Mammogram Image Segmentation by Watershed Algorithm and Classification through k-NN Classifier

B.N. Beena Ullala Mata and Dr.M. Meenakshi

Abstract--- *This paper presents a novel approach to detect the tumors in the mammogram images based on watershed algorithm. To increase the performance of the classifier, watershed algorithm combined with K-NN classifier is implemented. The gray level co-occurrence matrices (GLCM'S) are obtained from the mammogram images, through the extraction of Halarick's texture features are classified. American Society of cancer, UK, provides the benchmark data, MIAS (Mammographic Image Analysis Society) database for the validation of proposed algorithm. These images are used for further analysis by classification into three categories using the algorithms. Mammogram abnormalities are found to be detected using the proposed algorithm with the available ground truth given in the data base (mini-MIAS database), the accuracy obtained is as high as 83.33%.*

Keywords--- *Halarick's Texture Features, k-NN, MIAS.*

I.INTRODUCTION

BREAST imaging is considered as a medical imaging technique for the early detection of lesions, in recent days [1,2]. The survival rate of women improves by 95% with the early detection of breast cancer. The effective way of screening method for early detection, approved by FDA is mammography, a medical imaging modality. Watershed algorithm is used for mammogram analysis, as mentioned in the literature, with respect to various algorithms [3].

Micro calcification is detected by the of mammogram images, using the said algorithm by the authors of [3]. The computational time allotted increases for the analysis of each images, since this work does not include the classification of mammogram images. Thus a novel algorithm is proposed which combines k-NN classifier along with watershed algorithm, overcoming the above said problem [4].

The CAD system involves automatic classification into normal and suspicious mammograms as a pre-screening protocol for viewing the suspicious mammograms, viewed by radiologists as a pre-screening protocol. The breast density pattern classification of mammograms is required and necessary to increase the accuracy of the detection algorithm since initial screening is limited by sensitivity of the system.

The above two algorithms are combined to filter out the normal images from analysis process. For this process of implementation, with 60 mammogram images from the MIAS (Mammographic Image Analysis Society) database are evaluated. An expert radiologist has marked the ground truth as normal, benign and malignant for this database.UK research groups have produced a digital mammography database for understanding of mammograms and the organization is named as the Mammography Image Analysis Society. The specifications of the X-ray films in the database which have been digitized from the United Kingdom National Breast Screening Programme mentioned below. The process of digitization involves Joyce-Lobel scanning microdensitometer to a resolution of 50¦ˋIm ¡ Ά50 ¦ ˋIm, a device linear in the optical density range 0-3.2 and representing each pixel with an 8-bit word. The database contains left and right breast images for 161 patients, and is available on a DAT-DDS tape. It consists of 322 images categorized into three types namely Normal, benign and malignant. There are 208 normal, 63 benign and 51 malignant (abnormal) images. According to breast density, these mammograms have been classified and categorised as shown in Figure 1.

1. Fatty
2. Glandular and
3. Dense

A radiologist's 'truth'-markings on the locations of any abnormalities that may be present are also included. For each film, an experienced radiologist has given the type, location, scale, and other useful information of them. According to these experts' descriptions, the database is consisting of four kinds of abnormalities (architectural distortions, stellate lesions, circumscribed mass and calcifications).

II.ARCHITECTURE

The method used in implementation is shown in the block diagram in figure 1.

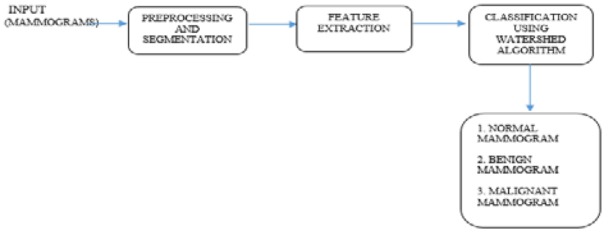


Figure 1: Steps Used in Implementation

B.N. Beena Ullala Mata, Associate Professor, B.M.S College of Engineering. E-mail:bansbeena@gmail.com

Dr.M. Meenakshi, Department of Instrumentation Technology, Dr.Ambedkar Institute of Technology. E-mail:meenakshi_mbhat@yahoo.com

DOI:10.9756/BIJAIP.8352

A. Pre-processing

The following steps are involved in the pre-processing of the images.

