INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue2)

Automatic Tumour Detection In Mammogram Using Supervised Learning Method

Chandana Saipriya .V
ECE, Jeppiaar SRR Engineering College
chandana.saipriya96@gmail.com

Dhanalakshmi .B
ECE, Jeppiaar SRR Engineering College
babudhana2468@gmail.com

Gnanasoundari .S
ECE, Jeppiaar SRR Engineering College
soundarisri96@gmail.com

Mercy Therasa .M
ECE, Jeppiaar SRR Engineering College
mercyjesudossa@gmail.com

Hemadevi .J
ECE, Jeppiaar SRR Engineering College
hemajana1011@gmail.com

Abstract: ***Breast cancer is the most occupied type of cancer in women that caused the most deaths among women. The early detection of breast cancer is more important for the chances of survival of patient. This work has mainly four modules: Pre-processing, Segmentation is carried out by Active Contour algorithm and Advanced K-means algorithm, Feature extraction is done by Gray Level Co-occurrence Matrix (GLCM), Expectation Maximization (EM) and Principle Component Analysis (PCA), finally classification is done by Random Forest Classification. To achieve the objective of this work, MIAS (Mammographic Image Analysis Society) and IN breast databases are used as input images. The Accuracy achieved in this system is 95.83%.***

Keywords: Active Contour Segmentation; K-Means Algorithm, Expectation Maximization, Principle Component Analysis, s Random Forest Classification.

I. INTRODUCTION

Breast cancer is the most common invasive cancer in females worldwide. It affects one in eight women during their live. The only possible way to find out the cancer is by Mammography test, which is an X-ray inspection of the breast, used to detect the breast tumour which may lead to breast cancer [1]. Supervised learning is the machine learning task of obtaining function from labelled training data. The training data consist of a set of training examples. The fundamental step in breast cancer detection is to remove the pectoral muscle. Then, texture is the main feature extracted here. Using the Random Forest Classification, the tumour is classified into benign or malignant. Mammographic Image Analysis Society (MIAS) is an organization of UK which analyses the mammograms and it has generated a database of digital mammograms. The database contains 322 digitized films.

II. METHODS AND MATERIALS

A. TECHNOLOGY

The technology used is MATLAB 8.3.0.532(R2014a). This high performance language, for technical computation, apparition and programming is an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation. The typical uses are Math and computation, Algorithm and application development, Data acquirement, Modeling, simulation and prototyping, Data analysis and visualization, engineering graphics.

B. EXISTING SYSTEM

The input image is given to five filters namely, Adaptive Histogram Equalization filter, Median filter, Butterworth filter, frost filter and Wavelet Denoising filter. The denoised image is then subjected to fuzzy C-Means algorithm to detect the suspicious lesion. Finally, morphological operation is used to remove the pectoral muscle [1].

DISADVANTAGES: The existing method is based on the threshold and region growing. In threshold based segmentation, the image is considered as having only two values either black or white. But the bit map image contains 0 to 255 gray scale values. So

sometimes it ignores the tumour cells. In region growing based segmentation, it needs more user interaction for the selection of the seed. Seed is nothing but the centre of the tumour Cells; it may cause intensity in homogeneity problem and also it will not provide the acceptable result for all the images. The execution time is much longer.

C. PROPOSED SYSTEM

The proposed system has mainly four modules: Pre-processing, Segmentation is carried out by Region Based Active Contour segmentation and advanced K-means algorithm. Feature extraction is done by GLCM, EM and PCA. Random Forest Classification is used to classify the tumour as benign or malignant. This system has 89.47% sensitivity.

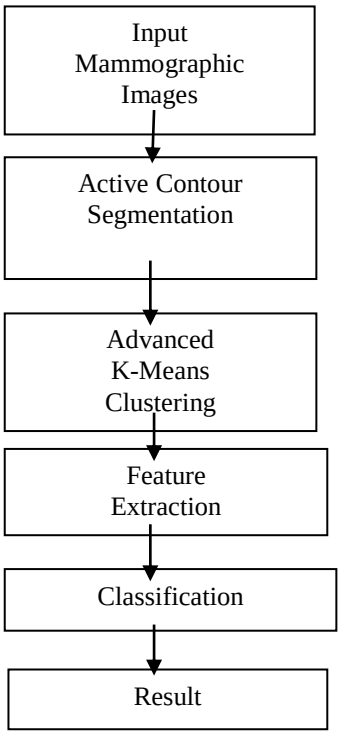


Figure 1.Block diagram of existing system

D. INPUT MAMMOGRAPHIC IMAGE

A grayscale image is made up of pixels, each of which holds a single number corresponding to the gray level of the image at a particular location. These gray levels contain the full range from black to white in a series of very fine steps, normally 256 different grays. Each black and white pixel can be stored in a single byte (8 bits) of memory. 100 mammographic x-ray images have been taken as input from MIAS database.

