ANALYSIS OF BREAST CANCER DIAGNOSIS UING OMBINATION OF TEXTURE FEATURES FROM MLO AND CC VIEW MAMMOGRAMS

A PROJECT REPORT

Submitt
ed by

RASHEEDHA.A

Reg. No: 13MAE13

In partial fulfillment for the requirement of award of the degree

of

MASTER OF ENGINEERING

APPLIED ELECTRONICS

Department of Electronics and Communication Engineering

KUMARAGURU COLLEGE OF TECHNOLOGY
(An autonomous institution affiliated to Anna University, Chennai)

COIMBATORE - 641 049

ANNA UNIVERSITY: CHENNAI 600 025

APRIL-2015

BONAFIDE CERTIFICATE

Certified that this project report titled **"ANALYSIS OF BREAST CANCER DIAGNOSIS USING COMBINATION OF TEXTURE FEATURES FROM MLO AND CC VIEW MAMMOGRAMS "** is the bonafide work of **RASHEEDHA A [Reg. No. 13MAE13]** who carried out the research under my supervision. Certified further that, to the best of my knowledge the work reported herein does not form part of any other project or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE	**SIGNATURE**
Ms. S.SASIKALA	**Dr. RAJESWARI MARIAPPAN**
PROJECT SUPERVISOR	**HEAD OF THE DEPARTMENT**
Department of ECE	Department of ECE
Kumaraguru College of Technology	Kumaraguru College of Technology
Coimbatore-641 049	Coimbatore-641 049

The candidate with university **Register No.13MAE13** is examined by us in the project viva-voce examination held on...........................

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

First I would like to express my praise and gratitude to the Lord, who has showered his grace and blessing enabling us to complete this project in an excellent manner. He has made all things in beautiful in his time.

I express my sincere thanks to my beloved Joint Correspondent, **Shri. Shankar Vanavarayar** for his kind support and for providing necessary facilities to carry out the project work.

I would like to express my sincere thanks to our beloved Principal **Dr.R.S.Kumar M.E., Ph.D.,** who encouraged me with his valuable thoughts.

I would like to express my sincere thanks and deep sense of gratitude to my HOD, **Dr. Rajeswari Mariappan M.E., Ph.D.,** for her valuable suggestions and encouragement which paved way for the successful completion of the project.

I greatly privileged to express my deep sense of gratitude to the Project Coordinator and Supervisor **Ms.S.Sasikala.M.Tech., (Ph.D)**, Associate Professor, for her continuous support throughout the course. In particular, I wish to thank and express my everlasting gratitude to her for the expert counselling in each and every steps of project work and I wish to convey my deep sense of gratitude to all teaching and non-teaching staff members of ECE Department for their help and cooperation.

Also I am very much thankful to the courtesy of TM Deserni, Dept. of Medical Informatics and RWTH Achen, Germany for providing their data to this research work.

Finally, I thank my parents and my family members for giving me the moral support in all of my activities and my dear friends who helped me to endure my difficult times with their unfailing support and warm wishes.

ABSTRACT

The classification of breast masses into benign and malignant categories plays an important role in the area of computer-aided diagnosis (CAD) of breast cancer. Normally the mammogram is available in two views namely MLO (Medio Lateral Oblique) & CC (Cranio Caudal) views. This proposed method deals with the preprocessing of the input images from MLO & CC views, segmenting them and extracting the features using Gabor filter, GLCM (Gray level Co occurrence Matrix) and the steerable pyramid separately. Combining those extracted features and classifying them using the (MLP) Multilayer Perception classifier. The main aim of this work is to improve the accuracy for early diagnosis of breast cancer. This work proves that the accuracy of the data sets are improved when the features extracted from the CC and MLO views are combined together.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBRVIATIONS

ABBREVIATION	NOMENCLEATURE
CC	cranio cardial
MLO	Mediolateral oblique
CAD	Computer aided diagnosis
PCA	Principal component analysis
ML	Multi layer perceptron
BPN	Back propagation neural network

CHAPTER 1

INTRODUCTION

The ultimate objective of this project is the diagnosis of the breast cancer in the early stage. The inclusion of information from multiple views of the breast may improve the performance of diagnosis . A few attempts have been made to incorporate the information from multiple mammographic views to improve the accuracy and sensitivity of the same**.**

1.1 BREAST CANCER

Breast cancer is one of the abnormality and most common form of cancer found among women. It is the second leading cause of death rate after lung cancer. Early detection of breast cancer is very important in increasing in survival rates and also helps in increasing treatment option mammography is a conventional detection method of breast cancer. Radiologist search for these abnormalities using mammogram.

1.2 MAMMOGRAPHY

Mammography is the study of Breast using low dose x-rays. The actual test is called Mammogram. The goal of mammography is the early detection of the breast cancer, typically through the detection of characteristics masses or micro classification. The two standard views of the mammogram are MLO and CC.

1.3 MLO and CC VIEW

There are several views in mammography but the Cranio-Caudal (CC) view and Medio-Lateral Oblique (MLO) view are commonly used for cancer detection analysis. MLO view covers a larger area than a CC view.

The MLO view allows visualization of the largest amount of breast tissue. A technically adequate exam has the nipple in profile, allows visualization of the inframammary fold and includes the pectoral is muscle extending down to the posterior nipple line (an oblique line drawn straight back from the nipple).

The Cranio Caudal **(CC)** view is the other standard view used in every screening exam. A technically adequate CC view will include as much breast tissue as possible. If you measure straight back from the nipple, the value you get should be within 1cm of measuring the posterior nipple line on the MLO view.

1.4 BREAST CANCER STATISTICS.

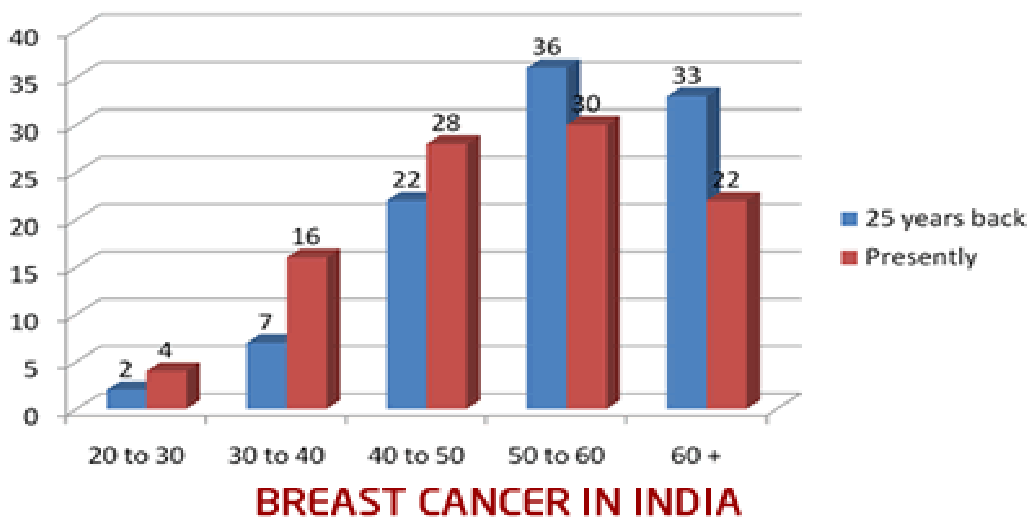


Fig.1 breast cancer statistics in India

In Fig.1, the x axis represents the age and the y axis represents the percentage of the affected people.

