means clustering we associate pixels with same intensities into a set of pre-defined groups. It is based on recursive iterations and is used to partition the whole image into k clusters. K-means is one of the unsupervised – simple learning algorithms. It divides a collection of regions into K groups. We find that k=5 is a best fit for all mammogram images provided in our database this k-means algorithm iterates mainly over 2 steps. Computing the mean of each given k-cluster and then computing the Euclidean measure of distance of each data point from each centroid of a cluster and assigning it to the nearest cluster. In the iteration course being executed, the Euclidean distance is minimized, in all the groups, which are far off the centroids from the respective data points. Confluence of the data points is such that they are geometrically compact as possible around their centroids.

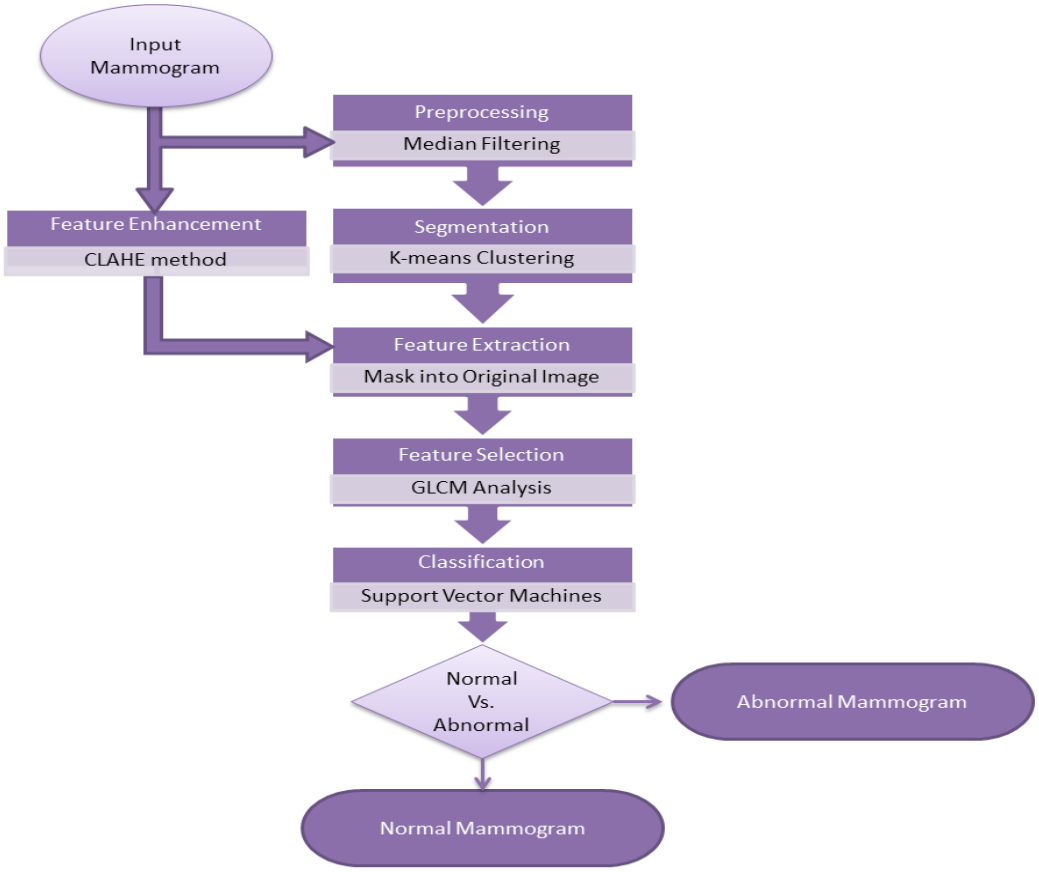

Figure 1: Block Diagram

3.3 Feature Extraction and Classification

When a radiologist is needed to inspect a tumor from a mammogram, the onus is on detecting the borders of it. The observations are relied on size, margin and shape. The decision on each mammogram is different as compared to each radiologist's opinion and experience. To solve this problem of indifferent interpretation, numerous features are extracted to obtain more information about the tumor. The preprocessed image is

segmented into a pre-defined 5 clusters by using K-means Clustering technique as shown in the image below.