E. PRE-PROCESSING

Mammograms are medical images that are difficult to understand. Hence pre-processing is essential to uplift the quality. It will prepare the mammogram for the next two process segmentation and feature extraction. According to the need of the next level, the pre processing step converts the image. It performs filtering of noise and other artifacts in the image and sharpens the edges in the image. RGB to grey conversion and Reshaping also takes place here

.

F. ACTIVE CONTOUR SEGMENTATION

Image segmentation is used to trace objects and boundaries like lines, curves, etc. in images. Image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual uniqueness. The result of image segmentation is a set of segments that collectively cover the entire image, or a set of contours extracted from the image [2]. Region based Active contour segmentation looks for the similarity within the sub region based on the desired property. The active contour model is defined by an energy function [3]. The energy function, which is minimized, is a combination of internal and external forces. The internal forces originate from the shape of the snake, while the external forces appear from the image and from higher-level image understanding process. The Chan and Vese proposed a new active contour model which segments the area without edges. The proposed model can easily recognize the individual segments in the input image with multiple segments and junctions, as compared the value with the initial segment, where the detected area belongs to the same segment.

G. K-MEANS SEGMENTATION

In this algorithm, each object is assigned to exactly one of a set of clusters. Objects in one cluster are similar to each other. The resemblance between objects is based on a measure of the distance between them [3]. The K-means operator assigns observations to clusters to increase the distances between the clusters. The main idea behind the K-means algorithm is the minimization of an object usually taken up as a function of the deviations between all patterns from their respective cluster centres [4]. The K-means

algorithm partitions a dataset into k predefined number of clusters that will try to reduce the intra-cluster distance based on Euclidean distance. K-means algorithm is very fast and simple.

H. GRAY LEVEL CO-OCCURRENCE MATRIX

A set of texture feature functions was applied to a set of 100 digitized mammograms in particular regions of interest. Here, we have focused on Grey Level co-occurrence matrices in four different directions to provide the texture- context information [5]. The digitized sample consists of 100 mammographic images originating from the MIAS dataset has been randomly selected. The region of interest has been selected which contains the suspicious region of interest (ROI). Then the feature selected from the ROI and statistical texture features are calculated for each ROI. GLCM are calculated by observing pairs of image cells distance d from each other and increasing the matrix position corresponding to the gray level of both cells. This is used to derive four matrices for each given distance and four different directions.

There is Ng dimension of square matrix of the Grey Level Co-occurrence matrix, where the dimension Ng values between 0 and 255 gray levels. The element (i, j), of the square matrix is generated by counting the number of times a pixel with value 'i' is nearby to a pixel with value 'j' and then dividing the entire matrix by the total number of such comparisons made. Each output is therefore considered to be the probability that a pixel with value 'i' will be found next to a pixel of value 'j'. Adjacency is said to occur in four directions in a 2D, square pixel image as shown in Fig 1.

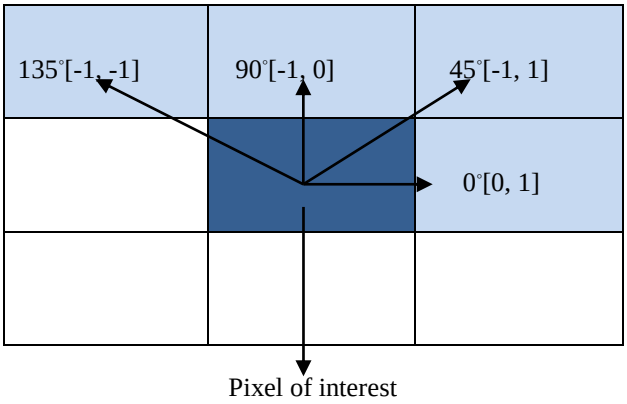


Fig. 1 Adjacency of pixel in four directions (horizontal, vertical, left and right diagonals)

I. EXPECTATION MAXIMIZATION

EM clustering algorithm detects the probabilities of cluster memberships based on one or more probability distributions. The goal of the clustering algorithm then is to increase the overall probability are likelihood of the data, given the (final) clusters [6]. Suppose that the sample consists of two clusters with different means within each sample, the distribution of values for the continuous variable will follow the normal distribution. EM clustering estimates the means and standard deviations for each clusters so as to increase the probability of the observed data. In other words, each observation belongs to each cluster with a same probability.