SOFTWARE USED

MATLAB R2012b

CHAPTER 2

LITERATURE SURVEY

This survey deals with all the feature extraction methods and the fusion techniques applicable for both MLO and CC views. The survey also proves that the multi view information provides more adequate information rather than a single view which improves the performance of accuracy.

I COMBINING FEATURES FROM MULTI VIEWS

A method proposed by Shalini gupta and Mia.k. markey 2005 [1] investigated the correspondence in Haralicks texture from MLO and CC view mammograms. Features were extracted using Spatial Gray Level Dependence (SGLD) matrices. Thirteen features viz., energy, correlation, inertia, entropy, inverse difference moment, sum average, sum variance, sum entropy, difference average, difference variance, difference entropy, and two information measures of correlation calculated from both the MLO and CC views were considered. After that the correlation coefficient of those features were calculated for both the views. A comparison was made between the correlated features between the two views and is taken as the first part of the study. Then a different subset of data were collected from two views and a comparison of correlation of the features were done as the second part. In this part, comparisons of feature correlations, comparisons of mass and calcification lesions, variation of correlation with distance and comparison of benign and malignant lesions were done. DDSM images are used. They observed that the texture features from MLO and CC view mammograms are less strongly correlated for classification than masses and less strongly correlated for benign than malignant. The result inferred that the inclusion of texture features from multiple views in a CAD ALGORITHM may impact the accuracy of diagnosis of classification and benign lesions. The effects in their study are specific only to DDSM but not to all type of digital mammograms.

A new method was proposed by Belal k elfara and Ibrahim S.I Abuhaiba [4] 2013 mainly to built the CAD model to discriminate between cancers benign and healthy parenchyma[4]. They used SCLGM (Square Centroid Line Gray Level Distribution Method) from the smallest square that includes the segmented mass with zero background as shown in Fig.2. and four centroid lines that pass through the square's center point.

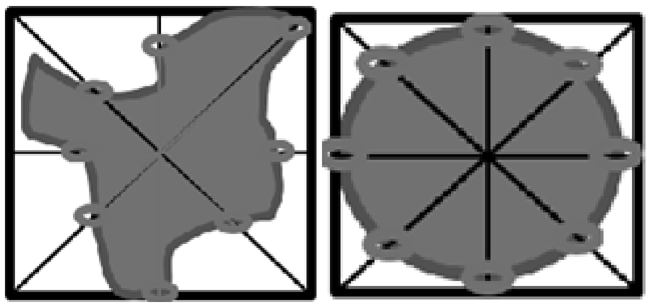


Fig2 .SCLGM method

They have extracted 145 features related to mean, variance and its difference vector, standard deviation and its difference vector, mean absolute deviation (MAD) and its differences, skewness (SK) and kurtosis(K). Then they applied sequential forward selection(SFS) and Genetic Algorithm(GA) for selection of features and formed the four feature sets namely, 1.Sequential Forward Features (SFF), selected by SFS technique, 2.Genetic Algorithm Features (GAF), selected by GA, 3.Union of Selected Features (USF): USF contains the union of features of SFF and GAF and 4.Intersection of Selected Features (ISF), represents the shared features between SFF and GAF. They used 610 mammographic images from DDSM as three datasets ie.410 images for training and two test sets with 100 images each. All the procedures were implemented for both CC and MLO views.

There are 27and 18 features were selected by SFS and GA respectively. There are 31 USF and 13 ISF. Each of these selected features were classified individually by artificial neural network with back propagation algorithm.

JC.FU et(al) 2005 [5] introduced the two stage detection procedure for the detection of breast cancer. In the first stage, the mammograms were preprocessed to enhance and locate the shape of the suspected micro calcification. In the second stage, 61 features (44 textural, 15 spatial and 2 spectral) were extracted for suspected micro calcification and Sequential Forward Search (SFS) was adopted to select the input vector sensitive only to the micro calcification. The GRNN (General Regression Neural network) and SVM (support vector machine) were used to classify the selected features. This proposed method was tested on the Nijmegen database. 7531 images are taken here in which 3674 were used for training data set.1838 for validation data set and 1839 for test data set. The SFS selection results in increase of Az value by 97.00 to 98.00 in SVM and 96.00 to 97.80 in GRNN.

The method proposed by shalini gupta, David Zhanga, Mehul P.Sampat, Mia K. Markey, [7] describes the combining information from MLO and CC mammographic views for the Computer Aided Diagnosis (CAD_x) algorithm. This approach involves in improving the CAD_e (Computer Aided Detection), also CAD_s(Computer Aided Diagnosis) by combining the features of two views with LDA(Linear Discriminant Analysis) that is taken individually and combining it in the feature selection level. The output classification level is done for classifying benign and malignant tumors. They used DDSM images in this approach. The region of interest (ROI) containing the lesion as delineated by a radiologist is provided in the database. From each ROI 260 different Haralick's texture features at different orientations and distances were extracted. Data set of 139 benign and 143 malignant lesions are trained and tested here. Experiments on 3 levels like combining the feature set from 2 views and with adding any additional feature to train a LDA model, combining the 2 feature set and new feature set obtained to train LDA model was added with the additional feature from the opposite view, and by averaging the feature values of MLO and CC view are taken in this method. The method also involves the output classification by combining both the views in five configurations considered like output_ AVG, output_ MIN, output_ MAX, output_ LDA and output_ PROD. various levels of LDA model are described and experiments were done to obtain better $A_{z..}$ The 5 levels of output combinations were obtained in the training phases. By combining both region, the better idea was obtained. Finally, the performance of multi-view reveals that they are at least same as single-view but the output classification proves better performance in multi-view than in the single view.

Another method done by shalini gupta, M.S., Priscilla F. Chyn, Mia K. Markey [8] deals with the performance analysis of CAD algorithm based on the Breast Image Report And Data System (BI-RADS). Different kinds of MLO and CC views are taken for this method. DDSM of 1626 cases are taken using kappa statistics. This study has two parts which includes investigation of (BI-RADS)^{TM}, (BI-RADS)^{TM} assessment and subtlety rating of the mammographic lesion in the first part. The (BI-RADS)^{TM} assessment descriptors are 'negative', 'benign finding', 'probability benign finding', 'suspicious abnormality', or 'highly suggestive of malignant' which are ordered from 1 to 5 respectively. The degree of agreement between MLO and CC view assessed by the kappa analysis as follows. A kappa value of less than or equal to 0.20 denotes slight agreement; 0.21 to 0.40, fair agreement; 0.41 to 0.60, moderate agreement; 0.61 to 0.80, substantial agreement; and 0.81 to 1.00, almost perfect agreement. The selection of cases for the LDA analysis is limited to those with low agreement between views. Thus all cases with identical descriptors for the two views were removed and hence the remaining 115 (51 malignant and 64 benign) were used in the LDA analysis. Encoding of the descriptors with the numerical value from 1 to 5 in the increasing order of the likely hood of malignancy was carried out. The subtlety gives the impression of the lesion. The images are taken to match with MLO and CC view for their lesion by their case number, side(R/L), lesion type (mass/micro-classification) and pathology (benign/ malignant). Linear classification is used to classify benign and malignant cases. The output lies between 0 and 1 that indicates likely hood of malignancy. (BI-RADS) descriptors of mass shape and mass margin and the patient age are taken as features for the LDA classifier. Leave-one-out sampling technique is used for the training cases. Implementation of LDA classifier with different architecture were employed. First set was implemented using features from any one view. In the 2nd set the output of 2 LDA classifiers are combined which are trained on feature from MLO and other trained from CC view. This methodology gives the final values of MAX, AVG, MIN of 2 individual classifiers. BOTH from 2 set of four descriptors as input to a single classifier (i.e.,) 2 from MLO and 2 from CC view. Receiving Operator Characteristics (ROC) curve methodology was used to evaluate the performance of classifier. The classes are differentiated by the area under the ROC that quantifies the efficacy of the classifier. The radiologist should report the descriptors separately from the MLO and CC view for this proposed method. The authors have shown only the small amount of improvement in the performance of (Bi-RADS) over the CAD_x system. A cost effective additional research is the future work of this method.

