ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

Early Diagnosis of Breast Cancer Using SVM

Manisha Gahirwal, Swita Fulwani, Kanchan Manik, Neha Navale, Aditi Parab.

Assistant Professor, Department of Computer, V.E.S.I.T, Chembur, Mumbai, India

B.E. Student, Department of Computer Engg, V.E.S.I.T, Chembur, Mumbai, India

B.E. Student, Department of Computer Engg, V.E.S.I.T, Chembur, Mumbai, India

B.E. Student, Department of Computer Engg, V.E.S.I.T, Chembur, Mumbai, India

B.E. Student, Department of Computer Engg, V.E.S.I.T, Chembur, Mumbai, India

ABSTRACT: Over years breast cancer has proven to be a major health problem in women.It is essential to treat breast cancer at its early stage so as to detect it successfully and reduce mortality. For this mammography is the essential diagnostic test for breast cancer tumor screenings wherein high-quality images with low x-ray dose are used to image the breast(s).As this disease presents immense danger it becomes necessary to build a Computer Aided Diagnosis(CAD) system to improve the accuracy of the diagnosis. This paper has proposed such a system. The developed method has the following four steps: pre-processing mammographic input for image enhancement, segmentation to detect Region of interest for breast abnormalities, then features are extracted and finally the mammographic image is classified.

KEYWORDS*:* Breast Cancer, Mammography, image pre-processing, segmentation, feature extraction, image classification.

I. INTRODUCTION

Breast cancer is the second frequently diagnosed cancer among women, especially in developed countries. In western countries about 53%-92% of the population has this disease [1]. Though breast cancer leads to death, early detection of breast cancer can increase the survival rate. The current diagnostic method for early detection of breast cancer is mammography. Mammography still remains the vital screening tool as it represents the most effective, low-cost and highly sensitive technique allowing the detection of a breast cancer at its initial stage. The mammograms are checked by the radiologist with the aim of detecting the abnormalities, but the complex structures and the signs of early disease are very small and subtle. It represents an abnormal cancer mass or microcalcifications. Microcalcifications (MC) are very tiny bits of calcium salt, and they may be present in clusters or in patterns and are associated with extra cell activity in breast tissue [1]. Usually the extra cell growth is not cancerous, but sometimes tight clusters of microcalcification can represent early breast cancer. MC in the breast shows up as white speckles on breast Mammograms. The calcifications are small; usually varying from 100 micrometer to 300 micrometer, but in reality may be as large as 2mm. It is very difficult to detect the microcalcifications in the mammogram, when more than 10 calcifications are clustered together, it becomes much easier to diagnose the disease. But the vitality still depends upon the stage at which the cancer is detected. Hence, any MC formation must be detected at the benign(initial) stage. A Computer Aided Diagnosis (CAD) system is used to inspect MC clusters. Various algorithms have been developed for automatic detection of breast cancer in mammographic images of breast(s). And finally, the features extracted from mammographic images are used to detect cancer.

II. RELATED WORKS

There are many methods of MCs and mass detection in MGs, mainly including traditional image processing methods, such as filtering, threshold algorithms, neural network, SVM, etc. In the breast area was first segmented via morphological filtering and thresholding algorithm, then, the difference image between the enhanced image and the

