ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 6, June 2017

"Development of Computer Aided Diagnosis System (CADx) for Detection of Anomalies in Breast using Textural Features with PNN classifier"

[1] Miss. Ankita Satyendra Singh, [2] Mrs. M. M. Pawar
Department of Electronics and Telecommunication Engineering
SVERI's COE Pandharpur, Solapur Maharashtra, India

Abstract: **High False Negative Rate (FNR) is a very significant problem in a Computer Aided Diagnostic System as false negative answer may lead to a very high increase in the number of deaths. The main aim of this paper lies in the development of a new Computer Aided Diagnosis (CADx) system for the proper identification of breast masses. It also focuses at extraction of textural features. The input images are pre-processed by using Adaptive Median Filter and then segmented by using Gaussian Mixture Model i.e. GMM segmentation and further are subjected to feature extraction, selection and finally classification by using PNN classifier. MIAS database is used for research purpose which contains 322 mammogram images out of which 60 images as 20 of benign, 20 malignant and 20 normal are taken into consideration for feature extraction. 22 texture features are extracted and are further classified. PNN classifier with 80-20 train-test partition is used for classification. The Sensitivity, Specificity and Accuracy obtained by the selected features are 100%, 100%, and 100% respectively.**

***Keywords*—Mammogram, Pre-processing, Adaptive Median Filter , Gaussian Mixture Model (GMM), EM algorithm, MAP algorithm, Image segmentation, Texture features, Classification, PNN classifier**

I. INTRODUCTION

Breast cancer is the cancer that develops from breast tissues. Signs of breast cancer may include a lump in the breast, a change in breast shape, dimpling of the skin, fluid coming from the nipple, or a red scaly patch of skin. In those with distant spread of the disease, there may be bone pain, swollen lymph nodes, shortness of breath, or yellow skin.[1] Breast cancer is found to be most common cancer occurring in women all over India and accounts for 25% to 31% of all cancers in women in Indian cities. In 2012 it resulted in 1.68 million new cases and 522,000 deaths. It Lumps found in lymph nodes located in the armpits can also indicate breast cancer. Indications of breast cancer other than a lump may include thickening different from the other breast tissue, one breast becoming larger or lower, a nipple changing position or shape or becoming inverted, skin puckering or dimpling. Breast cancer begins in the breast tissue that is made up of glands for milk production, called lobules, and the ducts that connect the lobules to the nipple. Most masses seen on a mammogram and most breast lumps turn out to be benign that is they are non-cancerous, do not spread and do not grow uncontrollably and are not life-threatening.

Mammography (also called mastography) is the process of using low-energy X-rays (usually around 30 kVp) to examine the human breast for diagnosis and screening. The goal of mammography is the early detection of breast cancer, typically through detection of characteristic masses or micro calcifications. Mammography is lightly accurate, but like most of the medical tests, it is not perfect. On average, mammography will detect about 80-90% of the breast cancers in women without symptoms. The tumours may be considered as either micro-calcification or masses.

II. LITERATURE SURVEY

Krawczyk et.al [2] proposed an approach to analyze breast thermo grams based on image features and a hybrid multiple classifier system. The employed image features provide indications of asymmetry between left and right breast regions that are encountered when a tumour is locally recruiting blood vessels on one side, leading to a change in the captured temperature distribution. The presented multiple classifier system is based on a hybridization of three computational intelligence techniques: neural networks or support vector machines as base classifiers, a neural fuser to combine the individual classifiers, and a fuzzy measure for assessing the diversity of the ensemble and removal of individual classifiers from the ensemble. In addition, they have addressed the problem of class imbalance that often occurs in medical data analysis, by training base classifiers on balanced object subspaces. This approach gives a sensitivity of 81.96% when tested on 150 breast thermo grams which is shown to be statistically better than those of all other methods, while resulting in only a slight drop in terms of specificity, and

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 6, June 2017

confirms that their hybrid ensemble algorithm provides an excellent classification method.