A new technique was developed by Berkman Sahinera et al [9] to improve the accuracy of computerized micro calcification detection by using the joint two-view information on cranio caudal (CC) and mediolateral-oblique (MLO) views. The cluster candidates were characterized by the single view classifier and the similarity classifier and the outputs of their results were fused together to get a true micro calcification results. Morphological features (such as size, mean density, shape, and contrast) extracted from each micro calcification cluster ,texture features extracted from the ROIs using the second-order statistics provided by the spatial gray-level dependence (SGLD) matrix and features derived from the CNN scores of the micro calcifications such as the average, standard deviation, maximum, and minimum of the individual micro calcification scores within the cluster which has been used extensively for the detection of micro calcifications were taken into account.

The single view classifier used the features from the clusters taken from MLO or the CC view where as the similarity classifier jointly uses the features from both the clusters of MLO and CC view of the same candidate . A high classification accuracy with small number of features was achieved by the leave-one-case out method and the area under the receiver operating characteristic curve (*Az*). The step wise selection includes 45 morphological features, 78 textural features and 4 CNN score features. This method searches the member of cluster pairs based on the NODs (nipple to object distance) of the CC and MLO views respectively. From the feature set two sets of features were generated for the similarity classifier that was also used by the single view classifier. The first set involves the corresponding feature of squared difference between them of two objects in a pair. The second set has the corresponding features of their average. Scores has been generated for cluster pairs and for the individual clusters for similarity classifier and the single view classifier respectively. Then the both scores were combined together. FROC (free response ROC) curves can be obtained for detection of the micro calcification clusters using constant paired-cluster threshold, and varying the decision threshold on the similarity scores . During the testing stage, when the scores of the two classifiers were fused in the joint view fusion , it results in the much high sensitivity for the detection of micro calcification. Finally to see the improvement of this joint view fusion, JA-FROC curve were analyzed and the results proved to have greater improvement in detecting both malignant and benign cases.

II MULTI-CLASSIFIER FUSION BASED TECHNIQUES

The method established by Li sun et (al) 2010 [2] involves the fusion of multi view mammograms leads to improve the accuracy and the robustness of classification. It also decrease the false positive rates. They considered a new multi classifier fusion approach by employing the measuring value as the inputs of the agents in a multi agent frame work. These measuring values were employed first for doing the multi classifier fusion. Multi agent fusion algorithm was applied here. Five different shape features circularity, the standard deviation of gray levels of boundary, variance of radius, gradient and compactness two contour features are used to describe the radius of irregular contours of edge-sharpness, i.e. standard deviation of the second derivative of the radius of the mass and local standard deviation of the second derivative of the radius of the mass were selected for this method. Four individual classifiers say Naive Bayes (NB),IBk, Multilayer Perceptron (MP) and Random Forest (RF), are used. And four fusion methods like multi-agent (MA),Average (Ave), Vote (Vot) and Average based on weight (WeiA) are used. From the DDSM data set, two groups of ROI were taken and in both the groups they observed that the classification accuracy is higher in multi agent (MA) when compared to all other fusion methods. Also the variance of accuracy (VARAC),variance of true positive (VARTN) and the variance of true negative (VARTP) are found to be smaller in MA. Finally they proved that MA increases the performance of classification in contrast to all the other methods.

A novel scheme based on multi-view information fusion [3] was developed by Lihuali et(al) in 2010 to improve the accuracy and to reduce the false positive rates. 304 ROI's of and 152 multi-view pairs from DDSM are were utilized in their work. Those pairs were divided into three groups with each group contains equal number of the benign and malignant cases. First group has 52 pairs and the masses in both views could be easily identified by radiologists. The second group contains 56 pairs and the masses are easy to identified only in one view. The third group consists of 44 pairs and the masses in both views are difficult to identify**.**

They have introduced two contour features to describe the radius of irregular contours of edge-sharpness, standard deviation of the second derivative of the radius of the mass and local standard deviation of the second derivative of the radius of the mass. In addition to these two new features, they have applied five shape features such as circularity, the standard deviation of gray levels of boundary, variance of radius, gradient and compactness to represent the masses. Four individual classifiers namely NB (Naive bayes) KNN (k-nearest-neighbor) MLP (multilayer perceptron), RF (random forest) were used to classify masses as benign and malignant. The multi-view classification result is defined as the average of the classification results in two mammographic views, which are obtained using an individual classifier. classification using seven features proves higher accuracy than using five features. The NB compute the posterior probability of the class y. It estimates the class conditional probability for the same. NB classifiers are robust to noise points and irrelevant attributes. In KNN ,the training examples that are relatively similar to the attributes of the test examples are found. They do not require a model building. A MLP which is specific type of multilayer feed forward network model is utilized here as a basic model and for its good non linear mapping ability. The RF algorithm combines the prediction made by multiple decisions trees that was built by the boot strap sample of the original data. A comparison is made between five and seven features respectively using the above four classifiers and proved that MLP performs best in giving the accuracy, True Positive and True Negative rates in using seven features than in five features.

III PERFORMANCE IMPROVEMENT TECHNIQUES IN MASS RETIVAL USING MULTI VIEW INFORMATION

A method proposed by Wei Liu et al [11] deals with the mass retrieval approach based on the multiple view information. A multi view example query is taken instead of a single view. several visual features, distance similarity measures (Euclidian distance) and non distance similarity measures (KNN regression model)were used here. They employed CBMIR(content based medical image retrieval) technique that uses the QBE (query based example) as the classical paradigm. A signature like color, texture, shape or layout feature was extracted from the query image or ROI of the system and it was compared to the signatures of the previously computed image or ROIs in the data base and a closest matches were returned.

The fourteen selected features includes average pixel value in the breast area, average local pixel value fluctuation in the breast area, standard deviation of the local pixel value fluctuation in the breast area, region conspicuity, normalized mean radial length of a region, standard deviation of radial length, skew of radial length, shape factor ratio, standard deviation of pixel values inside the mass region, standard deviation of the gradient of boundary pixels, skew of the gradient of boundary pixels, standard deviation of pixel values in the surrounding background, average local pixel value fluctuation in the surrounding background, normalized central position shift.

