Improvement in Detection Accuracy of Digital Mammogram Using Point Transform and Data Mining Technique

S.P. Meharunnisa, M. Ravishankar and K. Suresh
Department of Electronics and Instrumentation Engineering, Visvesvaraya Technological University, India

Abstract
Cancer is one of the dangerous diseases faced by humans. Every one out of 100 women is facing breast cancer. So, to overcome this huge ratio many researches are being carried out. Prevention is better than cure; this paper presents one such attempt of detecting breast cancer in the early stages. In proposed method exponential point transform is carried out for image enhancement and in preprocessing stage pectoral mass is removed from the mammogram image. As the next step we apply K-means algorithm and morphological processing to identify the infected region and removal of unwanted region. Finally, Decision Tree Data mining technique is used for classifying features to detect presence of tumor. Hence by this approach we get more accurate results. The experimental results gave an accuracy of 97.03%.

Keywords:
Breast Cancer, Point Transform, K-means, Decision Tree Classifier

1. INTRODUCTION

Breast cancer is the most and one of the major causes for death among women as per world health organization. The ratio of breast cancer in women is increases day by day. Digital Mammography is considered as the most efficient method for detecting the breast cancer at the remediable stage [3]. Noisy images make it difficult for the radiologists to detect the signs of breast cancer that is the massive tissue present in the breast. About 10 to 30% of breast lesions are missed due to routine screening due to over sightedness and reading these images is a time consuming and hard task for radiologists. Poorly defined masses, developing densities and masses not identified with mammograms may lead to false results. Hence by processing and enhancing the readability of these mammograms, CAD can play a significant role in the identification of early signs of breast cancer like microcalcifications, masses and bilateral asymmetry.

The goal of this work is to diagnosis and detect early signs of breast cancer to enhance the quality of life and to assist radiologist for misclassification of cancer.

Early researches have proposed various techniques for detection of massive tissue present in the breast mammogram image [4][5], so that preventive measures could be taken to avoid the affected women to suffer from breast cancer, so that the survival rate and recovery chances can be improved and the ratio of the infected women can be brought down. There are many algorithms proposed by researchers like Dual stage adaptive threshold algorithm [6], ID3 algorithm [7], Formation of homogeneous blocks for mammogram segmentation [12], Gabor filtering [13] etc.

Organization of the Paper: In the proposed method images are taken from MiniMias data base as the input image and preprocessing is done to remove artifacts present in the image. Later exponential transform is applied for image enhancements which makes the tumor region appears bright with reference to back ground. K-means algorithm is employed to segment the tumor part; later morphological processing steps are applied to remove unwanted regions. Statistical features like Mean, Kurtosis, and Variance are calculated for segmented image and the same is given to Decision Tree classifier to indicate the presence of tumor. The proposed method uses advance preprocessing, segmentation and classifier techniques that helps the radiologist to detect the infected region.

2. RELATED WORK

Jaleel et al. [14] proposes an algorithm for early detection of breast cancer. This work incorporates Manual segmentation and textural analysis for the mammogram mass classification. A total of 148 mammogram images were taken from mini MIAS database and classified into benign and malignant. The classifiers used are K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), Radial Basis Function Neural Network (RBFNN). It is found that RBFNN with DWT features outperform SVM and K-NN with 94.6% accuracy.

Washington et al. [2] proposes fuzzy morphological extreme learning machines with hidden layer kernel based on nonlinear morphological operators of erosion and dilation. The proposed methods were evaluated using 2,796 images from IRMA database, considering fat, fibroid, dense and extremely dense tissues. Zernike Moments and Haralick texture features are used as image descriptors. The proposed model classifies masses as benign, malignant or normal. Results shows comparison between Extreme Learning Machines using Sigmoid and Fuzzy Morphological Kernels, evaluated using classification rate and Kappa index. Using fuzzy morphological kernels, classification rate and Kappa value increases for most of cases analyzed.

Elmanaa et al. [15] used CAD system to distinguish between masses and normal breast tissue was proposed. DDSM database for mammogram images which were first preprocessed using image enhancement algorithm, then 100 regions of interest (ROIs) containing masses and normal breast tissue are extracted. They performed feature selection using Sequential Forward Selection and Sequential Floating Forward Selection. Finally used K-Nearest Neighbor (KNN) classifier, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Support Vector Machine (SVM) classifier for classification with leave-one-out method for testing. The obtained results show acceptable sensitivity and specificity for the system.