J. PRINCIPAL COMPONENT ANALYSIS

The main objective of feature selection is to minimize the dimensionality of the data in order to improve the sensitivity and reduce the computational time of the classifier. The minimization can be applied by Principal component analysis (PCA) which is widely applied on datasets. It is a linear dimensionality reduction method from which we determine a minimal feature subset from the entire set of features [8].

K. RANDOM FOREST CLASSIFICATION

It is an ensemble learning method for classification deterioration and other tasks that operate by forming a multitude of decision trees at training time and giving the class that is the mode of the classes (classification) or mean prediction of the individual trees. It was used to minimize the prediction error that is generated by standard trees in other classifications.

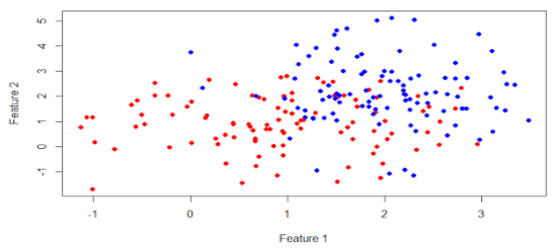
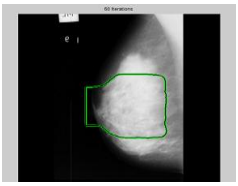
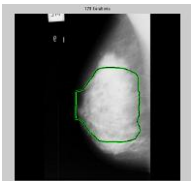
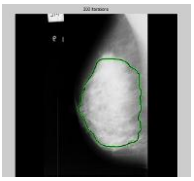
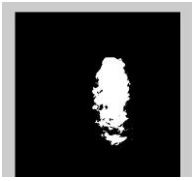


Fig. 2 Random Forest Classification

III. RESULT AND DISCUSSION

Thus, the tumour in the given images have been segmented from the pectoral muscle, features like contrast and homogeneity have been extracted and compared using GLCM, PCA , EM and the tumour cells are classified as normal, benign and malignant. Since the sensitivity is 89.47% and specificity is 100%, the tumour should be extracted from the root cells and the need for mastectomy can be avoided and for future enhancement, stem cells can be used for the regrowth of the normal breast cells. The following are the results of this work.

The following table represents the output of our proposed system. In the table, 'a', 'b', 'c' represents the active contour segmentation in 60, 120 and 200 iterations. Similarly, 'd' represents the K-means algorithm output. 'f' represents the morphological operation output where only the tumour cell is shown .Finally, 'g' represents the classification output whether the given mammography image contains normal, benign and malignant cells respectively.

PARAMETERS	OUTPUT(NORMAL)
60 Iterations (a)	
120 Iterations (b)	
200 Iterations (c)	
Segmentation (d)	
K-means algorithm (e)	
Morphological Operation (f)	

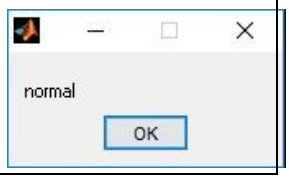
Classification (g)	

Fig. 3 output for NORMAL cells

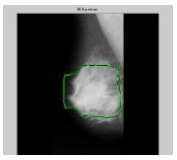
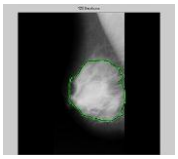
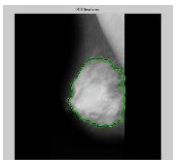
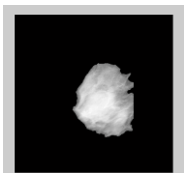
PARAMETERS	OUTPUT(BENIGN)
60 Iterations (a)	
120 Iterations (b)	
200 Iterations (c)	
Segmentation (d)	
K-means algorithm (e)	
Morphological Operation (f)	
Classification (g)	