Wavelet-based multi-scale fractal dimension feature was used for comparison. Haar wavelet was adapted that results in a 18 dimensional feature vector. The similarity measures includes cosine distance, Euclidean distance and the weighted k-Nearest Neighbor(k-NN) regression algorithm. The similarity measures depends on the following viz., AVG-r (the smaller the *AVG-r* is, the better the retrieval accuracy), AVG –p(the bigger the *AVG-p* is, the better the retrieval accuracy) and ANMRR (Average Normalized Modified Retrieval Rank) is the average of the normalized modified retrieval rank over all queries. The smaller the *ANMRR* is, the better the retrieval accuracy .The training data collection was done in such a way that the 126 images in the database were partitioned into 10 different groups by the k means clustering method based on the visual features and then a total of 300 intra-group pairs were randomly selected, of which each pair was formed by images from a same group. Finally, a total of 300 inter-group pairs were randomly selected, of which each pair was formed by images from two different groups. For k-NN regression model, we considered an image to be truly relevant to a query if its corresponding observer score is larger than a preselected threshold *T*. Here *T =6* was used.The retrieval results of the multi view was done by taking the fusion factor “F” and the retrieval results “R” from both the MLO and CC masses respectively. Four retrieval modes such as single-view MLO retrieval, single-view CC retrieval, multi-view corresponding retrieval and multi-view arbitrary retrieval modes were implemented here. Different retrieval measures were analyzed using cosine distance, Euclidean distance and the K NN regression method and the MCR (Multiview corresponding retrieval) proves the better results of all.

IV 3D FEATURE EXTRACTION FROM MLO AND CC VIEWS

S.M.Vijayarajan 2014 [6] developed a CAD system for the detection of masses and architectural distortions by utilizing the correspondence between MLO and CC projections[7]. The breast mass in both CC and MLO were detected separately and merged to get the 3D view. This was done in four steps 1.2D feature detection, 2.MLO and CC feature mapping in 3D, 3.MLO and CC component merge and 4.3D image generation. The parabolic boundary, the pectoral boundary and nipple positions of MLO and CC are detected as 2D features. From these features 3D co-ordinates are generated. Finally 3D mammogram with mass is generated from the 3D feature set and the mass component location value in MLO and CC.

The method proposed by Werapon Chiracharit and Rachada Kongkachandra [10] insisted in using the 3D shape and distribution features extracted from the view correspondence between MLO and CC mammograms. The proposed method found a new point of view to solve the mammographic view problem. The correspondence of both MLO and CC views were calculated first and then corresponding features of the matched clustered micro calcifications were extracted. The misrepresentations of the conventional classification system leads to the problems like the loss of the information depth of the micro calcification due to the different imaging poses. It leads to the confused informations like the shape distortion, disagreeable environment and contradictory arrangement. Here features like size, shape and distribution were taken in to account to classify the clustered micro calcification into benign or malignant.

The malignant type is typically different from the benign in shape, contour, size, number or distribution in the cluster. Totally twelve different features were selected here. Benign cases will have homogeneous rounded shape, sharply outline generally larger size, solitary or numerous number, diffusely distribution, and very fine and dense density. While the malignant is dot-like or elongated form, fragment with irregular, tiny, very numerous clustered and differ granular. For finding the 3D distribution features, they employed a stereo constraint to match the micro calcification pair from both the MLO and CC view. After this 3D dimensional features were extracted and they includes Volume of the Cluster (VC), Cluster Density (D), Distances between Micro calcification Pairs in the Cluster. This is followed by reconstructing a volume of single micro calcification cluster from two cross-section image (CC and MLO views) with different angle. After that the centroids of matched micro calcification in both CC and MLO view was determined followed by finding the differences in them. Micro calcification in the MLO image can be removed by the new position that has the same co ordinates of the centroid in the CC view image. This helps in calculating the volume and area of circumference of the single micro calcification. The eight 3D shape features found here are Fourier Descriptor (FFavg and FFSD),Compactness (Cavg and CSD), Volume or size (Vavg and VSD), Density of Intensity (Iavg and ISD). For classification three layered feed forward neural network with back probogation algorithm was used as a classifier. Mammographic view matching was done by taking the ROIs which are added with the ROIs of the translation points of both the MLO and CC views. Finally the classification of the micro calcification clusters was done which has the true positive and the false positive for the cancer cells and true negative and false negative for the non cancerous cells. The proposed method shows higher sensitivity and specificity 96 percent and 91 percent respectively than the single view.

CHAPTER 3

IMAGE DATABASE

Two image databases are used for this work. They are

1. DDSM
2. INBREAST

3.1 DDSM

The widely used database is the Digital Database for Screening Mammography (DDSM). It is the largest public database, with 2620 cases including two images from each breast (MLO and CC), for a total of 10,480 images, with all types of findings from normal images to images with benign and malign lesions. The database is divided into 43 volumes, and each volume is divided in a number of studies. It contains volumes with normal cases, volumes with cases containing benign abnormalities, and volumes containing cases with cancerous abnormalities.

3.2 INBREAST DATABASE

The image matrix was 3328 x 4084 or 2560 x3328 pixels, depending on the compression plate used in the acquisition (according to the breast size of the patient). Images were saved in the DICOM format. All confidential medical information was removed from the DICOM file, according to Supplement 55 of the DICOM standard; the correspondence between images of the same patient is kept with a randomly generated patient identification and collected from which 90 have two images (MLO and CC) of each breast and the remaining 25 cases are from women who had a mastectomy and two views of only one breast were included. This sums to a total of 410 images. Eight of the 91 cases with 2 images per breast also have images acquired in different timings.

The database includes examples of normal mammograms, mammograms with masses, mammograms with calcifications, architectural distortions, asymmetries, and images with multiple findings and their bi-rads rating. BI-RADS is an acronym for Breast Imaging-Reporting and Data System, a quality assurance tool originally designed for use with mammography.

BI-RADS Assessment Categories are:

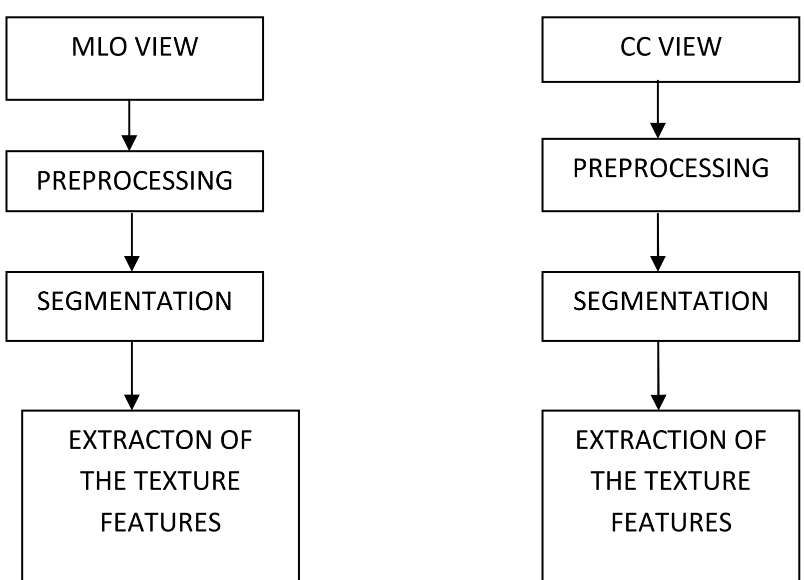
Category 0: Incomplete
Category 1: Negative
Category 2: benign findings;
Category 3: probably benign findings;
Category 4: suspicious findings;
Category 5: a high probability of malignancy; and
Category 6: proved cancer
In case of categories 4 and 5, a biopsy is needed to exclude or confirm malignancy.

CHAPTER 4

PROPOSED WORK

The first step of the proposed work includes the collection of MLO and CC view mammograms of DDSM data bases and preprocessing them. In the second step, preprocessed images have to be segmented and the texture features are to be extracted from them. Finally, classification of the images are done using the above features and performance measures are calculated.