Tai et al. [17] presents an automatic CAD system that uses local and discrete texture features for mammographic mass detection. This study also proposes two complex feature extraction methods based on co-occurrence matrix and optical density transformation to describe local texture characteristics.
and the discrete photometric distribution of each ROI. Linear discriminant analysis is used to classify abnormal regions by selecting and rating the individual performance of each feature. Results shows that the proposed system achieves satisfactory detection performance.

Rangaraj et al. [23] reviews the recent advances in CAD systems for detection of breast cancer by identifying different abnormal features and early signs like micro-calculcitations, masses, bilateral asymmetry and architectural distortions. Also, gives study of currently employed image enhancement techniques for contrast enhancement of mammographic images. Explains about different categories of technologies involved in development of algorithms for identification of early signs of breast cancer.

3. METHODOLOGY

3.1 PRE-PROCESSING AND IMAGE ENHANCEMENT

Pre-processing of image is required to extract the useful information from the given image and discard the unwanted data. For this we need a set of data to extract the special information from the given image that is we require a digital data format. In our proposed method we use MiniMias data base as means of set of data which has around 300 mammogram images and also around 150 mammogram images are collected from doctor for experiment purpose.

We consider an input image from MiniMias data base and resize that image into 512×512. The next step is to pre-process this resized image to remove the artifacts like date and time printed on the image during X-ray. The Fig.1 Represents the steps involved in preprocessing stage.

![Image](image.png)

**Fig.1. Steps involved in preprocessing stage**

The artifacts have been removed in the preprocessing stage for further processing. Fig.2 shows the mammogram input image and preprocessed image.

The pectoral mass is removed from the scanned image by global thresholding and BLOB analysis. Pectoral muscle [1] may increase the computational complexity of the detection process and also causes the reduction in detection accuracy. Hence to remove all these unnecessary parts from the breast region in the mammogram is a vital preprocessing task in CAD system of the breast cancer.

In global thresholding, a single threshold for all the image pixels is used. When the pixel values of the components and that of background are fairly consistent in their respective values over the entire image, global thresholding could be used.

Blob Analysis is a fundamental technique of machine vision based on analysis of consistent image regions. As such it is a tool of choice for applications in which the objects being inspected are clearly discernible from the background. Diverse set of Blob Analysis methods allows to create tailored solutions for a wide range of visual inspection problems.

Main advantages of this technique include high flexibility and excellent performance. Its limitations are: clear background-foreground relation requirement and pixel-precision.

The basic scenario of the Blob Analysis solution consists of the following steps:

- Extraction: in the initial step global thresholding technique is applied to obtain a region corresponding to the objects.
- Refinement: the extracted region in the refinement step is enhanced using point transformation techniques.
- Analysis: in the final step each pixel of the region is subject to measurements and the final results are computed. If the region represents multiple objects, it is split into individual blobs each of which is inspected separately.

3.2 IMAGE ENHANCEMENT

After pre-processing and Pectoral mass removal is done, the next step carried out is the image enhancement by exponent point transform. In proposed point transformation, we apply transformations on each and every pixel of the image to get the resulted image. The Fig.3 represents the steps involved in pectoral mass removal and image enhancement.

The Fig.4 illustrates about point transformation and is given by Eq.(1)

\[ EPT: S = C \times r^\gamma \]  

where, **EPT** is the Exponential points transform, **C** is the Multiplication factor, **r** is the input image, **s** is the output image and **\gamma** is the decaying factor.
For a given cluster assignment $C$ of the data points, compute the cluster means $m_k$ using Eq. (3)

$$
m_k = \frac{\sum_{i=1}^{N_k} x_i}{N_k}, \quad k = 1, \ldots
$$

For a current set of cluster means, assign each observation as in Eq.(4)

$$
C(i) = \arg \min_{k \in \{1, \ldots, K\}} \| x_i - m_k \|^2, \quad i = 1, \ldots
$$

Iterate above two steps until convergence. K-means is the fastest clustering algorithm when compared with other segmentation algorithms. Hence our proposed methodology is implemented using K-means algorithm. Fig.5 shows the result of K means segmentation.