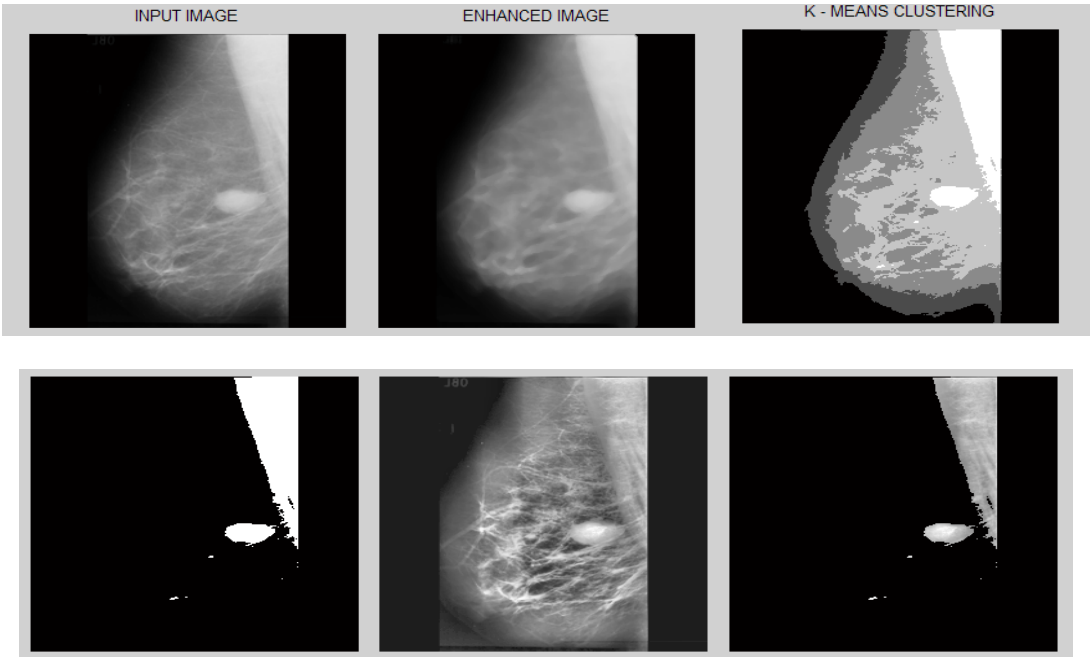

Figure 2: K-means segmentation & Feature Extraction of patient mdb025

The output of the K-means cluster which generally contains the cancerous mass and other regions is masked onto the CLAHE enhanced original mammogram image as shown. The GLCM features are extracted from this image and is cross validated from the glcm matrix of the training database of Normal & Abnormal images to classify this patient's condition. SVM is a binary classifier which utilizes a set of pre-defined data to extract features and build a graph with support vectors mapped on a vector space. Each support vector is classified into two categories separated by a hyper plane, where the two categories taken here are Cluster Shade & Energy in the list of GLCM features. The new test data extracts the same similar feature vector and is characterized by determining which category it belongs to by plotting on the either side of the hyper plane. A SVM with RBF (Radial Basis Function) kernel is used to perform this classification by fitting the best possible hyper plane as shown in the graph below.

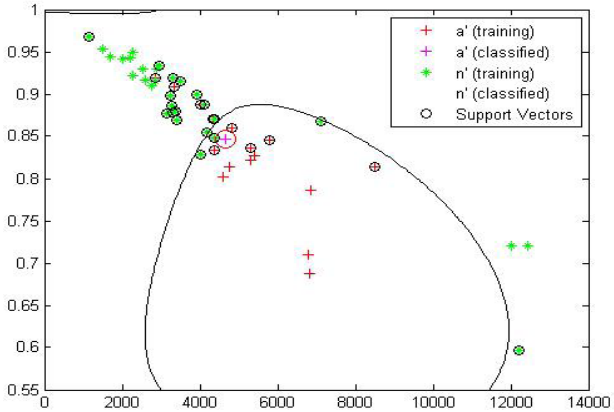

Figure 3: SVM graph classification for patient mdb0

4. RESULTS AND ANALYSIS

A few mammograms are randomly selected and stored as a testing database and are unique with respect to the training section. These testing database of mammograms are provided as an input for testing the accuracy of the SVM classifier. The outputs after classification along with their actual ground truth information, provided in table 1.