The proposed work has the following process flow:



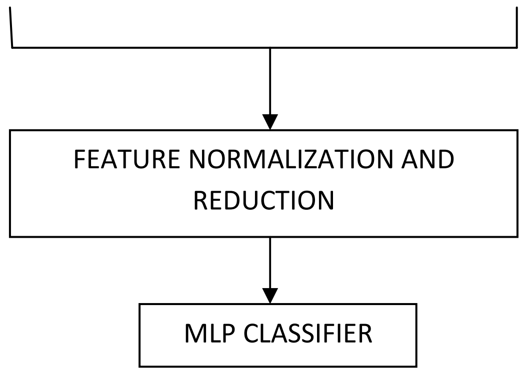


Figure 3 . Overall block diagram of proposed method.

- The pre processing of the images includes the following steps:

  1.Contrast enhancement

  2.Label removal and

  3.Pectoral muscle removal

- The images are segmented using the Adaptive k means clustering algorithm.
- Texture features like Steerable pyramid, Gabor and GLCM (Gray Level Co Occurrence Matrix) are extracted and combined together.
- The combined features are reduced using PCA(Principal Component Analysis) technique.
- Finally, the features are used to classify the images as benign or malignant using a MLP(Multi Layer Perceptron) with back propagation learning.

CHAPTER 5

PREPROCESSNG AND SEGMENTATION

5.1 Pre processing

Digital mammograms are medical images that are difficult to be interpreted, thus a preparation phase is needed in order to improve the image quality and make the segmentation results more accurate. It aims at separating the breast tissue from the background of the mammogram and it includes two independent segmentations. The first step segments the background region which usually contains annotations, labels and frames from the whole breast region, while the second step removes the pectoral muscle portion (present in Medio Lateral Oblique (MLO) views) from the rest of the breast tissue.

5.1.1 Contrast enhancement

The first part involves the contrast enhancement of the images in CC and MLO views, if required. The DDSM database has low contrast images and is in need of contrast enhancement.

5.1.2 Label removal

The next part is the removal of labels from the images. The label is removed, based on the area of the individual objects in the image. The areas of such objects are then calculated. The labels occupy only smaller area than the breast portion is removed by considering the maximum area condition. Thus the breast portion having maximum area is only present in the output image. Based on the area calculation, the desired part occupying larger area is separated out.

5.1.3 Pectoral muscle removal

The final stage involves the pectoral muscle removal. Pectoral muscles are the regions in mammograms that contain brightest pixels. These regions must be removed before detecting the tumor cells so that mass detection can be done efficiently. The Pectoral muscles located at the left top corner or right top corner which depends on the left or right view of the image. The label removed output is given as the input for removing the pectoral muscle. Initially the pectoral muscles are detected before removing it. For this searching for non zero pixels and zero pixels in the image. It is done by calculating the values of first five rows and last five rows and comparing them. By assigning a first five rows are greater than the last five rows then it is assumed that the breast region is right oriented otherwise it is left oriented and it is to be converted into right. Then the starting and ending point of the breast region is found and it is joined by using a straight line. Then the space above the line is covered by pectoral muscle and which is removed by assigning zero pixels to those values. The resulting image contains the pectoral muscles are removed from the original image.

5.2 . Segmentation

Segmentation is the process of partitioning the image into multiple segments. The goal of segmentation is to simplify and change the representation of an image into something that is more meaningful and easier to analyze. After the preprocessing has been completed, the images are to be segmented to diagnose the suspicious region by using suitable algorithm explained below.

5.2.1. Adaptive k means clustering algorithm

After the preprocessing of the images are completed, adaptive k means clustering algorithm has to be done for all the images. k-means clustering is a method of cluster analysis that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. The steps involved in this algorithm is as follows:

1) Pick k cluster centers, either randomly or based on some heuristic.

2) Assign each pixel in the image to the cluster that minimizes the distance between the pixel and the cluster center.

3) Re-compute the cluster centers by averaging all of the pixels in the cluster.

4) Repeat last two steps until convergence is attained (e.g. no pixels change clusters.

The minimum distance between the pixels can be calculated using the formula:

$\arg\min \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\|^2$ where $\boldsymbol{\mu i}$ is the mean. The closest mean calculation use the formula ,

$$S_i^{(t)} = \left\{ \mathbf{x}_j : \left\| \mathbf{x}_j - \mathbf{m}_i^{(t)} \right\| \leq \left\| \mathbf{x}_j - \mathbf{m}_{i^*}^{(t)} \right\| \text{ for all } i^* = 1, \ldots, k \right\}$$

and the new mean can be calculated by using the formula ,

$$\mathbf{m}_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{\mathbf{x}_j \in S_i^{(t)}} \mathbf{x}_j$$

.

CHAPTER 6

FEATURE EXTRACTION METHODS

6.1 GABOR FEATURE EXTRACTION

The features from the segmented regions are to be extracted using the Gabor filters tunable band pass filters with multi-scale, multi-resolution and have selectivity for orientation, spectral bandwidth and spatial extent. Gabor filters are employed to quantify the texture information from specific regions, tissues and internal structures of the images providing a concise representation for a richer image analysis.

Feature extraction is performed in two steps. In the first step, a new set of k images are obtained by clustering each original grayscale image using the k-means algorithm considering their pixel values. In the second step Gabor subspaces are extracted from each image generated in the first step. Then the two features, mean and standard deviation are computed from each Gabor subspace. The final feature vector is obtained by combining the features extracted from the set of clustered images obtained in the first step. Finally the feature vectors obtained from CC and MLO views are concatenated to get final combined feature vector. The size of the feature vector from one view of mammogram will be KxUxVxF, where K – number of clustered images from one original image, U&V - the number of scales and number of orientations used in Gabor filter and F- number of features captured from each Gabor subspace. Combining the features computed from both MLO and CC views doubles the size of the feature vector.

6.2.GLCM FEATURE EXTRACTION

GLCM (Gray level co occurrence probability matrix) is an array of modeling the information of a texture as a two dimensional array gray level variation. Fifteen statistical measures such as auto correlation, contrast, homogeneity, cluster prominence, correlation , sum entropy ,etc., were taken in to account. The GLCMs are stored in a i x j x n matrix, where n is the number of GLCMs calculated usually due to the different orientation and displacements used in the algorithm. GLCM is also called as *Gray level Dependency Matrix*. It is defined as “A two dimensional histogram of gray levels for a pair of pixels, which are separated by a fixed spatial relationship.” GLCM of an image is computed using a displacement vector d, defined by its radius δ and orientation θ. The extracted GLCM features are as follows:

- Autocorrelation
- Contrast
- Correlation

- Correlation
- Cluster Prominence
- Cluster Shade
- Dissimilarity
- Energy
- Entropy
- Homogeneity
- Homogeneity
- Maximum probability
- Sum of squares
- Sum average
- Sum variance
- Sum entropy
- Difference variance
- Difference entropy
- Information measure of correlation1 and 2
- Inverse difference (INV)
- Inverse difference normalized (INN)
- Inverse difference moment .

6.3 STEERABLE PYRAMID

The Steerable Pyramid is a linear multi-scale, multi-orientation image decomposition which provides a useful front-end for image-processing and also for computer vision applications. The decomposition is done in such a way that the images are separated in to low and high pass sub bands and a low pass sub band and this sub band gets sub sampled by a factor of 2 in the X and Y directions. A power map is created as the final stage that represents the responses of the oriented filters at each scale.8 filters are used in 3 levels that results in 24 power maps. The final result is a 48 element feature vector. The steerable pyramid features were extracted from CC and MLO views.