3.4 POST PROCESSING

Post processing is done to remove unwanted region and retain the tumor region. The method used for post processing is morphological processing. Morphological processing operates on a binary image after non-tumor region is discarded.

The structuring element is a small binary array. The basic structuring operation is of two types that is erosion and dilation. Erosion operation is given by Eq.(5)

$$
A - B = \{ z \mid B_z \subset A \}
$$

where, $A$ is the original set and $B$ is the structuring element. Thus erosion removes the thin layer, isolated dots and leaves gross details in an image.

Dilation is the opposite operation of erosion. Dilation operation is given by Eq.(6)

$$
A + B = \{ z \mid B_z \subset A \}
$$

One more operator we use in morphological processing is the WED operator that bridges gaps or fills holes in a given image but it does not changes overall size of objects. The two types of operation in WED operators are OPENING and CLOSING. OPENING will first erode the image and then dilate it back again which results in breaking narrow bridges and eliminate thin structures. CLOSING will first dilate the image and then erode it back. It fuses narrow bridge and eliminates the small holes present in the image. Thus, morphological processing will eliminate the entire noise region and retain back the tumor region. Fig.6 illustrates post processing methods output,
Code flow for the proposed method is as follows,

- The input image, taken from minimias database is read.
- The input image is then resized to 512×512.
- The artifacts present in the resized image are removed by morphological process.
- The next step is to set a threshold for the pectoral mass in the image, above the threshold value the region is white and below the threshold value the region is black. The maximum intensive white region is find out that is the binary large object is found that contains the human part and the remaining is considered as noise.
- Image enhancement is done for this BLOB region.
- K-means Segmentation algorithm is applied to the enhanced image for clustering the image.
- Morphological processing is applied to remove the noise region and retain the tumor region.
- After morphological process the features of the image is calculated.
- Once the features of the image is calculated, we apply the data mining decision tree technique to classify the tumor.

The Table.1 shows the features extracted from the processed image

Table.1(a). Features Extracted from Processed Image (Contrast, Variance and Standard Deviation)

| File name | Contrast  | Variance | Standard Deviation |
|-----------|-----------|----------|--------------------|
| mdb001    | 0.133348  | 0.014027 | 0.118437           |
| mdb013    | 0.092909  | 0.009705 | 0.098516           |
| mdb050    | 0.011283  | 0.001114 | 0.03337            |
| mdb023    | 0.091396  | 0.010591 | 0.102913           |
| mdb211    | 0.098138  | 0.010328 | 0.101626           |
| mdb206    | 0.055062  | 0.003697 | 0.060801           |
| mdb209    | 0.104506  | 0.016172 | 0.12717            |
| mdb236    | 0.048694  | 0.011746 | 0.108377           |
| mdb248    | 0.041577  | 0.003069 | 0.055397           |
| mdb290    | 0.076413  | 0.018181 | 0.134838           |
| mdb315    | 0.076787  | 0.013136 | 0.114613           |

Table.1(b). Features Extracted from Processed Image (Kurtosis, Mean and Smoothness)

| File name | Kurtosis   | Mean        | Smoothness |
|-----------|------------|-------------|------------|
| mdb001    | 66.32395   | 0.014439    | 0.013833   |
| mdb013    | 77.47777   | 0.011126    | 0.009612   |
| mdb050    | 596.8971   | 0.001365    | 0.001112   |
| mdb023    | 87.72898   | 0.010929    | 0.01048    |
| mdb211    | 90.79569   | 0.010612    | 0.010222   |
| mdb206    | 244.8553   | 0.003879    | 0.003683   |
| mdb209    | 57.29488   | 0.016664    | 0.015915   |
| mdb236    | 67.55838   | 0.013094    | 0.011609   |
| mdb248    | 303.3516   | 0.003176    | 0.003059   |
| mdb290    | 50.99059   | 0.018707    | 0.017857   |
| mdb315    | 71.95449   | 0.013422    | 0.012966   |

Fig.6. Post processing after segmentation process

The Table.1 shows the features extracted from the processed image for any one of the patient from the input database.