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

noise-suppressed one in the breast area was employed for MCs segmentation of this difference image by classification based on the neural network classifier. According to Breast Imaging Reporting and Data System of the American Radiology College , the characterizations of MCs include different parameters depending on size,shape and distribution properties. Currently, the descriptions of the extraction area are mainly gray features, shape features and texture features. Gray features reflect the density of breast tissue and contrast between the lesion and surrounding tissue. The commonly used features are the mean value, the variance, the contrast, etc. Kinoshita S.K. et al calculated the histogram statistical characters of the breast area, including mean, variance, skewness, kurtosis and entropy. They used particle size character to describe the structures distribution of the different size, and the Leyden domain characterization to describe the distribution of linear structure. The feature space describing the MGs is often large and complex. Therefore, feature selection is an essential work. The common methods in feature selection include PCA, linear decision analysis, logistic regression, backward selection, one dimensional analysis and genetic algorithms. Important part of the CAD system is a classification step. In literature, there are also described many classification methods, including linear discriminant analysis, artificial neural network, Bayesian methods,rule-base detection methods, decision tree, and mixed classification methods, etc. Rangayyan et al. proposed the use of shape factors and edge acutance for the classification of manually segmented masses as benign or malignant, and speculated or circumscribed, finally obtaining an overall classification accuracy of 95% on a database of 54 MGs.They obtained a sensitivity of 91% with 3.2 false positives per image in the training phase. Jian, W. et al. extracted manually 25 abnormal ROIs according to the center and diameter of the lesions and 25 normal ROIs selected randomly. Then, MCs were segmented by combining space and frequency domain techniques . Afterwards, three texture features based on wavelet (Haar, DB4, DT-CWT) were extracted. Systems for CAD have been designed with the aim of assisting radiologists in the analysis of MGs with purpose to increase the accuracy of diagnosis, as well as to improve the consistency of interpretation of images via the use of the computer's results as a reference. The results of computer-based image processing and CAD could also be useful in addressing other limitations in visual interpretation of MGs due to poor quality and low contrast of the images, superposition of breast structures due to the projection nature of MGs, visual fatigue in the screening context, and environmental distraction. It has been shown that double reading (interpretation of each MG exam by two radiologists) can increase the accuracy of diagnosis: the suggested use of CAD systems also includes the role as a second reader. Nevertheless, it is difficult to achieve high accuracy in deriving measures of breast tissue density due to intrinsic difficulties with MG images; furthermore, the estimates provided by radiologists based upon visual analysis are subjective. An approach to address the problems mentioned above is to realize unsupervised and automatic classification of images through the characterization of similarity based upon breast tissue density. Such a classification permits quantitative evaluation of the similarity of images independent of subjective factors. Usually such systems may involve three principal steps: (i) segmentation; (ii) parameter extraction and selection of the segmented lesions; (iii) lesions classification. In this study, we try to develop full automatic breast cancer lesions detection system. Figure 1 presents block diagram of the designed CAD system.

Approach for Breast Cancer Detection: The proposed modelincludes the layout of the functioning of the CAD system. It includes capturing a mammographic image. Pre-processing is done to enhance the image. Later it segments the suspicious mass region(s) that might contain malignant mass [1]. Feature extraction grabs the features of the image which include the texture features based on the Gray level scheme [1]. Finally, the classification of the mass region is done so as to distinguish it as normal or abnormal mass. After distinguishing results are brought upon and discussion is done keeping in account the case study. The proposed model is shown in Fig. 1.1.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

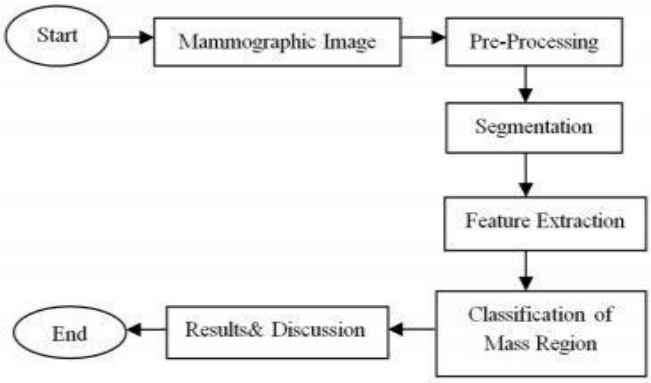


Fig 1.1 Model of Mammography for Cancer Detection

III. CAD SYSTEM

As breast cancer is the first cause of death among women, it has been proved that early diagnosis is the most efficient way to defeat the disease. Many studies show that double reading by two radiologists improves sensitivity up to 15% [2]. Unfortunately, due to the high cost, it is not always possible to have two radiologists in a radiological centre. Therefore, in recent years several computerized systems have been developed to aid radiologists working in mammography to reach a correct diagnosis. The main goal of this Computer-Aided Detection (CAD) systems is to focus the radiologist's attention on suspicious areas [2]. Preliminary studies have demonstrated that their use can lead to a more significant perception of abnormalities, with an increment in sensitivity[3]. Computer-aided detection (CAD) systems apply image processing and pattern recognition technique to detect and classify abnormalities in mammograms, which can provide an open-minded view to the radiologist. The abnormalities in mammograms include microcalcifications (MCs), architectural distortion, masses and asymmetry [2] [3].