A steerable pyramid, shown in figure 12, is an implementation of a multi-scale, multi-orientation band-pass filter bank used for applications including image compression, texture synthesis, and object recognition. It can be thought of as an orientation selective version of a Laplacian pyramid, in which a bank of steerable filters are used at each level of the pyramid instead of a single Laplacian of Gaussian filter. The basis functions of the steerable pyramid are Kth-order directional derivative operators (for any choice of K), that come in different sizes and K+1 orientations. As directional derivatives, they span a rotation-invariant subspace, and they are designed and sampled such that the whole transform forms a tight frame. An example decomposition of an image of a white disk on a black background is shown to the right. This particular steerable pyramid contains 4 orientation sub bands, at 2 scales. The smallest sub band is the residual lowpass information.

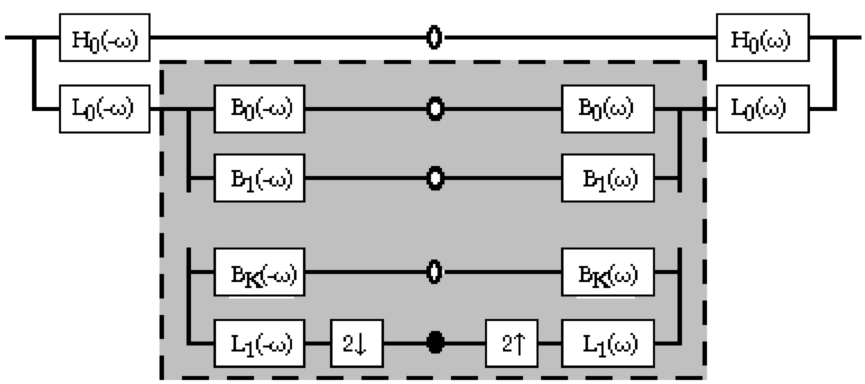


Figure 4: Steerable Pyramid

The block diagram for the decomposition (both analysis and synthesis) is shown in Figure 4. Initially, the image is separated into low and highpass sub bands, using filters L0 and H0. The lowpass sub band is then divided into a set of oriented bandpass sub bands and a low(er)-pass sub band. This low(er)-pass sub band is subsampled by a factor of 2 in the X and Y directions. The recursive (pyramid) construction of a pyramid is achieved by inserting a copy of the shaded portion of the diagram at the location of the solid circle (i.e., the lowpass branch).

CHAPTER 7

FEATURE REDUCTION TECHNIQUE

Principal component analysis (PCA), which is based on statistical procedures was employed. The extracted features from the Gabor filter, GLCM and the steerable pyramid are combined together and reduced using PCA for the benign and malignant cases of the DDSM data base.

7.1. PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to (i.e., uncorrelated with) the preceding components. The principal components are orthogonal because they are the eigenvectors of the covariance matrix, which is symmetric. PCA is sensitive to the relative scaling of the original variables.

In PCA, initially the n dimensional vector say x = [x1,x2,…..xn]^T is transformed to the y vector y = x (A-m_x). The mean vector m_x is calculated by $m_x = E\{X\} = \frac{1}{K}\sum_{K=1}^{K} X_K$. The covariance matrix C_x determined by the formula C_x=E{(x-m_x)(x-m_x)^T} $= \frac{1}{k}\sum_{k=1}^{k} x_k\, x_K^T - m_x m_x^T$. The rows in the matrix A are arranged Eigen vectors of the Eigen values accordingly in the descending order.The size of the C_x matrix is n x n. The co variance between the input and the output vectors x_i and x_j are obtained by the equation C_x (i,j) =E{(x_i-m_i) (x_j-m_j)}. Since the rows of the matrix A is orthonormal, The inversion of PCA is possible according to the equation x = A^T y + m_x. the PCA transformation is shown below

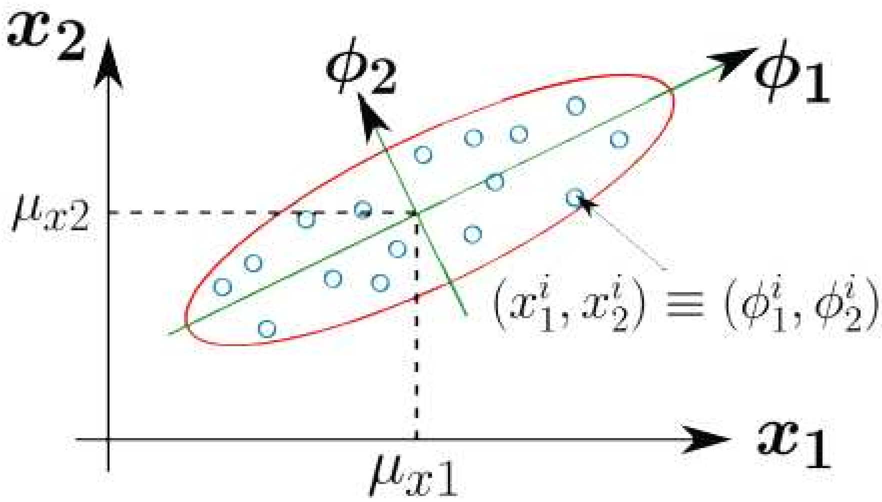


Fig.5 PCA transformation

The feature values of the steerable pyramid, Gabor and the GLCM are combined together and by using the PCA they are reduced to 100 values i.e., those 100 significant values are considered for the classification process.

CHAPTER 8

CLASSIFICATION

Classification is grouping of the objects or things that are similar. The benign and the malignant type images are classified using a multilayer perceptron classifier (MLP) or back propagation neural network (BPN) which is an important for the data mining tool for the classification algorithm. Basic Neural Network (NN) is composed of three layers, input, output and hidden layer. Each layer can have number of interconnected nodes (neurons which are mimicking the biological neuron).Nodes from input layer are connected to the nodes from hidden layer. Nodes from hidden layer are connected to the nodes from output layer. Those connections represent weights between nodes.

8.1. NEURON

Human brain has over 100 billion interconnected neurons. Most sophisticated application have only tiny fraction of that. It can only be imagined how powerful NN with this number of interconnected neurons would be. Neurons use this interconnected network to pass information's with each other using electric and chemical signals. Although it may seem that neurons are fully connected, two neurons actually do not touch each other. They are separated by tiny gap call Synapse. Each neuron process information and then it can "connect" to as many as 50 000 other neurons to exchange information. If connection between two neuron is strong enough information will be passed from one neuron to another.

Each neuron is not very bright but put 100 billion of them together and let them talk to each other, then this system becomes very powerful. A typical neuron would have 4 components Dendrites, Soma, Axon and Synapse. Dendrites gather inputs from other neurons and when a certain threshold is reached they generate a non-linear response (signal to other neurons via the Axon).

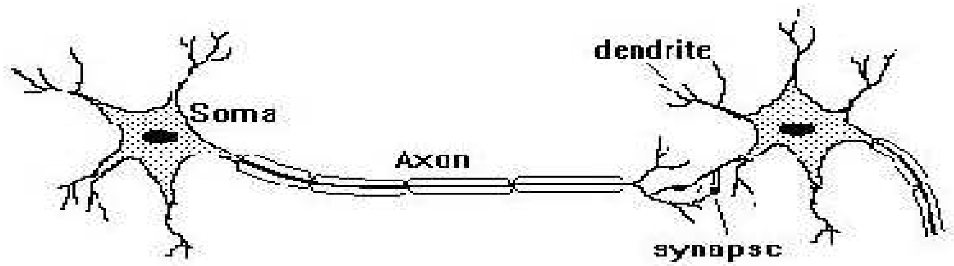


Fig 6.A simple neuron

8.2. MLP NEURAL NETWORK

A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. A MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called back propagation for training the network. MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable.