3.5 CLASSIFICATION

Classification is the process of classifying set of data into different classes or categories. Accurate classifications can be done using either by learning techniques or feature sets.

Several classification methods are developed in recent years. However, how to choose the best classifier with compatible features is a challenge. Non-parametric classifiers such as neural network, decision tree classifier and knowledge-based classification have increasingly become important approaches for multisource data classification. The classification process is divided into learning phase and testing phase. In the learning phase, known data are given and the feature parameters are calculated by the processing which precedes classification. The Predefined data on a candidate region which has already been identified by the doctor as tumor or normal as given, and the classifier is trained. In the testing phase unknown data are given and the classification is performed using the classifier after learning. We compared the features extracted from the proposed point transform enhanced image with various classifiers for the CAD system. Decision tree classifier, support vector machine classifier (SVM), linear quadratic discriminant analysis classifier
(LQDA) and Navy Bayesian (NB) classifiers were used and performance analysis has been done.

3.5.1 Support Vector Machines (SVM):

Support vector machines represent an extension to nonlinear models developed by Vapnik and Lerner. The SVM algorithm is based on statistical learning theory which describes the properties of learning machines that allow them to give reliable predictions.

3.5.2 Linear and Quadratic Discriminant Classifier (LQDA):

Quadratic discriminant analysis is closely related to linear discriminant analysis, where it is assumed that the measurements from each class are normally distributed, unlike LDA however, in QDA there is no assumption that the covariance of each of the classes is identical. LDA is used to find the linear combination of features which best separate two or more classes of objects or events.

3.5.3 Naive Bayesian Classifier (NB):

NB classifiers are simple probabilistic classifiers based on applying Bayes’ theorem with Naive (strong) independence assumptions between the features. They are highly scalable, requiring a number of parameters linear in a learning problem. It is machine learning algorithm for classification problems.

3.5.4 Decision Tree Classifier:

Decision tree algorithm is a regressive algorithm. It is the most fundamental technique that gives accurate results. Statistical feature like Contrast, Variance, and Standard deviation, Kurtosis, Mean and Smoothness are given as inputs to this classifier. A decision tree algorithm is a tree like structure in which internal node represents test on an attribute, each branch represents outcome of test results. We are going to split the attributes, for each value of attribute we create a new child node. Later we sort the child node and apply iteration on the child node. If the iteration is true then the decision is applied to say whether the taken image is affected by tumor or not. Tree diagram of features set is given in Fig.8.

In Principle, there are exponentially many decision trees that can be constructed from a given set of features or attributes. While some of the trees are more accurate than others, finding the optimal tree is computationally infeasible because of the exponential size of the search space. Nevertheless, efficient algorithms have been developed to induce a reasonably accurate, albeit suboptimal decision tree in a reasonable amount of time, one such algorithm is Hunt’s algorithm which is the basis of many existing decision tree induction algorithms, including ID3, C4.5 and CART.

In Hun’s algorithm, a decision tree is grown in a recursive fashion by partitioning the training records into successively purer subsets.

Let \( T_s \) be the set of training records that are associated with node \( s \) and \( y = \{y_1, y_2, \ldots, y_k\} \) be the class labels.

**Step 1:** if all the records in \( T_s \) belong to the same class \( y_i \), then \( s \) is a leaf node labeled as \( y_i \).

**Step 2:** if \( T_s \) contains records that belong to more than one class, an attribute test condition is selected to partition the records into smaller subsets. A child node is created for each outcome of the test condition and the records in \( T_s \) are distributed to the children based on the outcomes. The algorithm is then recursively applied to each child node.

4. PERFORMANCE ANALYSIS

Performance analysis of Point transform with decision tree data mining technique has been compared with SVM, LQDA and Navy Bayesian classifiers. Point transform with decision tree classifier gives more accurate results than other approaches. The comparison table is as shown below Table.2.

| Proposed Point Transform Algorithm with Classifiers | Specificity (%) | Sensitivity (%) | Accuracy (%) |
|--------------------------------------------------|-----------------|----------------|-------------|
| Decision tree classifier                         | 96.77           | 98.11          | 97.03       |
| SVM Classifier                                   | 96.31           | 79.24          | 92.96       |
| LQDA Classifier                                  | 95.39           | 81.11          | 92.59       |
| NB Classifier                                    | 87.57           | 96.22          | 75.54       |

From Table.2, we can inference that proposed point transform with decision tree classifiers gives 97.03% accuracy than other classifiers methods.