A) Pre-processing

Mammographic images with and masses microcalcifications are usually small and have low contrast thus making the abnormalities hard to be detected. Pre-processing block involves enhancing image, removal of noises, blood vessels and glandular tissues which become a cause of many False Positives during detection stage [1]. Fig. 1.2(a) shows a mammogram containing a mass in mediolateral-oblique (MLO) view, and the pre-processing method is described below. The pixel values that lie in this range are saturated to the upper or lower limit value, respectively as shown in Fig 1.2(b). Next, Filtering operation is applied as it transforms pixel intensity values to reveal certain image characteristics. Filters employed improved the: uneven contrast, smoothened the image, removed noise, gray level enhancement, highlights sharp gradients and improves the edges of the mammographic image as shown in Fig. 1.2(c). Because the pectoral muscles represent a brighter region, it can affect the detection process [1]. Thus, the whole foreground is transformed by gamma correction with a decoding gamma to preserve the brighter luminance and suppress the darker luminance. In other words, the gamma expansion enhances the pectoral muscle, the enhanced image is as shown in Fig 1.2(d)

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

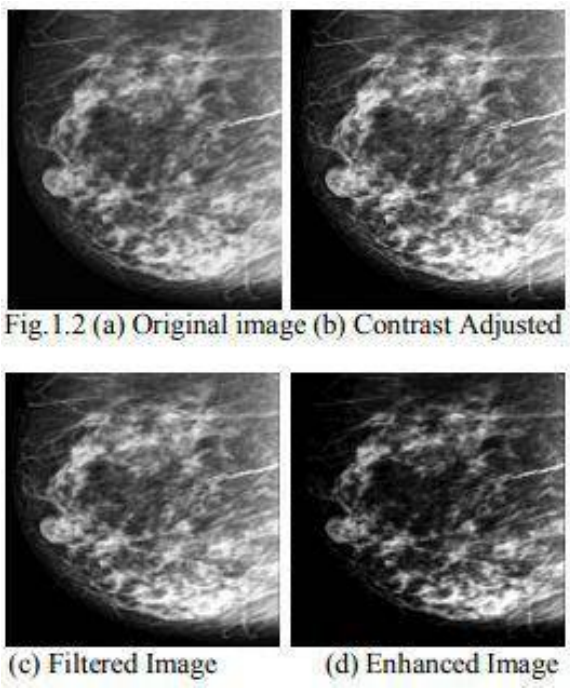


Fig 1.2 Image Pre-processing

B) Segmentation

This stage uses the output of the first stage to segment the ROIs. Segmentation is the division of the enhanced image Fig 1.2(d) into various non-overlapping regions. It corresponds to the extraction of objects from the background. The segmentation is done to extract locations of suspicious areas to aid and classify the abnormalities as malignant or benign [1]. Segmentation algorithms are based on intensity value, which are discontinuous, based on abrupt changes in the image, as edges and similarity. Thus, depending on the nature of images and the region of interest, the segmentation methods can attempt to detect the ROIs [1]. After the pre-processing of the image we get image in gray format and now we have to segment this enhanced image. According to the images and directions from the radiologist, tumor regions were selected and the regions had varying intensity values. Thus, various morphological operations are applied to extract the required regions as discussed below: 1. Connected Components: Removal of the connected components that have fewer than 50 pixels and produce another BW image. This operation is known as an area opening. 2. Measurement of Properties: After the removal of connected components we are interested in knowing various properties of the regions using „region props" like: Area, Euler Number, Orientation, Bounding Box, Extent, Perimeter, Centroid, Convex Area, Filled Area, Pixel List, Eccentricity etc. Then used the suitable parameters like Area, Centroid and Bounding Box [1]. Segmentation: Rectangular boxes of defined values as specified from above properties are segmented and saved. Thus, segmentation of image(s) is done now further process of tumour severity needs to be unlocked. The segmented regions from the enhanced image Fig. 1.2(d) are shown in Fig. 1.3(a, b, c, d, e, f, g, h, i, j).