Nascimento et.al [3] proposed a system for texture analysis and classification of lesions in mammographic images. Multi resolution analysis features were extracted from the region of interest of a given image. These features were computed based on three different wavelet functions, Daubechies 8, Symlet 8 and bi-orthogonal 3.7. For classification, they used the polynomial classification algorithm to define the mammogram images as normal or abnormal. They also made a comparison with other artificial intelligence algorithms (Decision Tree, SVM, K-NN). A Receiver Operating Characteristics (ROC) curve is used to evaluate the performance of the proposed system. Their system is evaluated using 360 digitized mammograms from DDSM database and the result shows that the algorithm has an area under the ROC curve Az of 0.98 ± 0.03. The performance of the polynomial classifier has proved to be better in comparison to other classification algorithms.

Zheng et.al [4] worked on diagnoses of breast cancer based on the extracted tumor features i.e., quantitative features such as continuous values (e.g. weight), discrete values (e.g. the number of features), interval values (e.g., the duration of an activity) and qualitative features such as nominal (e.g., colour) & ordinal using a hybrid of K-means and support vector machine algorithms. Feature extraction and selection are critical to the quality of classifiers founded through data mining methods. To extract useful information and diagnose the tumour, a hybrid of K-means and support vector machine (K-SVM) algorithms is developed. The K-means algorithm is utilized to recognize the hidden patterns of the benign and malignant tumours separately. The membership of each tumour to these patterns is calculated and treated as a new feature in the training model. Then, a support vector machine (SVM) is used to obtain the new classifier to differentiate the incoming tumours. Based on 10-fold cross validation, the proposed methodology improves the accuracy to 97.38%, when tested on the Wisconsin Diagnostic Breast Cancer (WDBC) data set from the University of California –Irvine machine learning repository.

III. METHODOLOGY

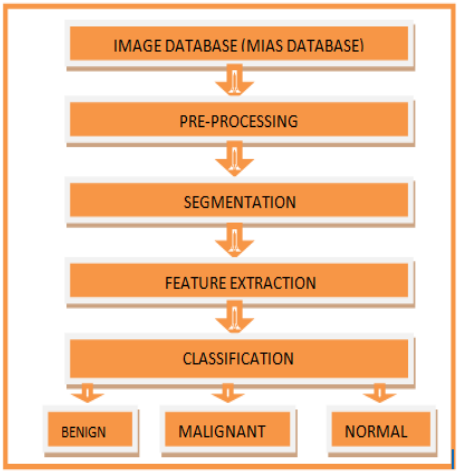


Image Database

The Mammographic Image Analysis Society, MIAS database contains all total 322 mammographic images in MLO which contains 207 normal, 63 benign and 52 malignant cases. In our experiment we have breast tissues which are fatty, fatty-glandular, dense-glandular and the abnormalities like well-defined /circumscribed, speculated masses and ill-defined masses. 60 images are considered for experimentation out of which 20 are benign, 20 malignant and 20 normal.

Image Pre-processing

The aim of the pre-processing technique is to enhance the image quality and make ready for further processing by removing the unrelated and surplus parts in the background of the mammogram images. The Adaptive Median Filter executes spatial processing to safeguard smooth non-impulsive and speckle noise. A chief advantage of adaptive approach to median filtering is repeated applications of this Adaptive Median Filter do not erode away edges and small structures in the image. Adaptive median filter smoothes the data where by keeping the minute and sharp details.

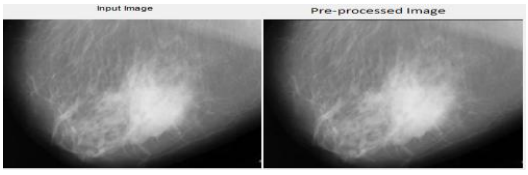


Fig. a) Input Image b) Pre-processed Image

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 6, June 2017

Segmentation

Gaussian Mixture Model is a category of clustering algorithm and makes use of uses iterative technique called Expectation Maximization, very similar to k-means clustering. The only difference is that the clusters are believed to each have an independent Gaussian distribution, each with their own mean and covariance matrix. The Gaussian Mixture Models approach will take cluster covariance into account when forming the clusters. The Gaussian Mixture Model gives a collection of independent Gaussian distributions, and so for each data point, we will have a probability that it belongs to each of these distributions / clusters.