5. CONCLUSION

The early detection of Breast Cancer based on image processing techniques is reliable and can prove to be an important investigative tool in the clinical evaluation of detection of cancer in early stages. Mammography is the best available inspection facility to detect the onslaught of breast cancer at the early stage. It can disclose information about abnormality, such as masses, micro-calculifications, bilateral asymmetry and architectural distortion. For the hundreds of mammographic images scanned by a radiologist, only a few are cancerous. While detecting abnormalities, some of them may be missed due to human error, as the detection of suspicious and abnormal images is a recurrent mission that causes fatigue and eyestrain. Using our proposed technique to enhance images, segment and extract the suspicious area, more accurate features can be analyzed. Besides, they can be characterized into appropriate category that our approach can contribute leading to significantly reduce in the false diagnosis of Breast Cancer.
REFERENCES

[1] Suhas Sapate and Sanjay Talbar, “An Overview of Pectoral Muscle Extraction Algorithms applies to Digital Mammograms”, Medical Imaging in Clinical Applications, pp. 19-54, 2016.

[2] Mussarat Yasmin, Muhammad Sharif and Sajjad Mohsin, “Survey Paper on Diagnosis of Breast Cancer using Image Processing Techniques”, Recent Journal of Recent Sciences, Vol. 2, No. 10, pp. 88-98, 2013.

[3] K. Ganesan et al., “Computer-Aided Breast Cancer Detection using Mammograms: A Review”, IEEE Reviews in Biomedical Engineering, Vol. 6, pp. 77-98, 2012.

[4] K. Hu, X. Gao and F. Li, “Detection of Suspicious Lesions by Adaptive Thresholding based on Multiresolution Analysis in Mammograms”, IEEE Transactions on Instrumentation and Measurement, Vol. 60, No. 2, pp. 462-472, 2011.

[5] B.N. Prathibha and V. Sadasivam, “Mammogram Analysis using SVM Classifier in combined Transforms Domain”, ICTACT Journal on Image and Video Processing, Vol. 01, No. 3, pp. 172-177, 2011.

[6] J Anitha, J Dinesh Peter and Immanuel Alex Pandian, “A Dual Stage Adaptive Thresholding (DuSAT) for Automatic Mass Detection in Mammograms”, Computer Methods and Medicine in Mammograms, Vol. 138, pp. 93-104, 2017.

[7] Aswini Kumar Mohanty, Manas Ranjan Senapathy, Swapnasibita Baberta and Saroj Kumar Lenka, “Texture based Features for Classification of Mammograms using Decision Tree”, Neural Computing and Applications, Vol. 23, No. 3-4, pp. 1011-1017, 2013.

[8] R.M. Haralick, K. Shanmugam and I.H. Dinstein, “Texture Features for Image Classification”, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 3, No. 6, pp. 610-621, 1973.

[9] E. Walid and H. Hakim, “A New Cost Sensitive Decision Tree Method Application for Mammograms Classification”, International Journal of Computer Science and Network Security, Vol. 6, No. 11, pp. 130-138, 2006.

[10] K. Polat and S. Gunes, “A Novel Hybrid Intelligent Method based on C4.5 Decision Tree Classifier and One-Against-all Approach for Multi-Class Classification Problems”, Expert Systems with Applications, Vol. 36, No. 2, pp. 1587-1592, 2009.

[11] M. Hall-Bey, “GLCM Texture: A Tutorial v.2.7.1”, Available at: www.fp.ucalgary.ca/mhallbey/tutorial.htm, Accessed on 2004.

[12] Samir Kumar Bandypodhiyay, “Formation of Homogeneous blocks for Segmentation of Mammograms”, International Journal of Engineering Science and Technology, Vol. 2, No. 12, pp. 7444-7448, 2010.

[13] Yufeng Zheng, “Breast Cancer Detection with Gabor Features from Digital Mammograms”, Algorithms, Vol. 3, No. 1, pp. 44-62, 2010.