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

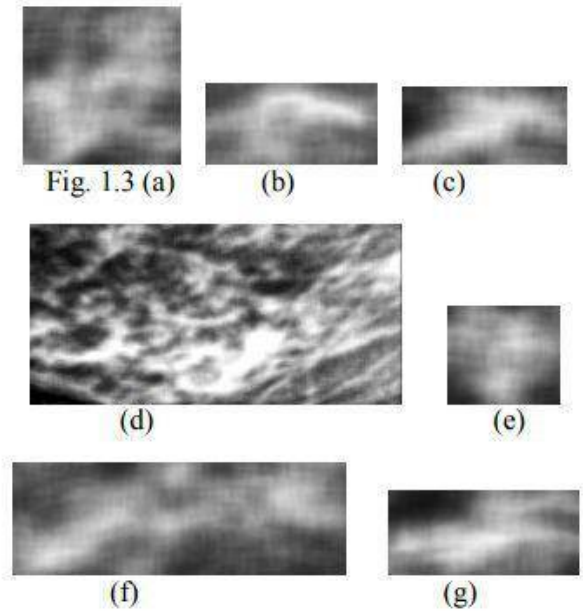


Fig 1.3 ROI extraction by segmentation

There are various types of image segmentation techniques such as Region-based, Edge detection, Feature-based clustering, Thresholding, Model-based. Among this Threshold is one of the widely methods used for image segmentation. It is useful in discriminating foreground from the background. By selecting an adequate threshold value T, the gray level image can be converted to the binary image. All the essential information about the position and shape of the objects of interest (foreground) should be included in the binary image. The benefit of obtaining first a binary image is that it simplifies the process of recognition and classification and also decreases the complexity of the data [1]. The most common way to change a gray-level image to a binary image is to select a single threshold value (T). Then all the gray level values below this T will be classified as black (0), and that above T will be white (1). The segmentation problem becomes one of selecting the correct value for the threshold T. A common method used to select T is by analysing the histograms of the type of images that want to be segmented. When the histogram presents only two dominant modes and a clear valley (bimodal) then it is considered as the ideal case [4]. In real-world applications, histograms are more complex, where valleys are not clear and which consist of many no of peaks and it is not always easy to select the value of T [4].

Algorithm of Thresholding:

1. Estimate value of T (start with mean)
2. Divide histogram into two regions, R1 and R2 using T
3. Calculate the mean intensity values 1 and 2 in regions R1 and R2
4. Select a new threshold T = (1 + 2)/2
5. Repeat 2-4 until the mean values 1 and 2 do not change in successive iterations 1, if f (x, y)>T G (x, y) = f(x) = 0, if f (x, y) ≤ T (1) Any point (x, y) in the image at which f (x, y)> T is called an object point; otherwise, the point is called a background point [4].

C) Feature Extraction

In image processing, processing huge amount of data is very tedious and time-consuming. It may not necessarily add efficiency for classification. However, to make it simpler, the input data is transformed into a reduced set of feature vector. This process is called as feature extraction. The feature vector is used as input for classification [5] [6].

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

Features can be classified into three classes, color, texture, and shape [5] [7] [9] [11]. Texture features plays a very important role in feature extraction phase, and for extraction of features; Gray Level Co-occurrence Matrix (GLCM) is used, as it has been proven to be a powerful tool for feature extraction. In this work, four types of textural features are considered namely contrast, correlation, energy and homogeneity and are extracted with and=00, 450, 950, 1350 [6] [7] [10]. Lastly the average is being taken of these four directions.