Fig. 4 output for BENIGN cells

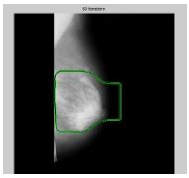
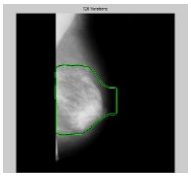
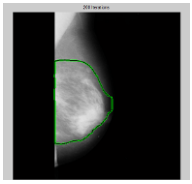
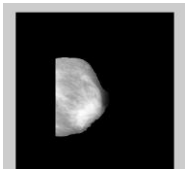
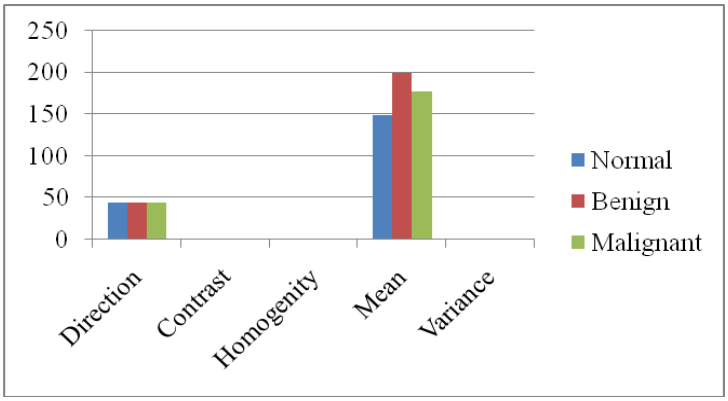
PARAMETERS	OUTPUT(MALIGNANT)
60 iterations (a)	
120 iterations (b)	
200 iterations (c)	
Segmentation (d)	
K-means algorithm (e)	
Morphological Operation (f)	
Classification (g)	

Fig. 5 output for MALIGNANT cells

Table 1.Actual and Proposed System Result

IMAGE NUMBER	ACTUAL RESULT OBTAINED FROM MIAS DATASET	PROPOSED SYSTEM RESULT
mdb001	B	B
mdb002	B	B
mdb003	N	N
mdb004	N	N
mdb005	B	N
mdb006	N	N
mdb007	N	N
mdb008	N	N
mdb009	N	N
mdb011	N	N
mdb012	B	B
mdb013	B	B
mdb014	N	N
mdb016	N	N
mdb017	B	B
mdb018	N	N
mdb020	N	N
mdb022	B	N
mdb023	M	M
mdb024	N	N
mdb025	B	B
mdb026	N	N
mdb028	M	M
mdb029	N	N
mdb030	B	B
mdb033	N	N
mdb034	N	N
mdb035	N	N
mdb036	N	N
mdb037	N	N
mdb082	N	N
mdb084	N	N
mdb089	N	N
mdb090	M	M
mdb091	B	B
mdb092	M	N
mdb094	N	N
mdb097	B	B
mdb098	N	N
mdb099	B	B
mdb101	N	N
mdb102	N	M
mdb103	N	N
mdb104	B	N
mdb105	M	M
mdb106	N	N
mdb107	B	B
mdb108	N	N
mdb109	N	N
mdb110	M	M



N-Normal B-Benign M-Malignant
TP (Sick people correctly identified as sick) = 17
FP (healthy people incorrectly identified as sick) = 0
TN (healthy people correctly identified as healthy) = 29
FN (sick people incorrectly identified as healthy) = 2
Sensitivity = TP/ (TP+TN) = 89.47%
Specificity = TN/(FP+TN) = 100%
Accuracy = (TP+TN)/(TP+FP+FN+TN) = 95.83%

Table2.GLCM features

	Direction	Contrast	Homogeneity	Mean	Variance
Normal	0	0.1225	0.99724	132.1165	1.4373
Benign	0	0.09119	0.99814	181.4741	1.4414
Malignant	0	0.10905	0.997475	154.2169	1.4061
Normal	45	0.14929	0.99697	148.9484	1.4594
Benign	45	0.11437	0.99794	199.4182	1.4506
Malignant	45	0.123825	0.99745	177.792	1.5542
Normal	90	0.09471	0.99775	152.7146	1.5643
Benign	90	0.07296	0.99844	125.6088	1.5197
Malignant	90	0.093525	0.99775	160.1459	1.4115
Normal	135	0.15956	0.99676	157.7334	1.3883
Benign	135	0.11783	0.99777	154.9458	1.3997
Malignant	135	0.156025	0.996775	230.4593	1.5162