[14] J. Abdul Jaleel, S. Salim and S. Archana, “Mammogram Mass Classification Based on Discrete Wavelet Transform Textural Features”, Proceedings of International Conference on Advances in Computing, Communications and Informatics, pp. 24-27, 2014.

[15] Washington W. Azevedo et al., “Fuzzy Morphological Extreme Learning Machines to Detect and Classify Masses in Mammograms”, Proceedings of IEEE International conference on Fuzzy Systems, pp. 1-6, 2015.

[16] M.E. Elamna and Y.M. Kadah, “Implementation of Practical Computer Aided Diagnosis System for Classification of Masses in Digital Mammograms”, Proceedings of International Conference on Computing, Control, Networking, Electronics & Embedded Systems Engineering, pp. 7-13, 2015.

[17] Shen-Chuan Tai, Zhi-Siou Chen and Wei-Ting Tsai, “An Automatic Mass Detection System in Mammograms Based on Complex Texture Features”, IEEE Journal of Biomedical and Health Informatics, Vol. 18, No. 2, pp. 618-627, 2014.

[18] Uma Ojha and Savitha Goel, “A Study on Prediction of Breast Cancer Recurrence using Data Mining Techniques”, Proceedings of 7th International Conference on Cloud Computing, Data Science Engineering, pp. 107-113, 2017.

[19] B. Shradhananda, B. Majhi and R. Dash, “Mammogram Classification using Two Dimensional Discrete Wavelet Transform and Grey Level Co-occurrence Matrix for Detection of Breast Cancer”, Neurocomputing, Vol. 154, pp. 1-14, 2014.

[20] Zhe Jiang, Shashi Shelkar, Xin Zhou, Joseph Knight and Jennifer Corcoran, “Focal-Test Based Spatial Decision Tree Learning”, IEEE Transactions on Knowledge and Data Engineering, Vol. 27, No. 6, pp. 1547-1559, 2015.

[21] Ibrahim Mohamed Jaber Alamin, W. Jeberson and H K Bajaj “Improved Framework for Breast Cancer Detection using Hybrid Feature Extraction and FFNN”, International Journal of Advanced Research in Artificial Intelligence, Vol. 5, No. 8, pp. 1-6, 2016.

[22] Rodrigo Coelho Barros, Marcio Porto Basgaluppe, Andre C.P.L.F. De Carvalho and Alex A. Freitas, “A Survey of Evolutionary Algorithms for Decision Tree Induction”, IEEE Transactions on Systems, Man, Cybernetics-Part C: Applications and Reviews, Vol. 42, No. 3, pp. 112-132, 2012.

[23] Jinshan Tang, Rangaraj M. Rangayyan, Issam El Naqa and Yongyi Yang, “Computer-Aided Detection and Diagnosis of Breast Cancer with Mammography: Recent Advances”, IEEE Transactions on Information Technology in Biomedicine, Vol. 13, No. 2, pp. 236-251, 2009.

[24] S.P. Mehranunisa, K Suresh and M. Ravishankar, “Detection of Breast Masses in Digital Mammograms using SVM”, International Science Press, Vol. 8, No. 3, pp. 899-906, 2015.

[25] S.P. Mehranunisa and K Suresh, “Identification of Microcalcifications for Early Signs of Breast Cancer”, STM Journals, Vol. 4, No. 3, pp. 7-12, 2014.

[26] S.P. Mehranunisa, B. Amith and K Suresh, “Early Detection of Breast Cancer using Computer Aided Detection and Diagnosis Recent Advances”, International Journal of Engineering Research and Technology, Vol. 5, No. 3, pp. 43-51, 2015.

[27] S.P. Mehranunisa, B. Amith and K Suresh, “Detection and Classification of masses in Breast Cancer using Computer Aided Detection and Diagnosis”, International Journal for Research and Development in Technology, Vol. 3, No. 6, pp. 1-7, 2015.

[28] Xiang Zhong, Jingshan Li, Susan M. E Reddy, “System-Theoretic Approach to Modeling and Analysis of Mammography Testing Process”, IEEE Transactions On Systems, Man, and Cybernetics: Systems, Vol. 46, No. 1, pp. 126-138, 2016.