Expectation Maximization

For GMMs, the clusters are found out using a technique called "Expectation Maximization". In the "Expectation" step, the probability that each data point belongs to each cluster is calculated using estimated mean vectors and covariance matrices. In the "Maximization" step, the cluster means and covariance are recalculated based on the probabilities calculated in the expectation step. To begin with the EM algorithm, the data points are randomly selected to use as the initial means, and the covariance matrix for each cluster is set equal to the covariance of the full training set while each cluster is given equal "prior probability". A cluster's "prior probability" is a fraction of the dataset that belongs to each cluster. In the "Expectation" step, probability that each data point belonging to each cluster is calculated. These properties are stored by the covariance matrix.

In ''Maximization'' step, to find the average value of a set of m values, where you have a weight _w _defined for each of the values, you can use the following equation:

$$\bar{y} = \frac{\sum_{i=1}^{m}(w_i y_i)}{\sum_{i=1}^{m} w_i}$$

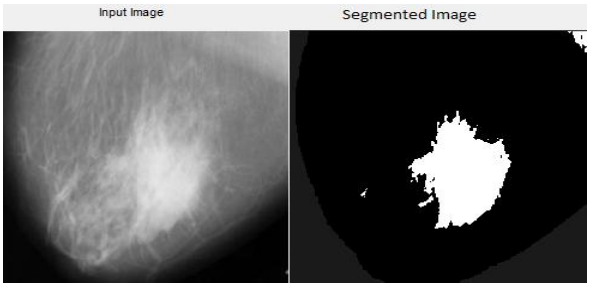


Fig. a) Input Image b) Segmented Image

IV. FEATURE EXTRACTION

Texture Features

Texture is a conception that is easy to recognize but very difficult to define. This difficulty is demonstrated by the number of different texture definitions attempted by vision researchers, some of them are as follows. Texture is visual patterns with properties of homogeneity that do not result from the presence of only a single color such as clouds and water. The texture relates mostly to a specific, spatially repetitive structure of surfaces formed by repeating a particular element or several elements in different relative spatial positions.

John R. Smith defines texture as visual patterns with properties of homogeneity that do not result from the presence of only a single color such as clouds and water .In CBIR, there are many techniques to measure texture similarity, the best-established rely on comparing values of what are known as second-order statistics calculated from query and stored images. Essentially, they calculate the relative brightness of selected pairs of pixels from each image. From these, it is possible to calculate measures of image texture such as the degree of contrast, coarseness, directionality and regularity, or periodicity, directionality and randomness.

Gray level co-occurrence matrix

Texture is one of the important characteristics used in identifying objects or regions of interest in an image. The fourteen textural features proposed by Haralick contain information about image texture characteristics such as homogeneity, gray-tone linear dependencies, contrast, number and nature of boundaries present and the complexity of the image. Co-textual features contain information derived from blocks of pictorial data surrounding the area being analyzed.

Texture feature calculations use the contents of the GLCM to give a measure of the variation in intensity at a pixel of interest. First proposed by Haralick et al. in 1973, they characterize the texture using a variety of quantities derived from second-order image statistics. Co-occurrence texture features are extracted from an image in two steps. First, the pair wise spatial co-occurrences of pixels separated by a particular angle and distance are tabulated using a gray level co-occurrence matrix (GLCM). Second, the GLCM is used to compute a set of scalar quantities that characterize different aspects of the underlying texture.

The GLCM is a tabulation of how often different combinations of gray levels co-occur in an image. The GLCM is an N × N square matrix, where N is the number of different gray levels in an image. An element p (i, j, d, θ) of a GLCM of

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 6, June 2017

an image represents the relative frequency, where i is the gray level of pixel p at location (x, y) and j is the gray level of a pixel located at a distance d from p in the orientation θ. While GLCMs provide a quantitative description of a spatial pattern, they are too unwieldy for practical image analysis. 22 texture features were found out.

V. FEATURES CLASSIFICATION

If we think of classifying any object or people we classify them on the basis of certain features they possess. Suppose, we consider the example of a motorcycle and a human being, humans have legs, hands that are the features that motorcycles do not have. While motorcycles have wheels that are the features that humans do not have. By selecting appropriate set of features, it is easy to classify the objects and people into distinct classes.