8.2.1 Activation function

If a multilayer perceptron has a linear activation function in all neurons, that is, a linear function that maps the weighted inputs to the output of each neuron, then it is easily proved with linear algebra that any number of layers can be reduced to the standard two-layer input-output model. What makes a multilayer perceptron different is that some neurons use a *nonlinear* activation function which was developed to model the frequency of action potentials, or firing, of biological neurons in the brain. This function is modeled in several ways.

The two main activation functions used in current applications are both sigmoid , and are described by

$$y(v_i) = \tanh(v_i) \quad \text{and} \quad y(v_i) = (1 + e^{-v_i})^{-1},$$

in which the former function is a hyperbolic tangent which ranges from -1 to 1, and the latter, the logistic function, is similar in shape but ranges from 0 to 1. Here $y_i$is the output of the ith node (neuron) and $v_i$is the weighted sum of the input synapses. Alternative activation functions have been proposed, including the rectifier and soft plus functions. More specialized activation functions include radial basis functions which are used in another class of supervised neural network models.

8.2.2 Layers

The multilayer perceptron consists of three or more layers (an input and an output layer with one or more *hidden layers*) of nonlinearly-activating nodes and is thus considered a deep neural network. Each node in one layer connects with a certain weight $w_{ij}$to every node in the following layer. Some people do not include the input layer when counting the number of layers and there is disagreement about whether w_{ij}should be interpreted as the weight from i to j or the other way around.

8.2.3 Terminology

The term "multilayer perceptron" often causes confusion. It is argued the model is not a single perceptron that has multiple layers. Rather, it contains many perceptrons that are organised into layers, leading some to believe that a more fitting term might therefore be "multilayer perceptron network". Moreover, these "perceptrons" are not really perceptrons in the strictest possible sense, as true perceptrons are a special case of artificial neurons that use a threshold activation function such as the Heaviside step function, whereas the artificial neurons in a multilayer perceptron are free to take on any arbitrary activation function. Consequently, whereas a true perceptron performs binary classification, a neuron in a multilayer perceptron is free to either perform classification or regression, depending upon its activation function.

The two arguments raised above can be reconciled with the name "multilayer perceptron" if "perceptron" is simply interpreted to mean a binary classifier, independent of the specific mechanistic implementation of a classical perceptron. In this case, the entire network can indeed be considered to be a binary classifier with multiple layers. Furthermore, the term "multilayer perceptron" now does not specify the nature of the layers; the layers are free to be composed of general artificial neurons, and not perceptrons specifically. This interpretation of the term "multilayer perceptron" avoids the loosening of the definition of "perceptron" to mean an artificial neuron in general

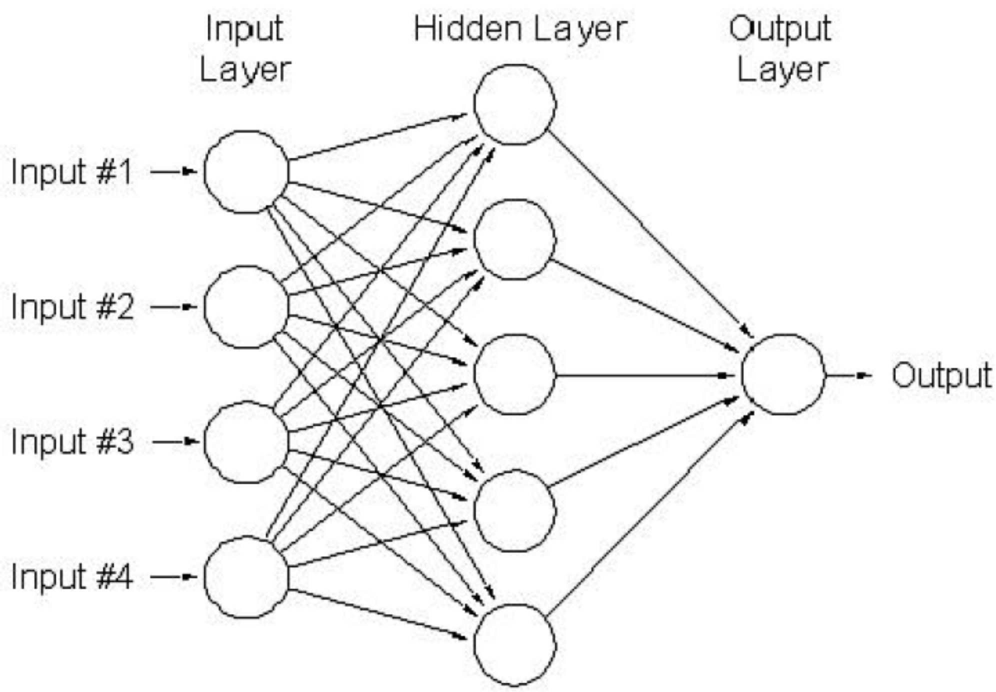


Fig.7. A simple neural network

8.3. BACK PROPAGATION (BP) ALGORITHM

One of the most popular NN algorithms is back propagation algorithm. back propagation algorithm is used to compute the necessary corrections. The algorithm can be decomposed in the following four steps:

i) Feed-forward computation

ii) Back propagation to the output layer

iii) Back propagation to the hidden layer

iv) Weight updates

The algorithm is stopped when the value of the error function has become sufficiently small.

Learning occurs in the perceptron by changing connection weights after each piece of data is processed, based on the amount of error in the output compared to the expected result.

This is carried out through back propagation, a generalization of the least mean squares algorithm in the linear perceptron.

We represent the error in output node j in the nth data point (training example) by $e_j(n) = d_j(n) - y_j(n)$, where d is the target value and y is the value produced by the perceptron. We then make corrections to the weights of the nodes based on those corrections which minimize the error in the entire output, given by

$$\mathcal{E}(n) = \frac{1}{2} \sum_j e_j^2(n).$$

Using gradient descent, we find our change in each weight to be

$$\Delta w_{ji}(n) = -\eta \frac{\partial \mathcal{E}(n)}{\partial v_j(n)} y_i(n)$$

where y_i is the output of the previous neuron and η is the *learning rate*, which is carefully selected to ensure that the weights converge to a response fast enough, without producing oscillations. In programming applications, this parameter typically ranges from 0.2 to 0.8.

The derivative to be calculated depends on the induced local field v_j, which itself varies. It is easy to prove that for an output node this derivative can be simplified to

$$-\frac{\partial \mathcal{E}(n)}{\partial v_j(n)} = e_j(n)\phi'(v_j(n))$$

where ϕ' is the derivative of the activation function described above, which itself does not vary. The analysis is more difficult for the change in weights to a hidden node, but it can be shown that the relevant derivative is

$$-\frac{\partial \mathcal{E}(n)}{\partial v_j(n)} = \phi'(v_j(n)) \sum_k -\frac{\partial \mathcal{E}(n)}{\partial v_k(n)} w_{kj}(n).$$

This depends on the change in weights of the kth nodes, which represent the output layer. So to change the hidden layer weights, we must first change the output layer weights according to the derivative of the activation function, and so this algorithm represents a *back propagation of the activation function*.