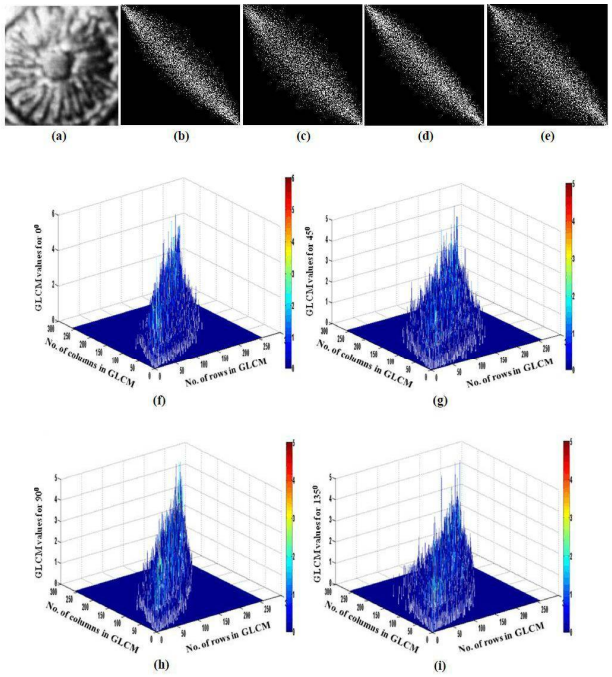


Fig. 4 Shows the (a) sample image and the GLCM matrix constructed out of it along four possible directions (b) 0^0 (c) 45^0 (d) 90^0 and (e) 135^0 together with its corresponding mesh plots ((f)-(i))

Gray level co-occurrence matrix(GLCM):

Gray level co-occurrence matrix explains the occurrence of certain gray levels in relation to other gray levels using statistical sampling [8] [9] [10] [11]. The process statement is reproduced as it is in the following paragraph.

Assume that an image to be analysed is rectangular and has Nx rows and Ny columns. The gray level appearing at each pixel is quantized to Ng levels.Letbe the $Lx = \{1, 2, \ldots, Nx\}$ rows, $Ly = \{1, 2, \ldots, Ny\}$ be the columns and $G = \{0, 1, 2, \ldots, Ng-1\}$ is the total number of

gray levels quantized up to Ng levels. The set $Lx \times Ly$ is the set of pixels of the image ordered ther row-column designations. Then, the image can be presented as a function of co-occurrence matrix that assigns some gray level in $Lx \times Ly$ as I: $Lx \times Ly \rightarrow G$.

The texture-context information is represented by the matrix of relative frequencies Pij with two neighbouring pixels separated by distance, one with gray level i and the other with gray level j. Such matrices of gray-level co-occurrence frequencies are a function of the angular relationship θ and distance d between the neighbouring pixels. By using a distance of one pixel and angles quantized to 45° intervals, four matrices of horizontal, first diagonal, vertical, and second diagonal (0, 45, 90 and 135 degrees) are used. Then, the unnormalized frequency in those four directions is defined by equation (2) [6] [7] [11].

$$p(i, j, \theta) =_{\#}$$

$$(k, l), (m, n) \epsilon \, |(Lx \times Ly) \times (Lx \times Ly)|$$

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018

$$I(k,l)=i, I(m,n)=j$$

$$(k-m=0,|l-n|=d) or (k-m=d, l-n=-d) \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$or(k-m=-d, l-n=d) or(|k-m|=d, l-n=0),$$

$$\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where # is the number of elements in the set, (k,l) the coordinates with gray level $i, (m,n)$ the coordinates with gray level j .Consider $p(i,j)$ be the $(i,j)th$ entry in a normalized GLCM. G is the number of gray levels range from 0 to $Ng-1$. μ is the mean value of p. $\mu x, \mu y, \sigma x, \sigma y$ are the means and standard deviations of Px and Py and presented in Equations (3), (4), (5) and (6) respectively [7] [9] [10].

$$\mu x = \sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} i.p(i,j) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

$$\mu y = \sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} j.p(i,j) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

$$\sigma x = \sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} (i-\mu x)^2.p(i,j) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

$$\sigma y = \sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} (j-\mu y)^2.p(i,j) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

D) Classification

Real-time image segmentation is a well-known problem that can be solved using pixel-wise classification and specific classifiers.Classification is a central problem of pattern recognition and many approaches to this problem have been proposed, e.g. neural networks, Support Vector Machines (SVM), k-nearest neighbors (KNN) and kernel-based methods . In many practical cases, it has been shown that the SVM method gives very good results [12] [13].

Texture-based classification

In computer version, it is very hard to describe and recognize a texture and it is also considered as a vital feature for remote sensing image classification. Support vector machine (SVM) is one of the most successfully used algorithms that undertakes advantages of avoiding local optimum, conquering dimension disaster with small samples [12] [13]. The features can be extracted from both the width across the mass contour and the whole mass region but the best results are obtained from the features computed through the width.