To make this kind of feature-based classification work, we need to have some knowledge of what feature make good predictors of class membership for the classes we are trying to distinguish. For example, having wheels or not distinguishes motorcycles from humans, but doesn't distinguish them from cars. These are two different distinguishing tasks. Depending on the classification task we are facing, different features or set of features may be important, and knowing how we arrive at our knowledge of which features are useful to which task is essential.

Texture features are extracted and classified using a classification algorithm which is supervised method that is first trained on a set of sample images (whose classification is known) called the training set. The performance of the algorithm is then tested on a separate testing set.

The classification method involves two steps:

1. **Training (with assessment)** - this is where we discover what features are useful for classification by looking at many pre-classified examples.
2. **Testing (with assessment)** - this is where we look at new examples and assign them to classes based on the features we have learned about during training.

During this process we tried to utilize an equal number of images taken from the MIAS database.60 images were used for experimentation, among them, 20 were benign and 20 were malignant.

Following are the features derived:

I. **Energy**=$\sum_{i=1}^{Ng}\sum_{j=1}^{Ng} P^2 d(i,j)$

II. **Entropy**=$-\sum_{i-1}^{Ng}\sum_{j-1}^{Ng} P_d(i,j)\log(P_d(i,j))$

III. **Contrast**=$\sum_{n=0}^{Ng-1} n^2 \sum_{|i-j|} Pd(i,j)$

IV. **Variance** = $\sum_i \sum_j (i-\mu)^2 p(i,j)$

V. **Homogeneity**=$\sum_{i=1}^{Ng}\sum_{j=1}^{Ng}\frac{P_d(i,j)}{1+|i-j|}$

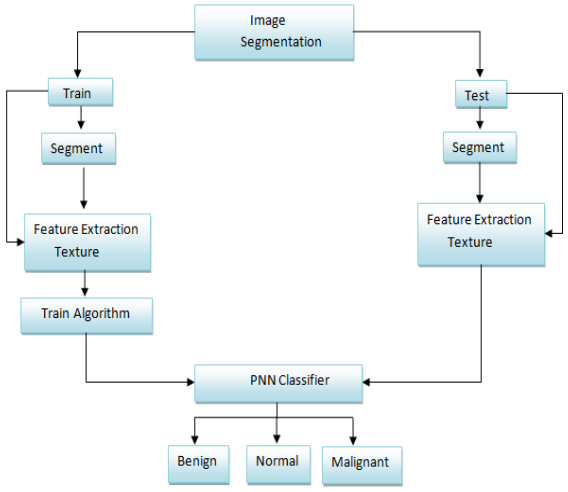
VI. **Correlation**= $\sum_{i=0}^{G-1}\sum_{j=0}^{G-1}\frac{\{i\times j\}\times p(i,j)-\{\mu_x \times \mu_y\}}{\sigma_x \times \sigma_y}$

VII. **Autocorrelation**=$\frac{1}{N}\sum_{i=1}^{N-k}(Z_i - Z)(Z_{(i+k)} - Z)$

VIII. **Cluster Prominence**=
$$\sum_{i=0}^{G-1}\sum_{j=0}^{G-1}(i+j-\mu x-\mu y)^4 P(i,j)$$

IX. **Dissimilarity** = $\sum_{i,j}|i-j|p(i,j)$

X. **Maximum probability**
= Max (p(x, y))



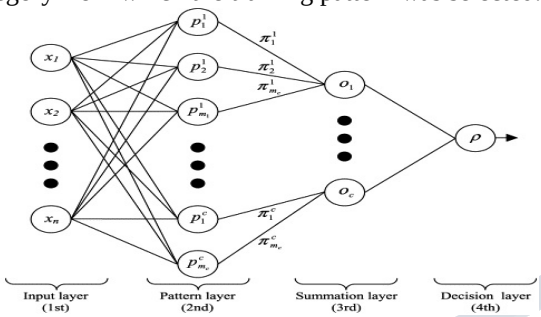
VI. PNN CLASSIFIER

The probabilistic neural net (PNN) is based on the theory of Bayesian classification and the estimation of probability density function (PDF). The PNN architecture consists of four layers: input layer, pattern layer, summation layer, and decision layer. The first layer shows the input pattern with n features. The number of nodes in the pattern layer is equal to the number of training instances. The number of

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 6, June 2017

nodes in the summation layer is equal to the number of classes in the training instances. The input layer is fully connected to the pattern layer. The input layer does not perform any computation and simply distributes the input to the neurons in the pattern layer. The pattern layer is semi- connected to the summation layer. Each group of training instances corresponding to each class is just connected to one node in the summation class. In other words, the summation units simply sum the inputs from the pattern units that correspond to the category from which the training pattern was selected.