8.4.PERFORMANCE MEASURES:

The performance measures such as accuracy, specificity and sensitivity were taken for both the DDSM and INBREAST data bases and they are compared. The mathematical expressions for all the performance measures are listed below:

SENSITIVITY = TP/(TP+FN)

SPECIFICITY = TN/(TN+FP)

ACCURACY = (TP+TN)/(TP+FP+TN+FN)

TP = true positive

TN = true negative

FP = false positive

FN = false negative. Totally 68 images that includes 25 benign cases and 43 malignant cases were taken in to account and they are used for the classification process.40 images are used for the training, and the remaining are utilized for the testing purpose for the DDSM image set.

CHAPTER 9

RESULTS & CONCLUSIONS

9.1 .RESULTS

The results of the proposed work which includes pre processed, clustered, and the segmented images of the DDSM dataset are shown below.

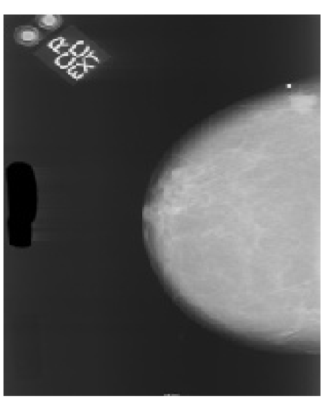


Fig 8.Input image

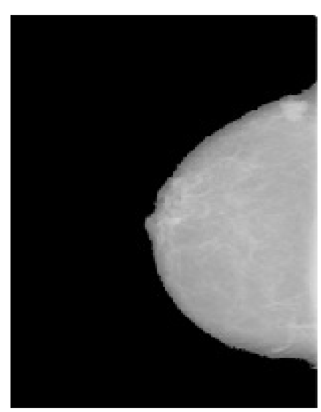


Fig.9 preprocessed image

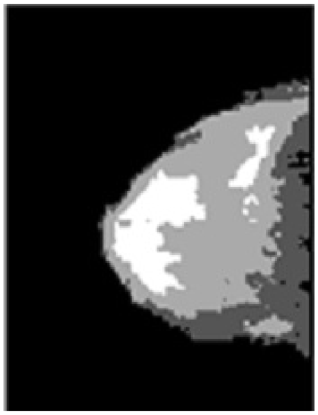


Fig.10. Clustered image

Fig.11.segmented image

The performance measures namely accuracy, specificity and sensitivity for the benign and malignant cases using the combined texture features were calculated effectively. The performance measures were calculated indivually for MLO and CC cases and they are compared with combining both the views. The results are also tabulated and shown below in the following table.1,

Table.1. Performance measures

FEATURES	ACCURACY	SPECIFICITY	SENSITIVITY
CC	88.2353	87.8788	88.7446
MLO	86.7647	87.6227	98.0796
COMBINED	91.1765	93.8776	93.8760

9.2 .CONCLUSION

The classification results shows that the accuracy of the combinations of MLO and CC view features is greater when compared to that of either MLO view features or CC view features. This work well proves that the accuracy of the classification of the benign and malignant breast masses is improved when the combination of the texture features from both MLO and CC views are combined together. Hence it is concluded that the use of combination of multi view features will improve the diagnosis capability of any CAD system.

9.3.FUTURE WORK

- The proposed work may be validated with IN BREAST and some other data bases.
- Some more texture features may be extracted and included to further improve the classification results.
- The same work may be used for Ultra Sound (US) breast images.

REFERENCES

[1] Shalini Guptaa & mia.markey **correspondence in texture features between two mammographic views** 2005.

[2] Li Sun, Weidong Xu, Lihue Li, Wei Liu, Juan Zhang, Guoliang Shao **A New Multi-view Identification Scheme for Breast Masses in Mammography Using Multi-agent Algorithm** 8th IEEE International Conference on Control and Automation ,Xiamen, China, June 9-11, 2010.

[3] Li Sun1, Lihue Li1, Weidong Xu, Wei Liu Juan Zhang, Guoliang Shao **A Novel Classification Scheme for Breast Masses Based on Multi-view Information Fusion**

[4] Belal k. Elffara and Ibrahim S.I.**Abuhaiba New feature extraction for mammogram computer aided diagnosis** , international journal of signal processing and pattern recognition vol.6 no.1 2013

[5]Zhimnin huo, Maryellen l. giger ,and Carl j.Vyborny **Computerized analysis of multiple mammographic views: potential usefulness of special view mammograms in computer aided diagnosis,** ieee transactions on medical imaging,vol.20 no.12, 2001

[6] J.C. Fu,, S.K. Lee, S.T.C. Wong, J.Y. Yen, A.H. Wang, H.K. Wu **Image segmentation feature selection and pattern classification for mammographic micro calcifications**., Computerized Medical Imaging and Graphics 29 (2005) 419–429

[7] S.M.Vijayarajan **Breast cancer detection and segmentation using multiview mammogram** academic journal of cancer research, 2014

[8] Shalini Guptaa, David Zhanga, Mehul P. Sampath, Mia K. Markey **Combining texture features from the MLO and CC views for mammographic CAD.**

[9] Shalini Gupta, M.S., Priscilla F. Chan, Mia K. Markey, **Breast cancer CAD*x* based on BI-RADS descriptors from two mammographic views.**

[10] David A. Clausi and Huang Deng **Design-Based Texture Feature Fusion Using Gabor Filters and Co-Occurrence Probabilities**, ieee transactions on image processing, vol. 14, no. 7, July 2005

[11] Wei Liu, Weidong Xu, Lihua Li, Shuang Li, Huanping Zhao **Improved Mammographic Mass Retrieval Performance Using Multi-View Information** 2010 IEEE International Conference on Bioinformatics and Biomedicine.

[12] Wenfeng Han, Jianwei Dong, Yuting Guo, **Identification of masses in digital mammogram using an optimal set of features** 2011 International Joint Conference of IEEE TrustCom-11/IEEE ICESS-11/FCST-11

[13] Jun Wei,a_ Heang-Ping Chan, Berkman Sahiner, Chuan Zhou, Lubomir M. Hadjiiski, Marilyn A. Roubidoux, and Mark A. Helvie **Computer-aided detection of breast masses on mammograms: Dual system approach with two-view analysis.**

[14] M. Lalitha, M. Kiruthiga, C. Loganathan **A Survey on Image Segmentation through Clustering Algorithm** International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064

[15] Alexander Barley, Christopher Town **Combinations of Feature Descriptors for Texture Image Classification** Journal of Data Analysis and Information Processing, 2014.

[16] Bhagwati Charan Patel and Dr. G.R.Sinha **An Adaptive K-means Clustering Algorithm for Breast Image Segmentation** International Journal of Computer Applications

[17] S.Sasikala, M.Ezhilarasi, R.Vivitadurga, S.Radhika, **"Detection and 3D Reconstruction Of Tumor Based On Mammographic Images: A Survey"** International Journal of Applied Engineering Research, Volume 9, Number 20 (2014).

[17] M. Mudrov´a, A. Proch´azka **principal component analysis in image processing.**

[18] Mirza Cilimkovic **Neural Networks and Back Propagation Algorithm.**

LIST OF PUBLICATIONS

Sasikala.S, Ezhliarasi.M, Rasheedha.A, ***"BREAST CANCER DIAGNOSIS USING BOTH MLO & CC VIEW MAMMOGRAMS: A SURVEY".*** International conference on Eletrical,Instrumentation and Communication Engineering – Recent Trends and Research Issues (ICE^2-RTRI) in SRI KRISHNA COLLEGE OF TECHNOLOGY.