Table 1

Testing database details

Mammograms	*Ground Truth*	*Classifier Output*
mdb019.pgm	Abnormal	Abnormal
mdb021.pgm	Abnormal	Abnormal
mdb023.pgm	Abnormal	Abnormal
mdb025.pgm	Abnormal	Abnormal
mdb028.pgm	Abnormal	Abnormal
mdb141.pgm	Abnormal	Abnormal
mdb195.pgm	Abnormal	Abnormal
mdb198.pgm	Abnormal	Abnormal
mdb199.pgm	Abnormal	Abnormal
mdb202.pgm	Abnormal	Abnormal
mdb206.pgm	Abnormal	Abnormal
mdb207.pgm	Abnormal	Abnormal
mdb302.pgm	Normal	Normal
mdb303.pgm	Normal	Normal
mdb305.pgm	Normal	Abnormal
mdb306.pgm	Normal	Normal
mdb307.pgm	Normal	Normal
mdb308.pgm	Normal	Normal
mdb309.pgm	Normal	Normal
mdb310.pgm	Normal	Normal

The support vector machine successfully predicted 19 out of 20 patients into Normal and Abnormal cases. This was verified using the publicly available mini-MIAS database. The SVM graph is plotted using the support vectors and a curve is fitted to perform the classifications. We can draw inferences from the classification results by using a confusion matrix as given below:

A confusion matrix is a classification table which is often used to express the performance of a classifier whose true values are known. The basic terms derived from the confusion matrix are

- True positive(TP); correctly predicting a label (here TP = 12)
- True negative(TN) ; correctly predicting the opposite label(here TN = 7)
- False positive(FP) ; falsely predicting a label (here FP = 1)
- False negative(FN) ; missing a label(here FN = 0)

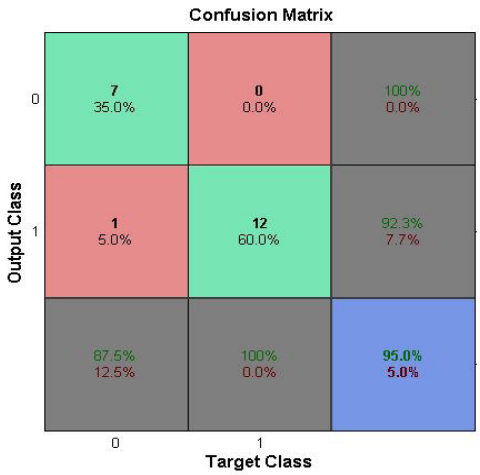

Figure 4: Confusion Matrix

For this binary classifier, we validate by using two metrics called sensitivity and specificity. The sensitivity is a proportion of information that are positive of all the information that are actually positive. Specificity deals with the proportion of information that are negative of all the information that are actually negative. F Score is the weighted average of the true positive rate (recall) and precision. The comparison of Sensitivity, Specificity & F Score are as given below in Table 2.

Table 2

Comparison of efficiency of the CAD model

Kernal Type	*Accuracy*	*Sensitivity*	*Specificity*	*F Score*
RBF/Polynomial	95%	87.5%	100%	0.933
Quadratic	90%	75%	100%	0.857
Linear/MLP	62.5%	62.5%	100%	0.769

Table 3

Comparing existing and proposed methods

Method	*Accuracy*
Proposed GLCM & SVM method	95%
Co-occurrence & Bayesian Neural Network [7]	90%
GLCM & ANN [8]	87.5%
Multiwavelet & Statistical Analysis [9]	65%