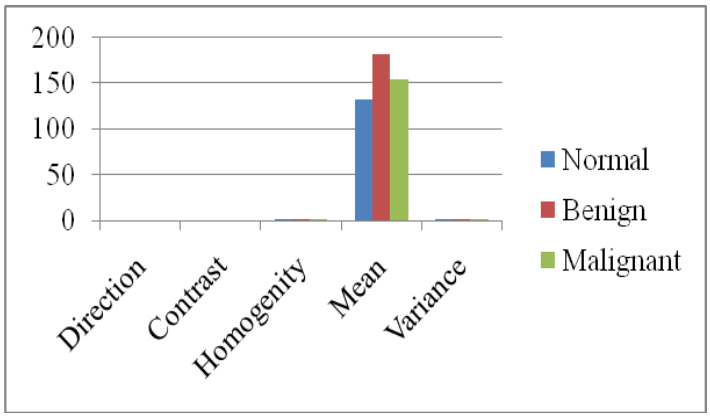


Fig. 6 GLCM features at 0° direction

Fig. 7 GLCM features at 45° direction

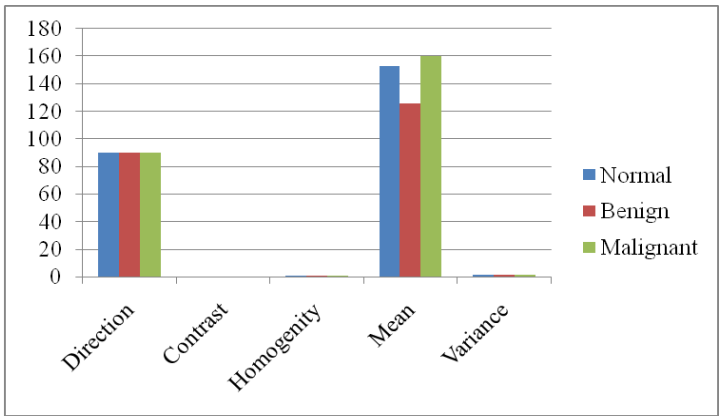


Fig. 8 GLCM features at 90° direction

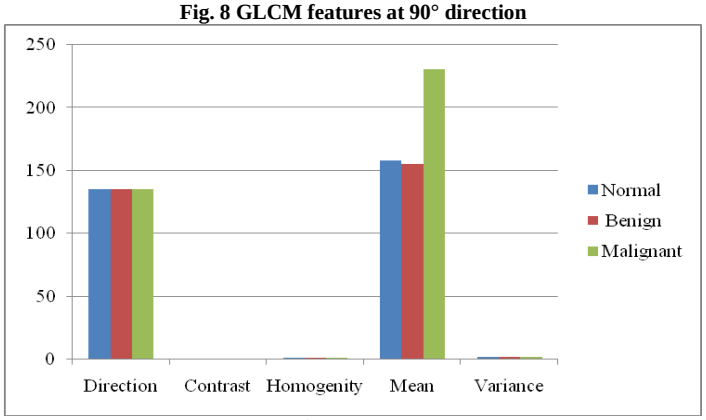


Fig. 9 GLCM features at 135° direction

The Table 2 represents the excel sheet calculation of GLCM features such as contrast, homogeneity, mean and variance at 0°,450°,90°,135°,respectively.The figures 6, 7 ,8 and 9 are the graph plotted for the table.

REFERENCE

[1] Detection of Suspicious Lesions in Mammogram using Fuzzy C-Means Algorithm Mukesh Kumar1, V.M. Thakkar2, Upendra Bhatt3, Neema Soliyal, 2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016, Jaipur, India

[2] International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 2, February 2012) 204 , Analysis of Mammogram Images using Active Contour Segmentation process and Level Set Method Mr. L.Megalan Leo1, Mr. A.Ranjith2, Mr. R.Thandaiah Prabu3

[3] Active Contours and Image Segmentation: The Current State of the Art, Global Journal of Computer Science and Technology Graphics & Vision Volume 12 Issue 11 Version 1.0 Year 2012 Type: Double Blind Peer Reviewed International Research Journal Publisher: Global Journals Inc. (USA) Online ISSN: 0975-4172 & Print ISSN: 0975-4350

[4] Implementation and analysis of k- means and fuzzy c means clustering techniques Snehali d. sable ,Student, M.E. (VLSI and Embedded) GHRIET, Pune India

[5] Detection of Masses in Digital Mammogram Using Second Order Statistics and Artificial Neural Network, International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 3, June 2011, Al Mutaz M. Abdalla1, Safaai Dress2, Nazar Zaki3

[6] Identification of masses in digital mammograms with MLP and RBF Nets, Keir Bovis1, Sameer Singh1, Jonathan Fieldsend1, Chris Pinder2, 1pann research, department of computer science, university of exeter, exeter, uk 2breast care unit, royal devon & exeter hospital, Exeter, UK

[7] Automatic Mammographic Mass Segmentation based on Region Growing Technique, 3rd International Conference on Electronics, Biomedical Engineering and its Applications (ICEBEA'2013) April 29-30, 2013 Singapore, , K.Yuvaraj, and U.S.Ragupathy

[8] Image Processing using Principle Component Analysis, International Journal of Computer Applications (0975 – 8887) Volume 15– No.4, February 2011 37