TP: true positive, the classification result is positive in presence of malignancy.
TN: true negative, the classification result is negative in being benign.
FP: false positive, the classification result is positive in being benign.
FN: false negative, the classification result is negative in presence of malignancy.
The classification performed is measured by the true positive rate (TPR), the true negative rate (TNR), and the accuracy (ACC). The number of true positive in a classifier is represented as TP, false positive as FP, true negative as TN, and false negative number as FN. In order to estimate the prediction performance of PNN classifier, we calculate the sensitivity, specificity, classification accuracy respectively. According to above definitions the equations related to sensitivity (accuracy of positive class), specificity (accuracy of negative class) and accuracy of recognize both negative and positive classes.
The definitions are as follows:

$$\textbf{TPR}=\frac{TP}{TP+FN} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$\textbf{TNR}=\frac{TN}{TN+FP} \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

$$\textbf{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \ldots\ldots (3)$$

Sensitivity and specificity are the statistical measures of the performance of a binary classification test.

Sensitivity

Sensitivity (also called as the true positive rate or the recall in some fields) measures the proportion of positives which are correctly identified as such (e.g. percentage of sick people who are correctly identified as having the condition), and is complementary to the false negative rate.

Specificity

Specificity (also called the true negative rate) measures the proportion of negatives which are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition), and is complementary to the false positive rate.

Metrics	Formula	Description
Sensitivity	TPR= $\frac{TP}{TP+FN}$ x 100 %	Percentage of abnormalities correctly detected/classified as abnormalities
Specificity	TNR= $\frac{TN}{TN+FP}$ x 100 %	Percentage of normal structures correctly detected/classified as normal
Accuracy	ACC= $\frac{TP+TN}{TP+FP+TN+FN}$ x 100%	Percentage of abnormalities and normal structures correctly detected/classified.

VII.RESULTS AND CONCLUSION

In this paper, we have calculated and presented the results of the classification of breast masses with a data set of 322 images. We have considered 20 benign and 20 malignant masses of circumscribed, speculated and ill-defined masses, asymmetry, architectural distortion and 20 normal mammograms. After the ROI extraction, each mass was

represented with 22 texture features. Before classification, feature selection was performed with 60 ROIs. The PNN classifier was used to classify the 60 images With the PNN classifier, Sensitivity, Specificity and Accuracy achieved in this paper is 100%, 100%, and 100% respectively.

REFERENCES

1) http://www.babymed.com/cancer/introduction-breast-cancer-causes-and-risk-factors

2) https://www.cancer.org/content/dam/cancerorg/research/cancer-facts-and-statistics/annual-cancer-facts and-figures/2017/cancer-facts-and-figures-2017.pdf

3) https://www.researchgate.net/publication/23081760 _Basics_of_Oncology

4) http://www.wcrf.org/int/cancer-facts-figures/data specific-cancers/breast-cancer-statistics

5) Bartosz Krawczyka, Gerald Schaefer "A hybrid classifier committee for analyzing asymmetry features in breast thermograms" Applied Soft Computing 20 (2014) 112–118

6) Marcelo Zanchetta do Nascimento, Alessandro Santana Martins, Leandro Alves Neves, Rodrigo Pereira Ramos, Edna Lucia Flores, Gilberto Arantes Carrijo "Classification of masses in mammographic image using wavelet domain features and polynomial classifier "Expert Systems with Applications 40 (2013) 6213–6221.

7) Danilo Cesar Pereiraa, Rodrigo Pereira Ramosb, Marcelo Zanchetta do Nascimento "Segmentation and detection of breast cancer in mammograms combining wavelet analysis and genetic algorithm" Computer methods and programs in biomedicine 114(2014) 88-101.