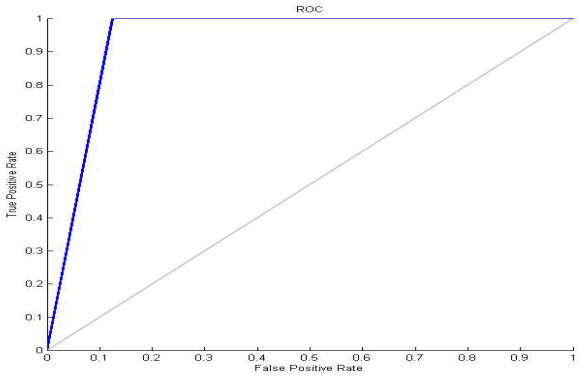

Figure 5: ROC Curve for the proposed SVM model

Table 3 denotes the comparison of our proposed work to that among several existing methods. ROC Curve is the most common graph that is used to report the performances of all classifiersacross many threshold. It formed by a plot of True Positive Rate (in the y-axis) vs. the False Positive Rate (in the x-axis) for various observations in a given class. Ideally, the curve is expected to move toward the left-top portion for correct prediction of the cases in the model. The line at the center is for a random output generationmodel. The ROC for the SVM model with RBF kernel is as shown in Figure 5

5. CONCLUSION & FUTURE SCOPE

Breast cancer is one of the major reasons of death in women. This paper presents a SVM model for determining the condition of breast cancer based on glcm features of the cancerous region in the mammograms. The SVM achieves an overall accuracy of 95% with a specificity of 87.5%. With the precision being maximum we are able to successively classify the unknown patient's mammograms and obtain satisfactory results. Several inferences can be drawn over the collection of GLCM features from the data stored in the overall database.

Future work can be develop by using hybrid models for obtaining higher sensitivity and accuracy in tumor identification rate. Other future work can be improving existing methods such as enhancing low contrast mass calcifications and method can also be extended to detect advanced stages of cancer and PCA analysis of GLCM features. Further automatic feature selection using data mining algorithms can be provided as an input to the support vector machines for better results in different categories.

References

1. Bommeswari Barathi, March 2014. Effective Filtering Algorithm for Enhancing Mammogram Images, Volume 2, Issue 3, *IJARCSMS*.
2. "Mass Detection in Digital Mammograms System Based on PSO Algorithm", Ying-CheKuo, Wei-Chen Lin, *Shih-Chang Hsu & An-Chun Cheng,* 978-1-4799-5277-9/14.
3. Jawad Nagi, Sameem Abdul Kareem, Farrukh Nagi, Syed Khaleel Ahmed, "Automated Breast Profile Segmentation for ROI Detection Using Digital Mammograms", 2010 IEEE EMBS Conference on Biomedical Engineering & Science (IECBES 2010) 30th November-2nd December 2010.
4. R. Ramani, Dr. S. Suthanthiravanitha, S.Valarmathy, "A Survey of Current Image Segmentation Technique for Detection of Breast Cancer", *International Journal of Engineering Research and Applications (IJERA),* Volume 2, Issue 5, September-October 2012.
5. D. Lavanya & Dr. K. Usha Rani, 2011. Analysis of Feature Selection With Classfication: Breast Cancer Datasets. *Indian Journal of Computer Science and Engineering* (IJCSE) Vol. 2 No. 5 ISSN : 0976-5166.
6. Sheng-Wen Zheng, Jui Liu, Chen-Chung Liu, 2013. A Random-Walk Based Breast Tumors Segmentation Algorithm for Mammograms. *International Journal on Computer, Consumer and Control (IJ3C),* Vol. 2, No.2.
7. "Classification of Normal, Benign and Malignant Tissues Using Fuzzy Texton and Support Vector Machine in Mammographic Images", Venkataragha, Sudhakar Putheti, *International Journal of Applications*, Vol. 82-No.15, November 2013.
8. "Classification of Normal and Abnormal Patterns in Digital Mammograms for Diagnosis of Breast Cancer", R. Nithya, B. Santhi, *International Journal of Computer Applications* (0975 – 8887), Volume 28– No.6, August 2011.
9. "Early Stage Detection of Cancer in Mammogram Using Statistical Feature Extraction", M. Vidya, N. Sangeetha, Vimal Kumar, Helen Prabha, International conference on recent advancements in electrical, electronics and control engineering; 2011 *IEEE*.