Support Vector Machine (SVM

In SVM algorithm the "Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for classification challenges. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyperplane that differentiate the two classes very well. SVM models havesimilar functional form to neural networks and radial basis functions, both popular data mining techniques.However, neither of these algorithms has the well-founded theoretical approach to regularization that forms the basis of SVM. The quality of generalization and ease of training of SVM is far beyond the capacities of these more traditional methods.

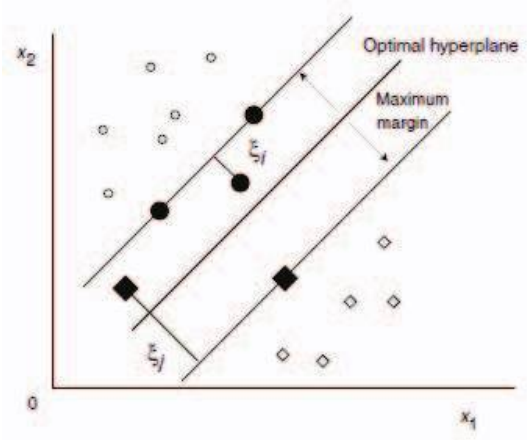
ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 2, February 2018



SVM optimal hyperplane localization.

IV. CONCLUSION

The proposed system is developed for the diagnosis of breast cancer from mammographic images. It includes pre-processing of images to reduce the computational cost and to maximize the probability of accuracy. To summarize the developed method the initial step based on gray level enhancement and gamma correction performs image enhancement and segments the mammographic image. In the second stage GLCM features are extracted and based on these features the mammographic image is classified as: Normal, Benign or Malign.

REFERENCES

[1] Jasmeen Kaur, Mandeep Kaur "Automatic Cancer Detection in Mammographic Images"International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 5, Issue 7, July 2016.
[2] Volodymy Ponomaryov "Computer-aided detection system based on PCA/SVM for diagnosis of breast cancer lesions" Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), 2015 CHILEAN Conference on 28-30 Oct. 2015.
[3] Yanfeng Li, Houjin Chen*, Lin Cao and Jinyuan Ma "A Survey of Computer-aided Detection of Breast Cancer with Mammography" an article by Journal of health and medical informatics (J Health Med Informat 2016).
[4] V. Sivakumar, V.Murugesh "A Brief Study of Image Segmentation using Thresholding Technique on a Noisy Image"
[5] Rajkumar Goel, Vineet Kumar, Saurabh Srivastava, A. K. Sinha "A Review of Feature Extraction Techniques for Image Analysis" IJARCCE-International Conference on Advances in Computational Techniques and Research Practices Noida Institute of Engineering & Technology, Greater Noida Vol. 6, Special Issue 2, February 2017.
[6] Retno Kusumaningrum and Aniati Murni Arymurthy "Color and Texture Feature for Remote Sensing – Image Retrieval System: A Comparative Study" IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 2, September 2011.
[7] Seyyid Ahmed Medjahed "A Comparative Study of Feature Extraction Methods in Images Classification" I.J. Image, Graphics and Signal Processing, 2015, 3, 16-23 Published Online February 2015 in MECS.
[8] Snehal A. Mane, Dr. K. V. Kulhalli "Mammogram Image Features Extraction and Classification for Breast Cancer Detection." International Research Journal of Engineering and Technology (IRJET) Volume: 02 Issue: 07 | Oct-2015.
[9] Computerized Analysis of Mammographic Images for Detection and characterization of breast cancer,a book by Paola Casti, Arianna Mencattini, Marcello
[10] Ranjit Biswas, Abhijit Nath, Sudipta Roy "Mammogram Classification using Gray-Level Co-occurrence Matrix for Diagnosis of Breast Cancer"
[11] Ms. P. V alarmathi & Dr. V. Radhakrishna "Tumor Prediction in Mammogram using Neural Network"Global Journal of Computer Science and Technology Neural & Artificial Intelligence Volume 13 Issue 2 Version 1.0 Year 2013.
[12] K.Menaka, S.Karpagavalli "Breast Cancer Classification using Support Vector Machine and Genetic Programming" International Journal of Innovative Research in Computer and Communication Engineering.
[13] J. Mitéran, S. Bouillant, E. Bourennane "SVM Approximation for Real-time Image Segmentation by Using an Improved Hyper Rectangles-based Method" Le2i - FRE CNRS 2309 Aile des Sciences de l'ingénieur Université de Bourgogne.