a).Enhancement

A contrast enhancement filter is a two dimensional filter h, of type unsharp is an appropriate filter used in this process this returns the correlation kernel which is a multidimensional array and has an image to be filtered. The output of each element is a double precision floating point value that has to be computed. The above said filter is of type replicate, where input array values outside the bounds of the array equals the nearest array border value as assumed.

b).Filtering

Sobel, Prewitt and Krish are the three, two-dimensional filter h which emphasizes the horizontal edges. Among them, Sobel filter, which returns a correlation kernel, is used in the appropriate form as a filter. Using this kernel, in both X and Y directions, the image is filtered. After filtering, the gradient magnitude is calculated by combining the above filters. The results of pre-processing steps are shown in figure 2.1

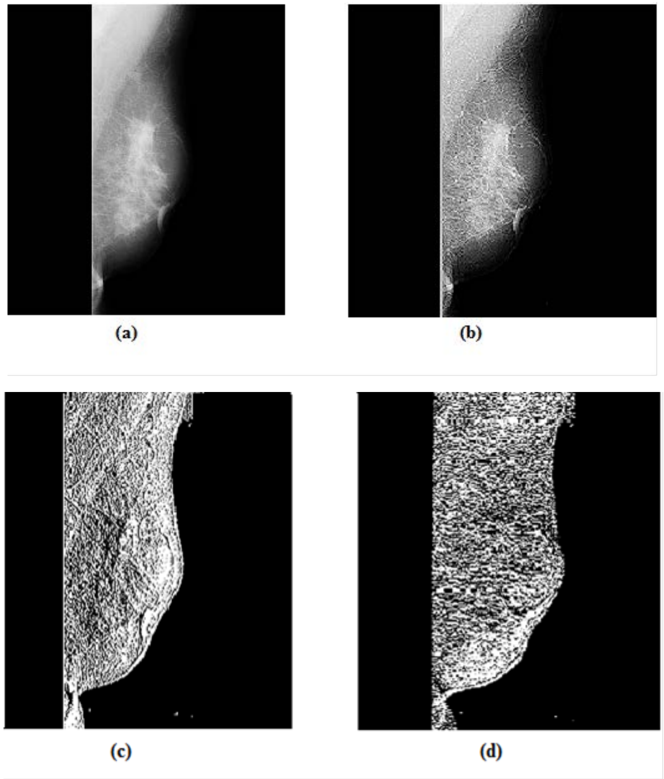


Figure 2.1 Pre-processing Images (a) Original Image

(b) Sharpened Image(c) Filtered Image in X Axis

(d) Filtered Image in Y Axis

B. Segmentation

The process of segmentation is used to detect abnormalities and to detect abnormalities about the breast. In general, in mammogram images, the gray level between pectoral muscles and the tumour is very close. It is generally difficult to apply simple region based segmentation technique. In the watershed algorithm, the exact boundary of any abnormal objects in the images can be determined through the gray scale images. For the purpose of region growing in a class of image segmentation techniques, recursive watersheds are used.

The application of watershed algorithms repetitively on gray scale images is discussed, so that it finally gives the exact boundary of any abnormal objects in the images. The recursive watersheds belong to the region growing class of image segmentation scheme.

Watershed Principle

The introduction of watershed as a morphological tool is due to H Digabel and C Lantuejoul. In terms of topography, the watershed can be imagined as the high mountain that separates two regions. Each region has its own minimum and if a drop of water falls on one side of watershed, it will reach the minimum of the region. The regions that the watershed separates are called the catchment basins and, at points where water coming from different basins would meet, dams are built.

The principle of the watershed technique is to transform the gradient of a gray level image in a topographic surface. This algorithm stops when catchment basins from two different sources meet. As a result, the landscape is partitioned into regions or basins separated by dams, called watershed lines or simply watersheds. Figure2.2.1 shows one-dimension flooding process that shows the catchment basin points. The gray scale image is considered as a topographic relief where brightness value is considered as physical elevation. The technique may be explained as if water rush through a hole if a paper with holes is submerged in water, hence the name watersheds.

There are a lot of local minima formed in the real images and are often noisy. This is the main drawback of initial watershed methods. This leads to over-segmentation. In order to avoid the over-segmentation problem, a classical solution is to use markers inside and outside the tumour. When simulating this process for image segmentation, two approaches may be used: either one first finds basins, then watersheds by taking a set complement; or one computes a complete partition of the image into basins, and subsequently finds the watersheds by boundary detection. Following are the steps used in this algorithm.

1. Read the image and convert to gray scale.
2. Develop Gradient images using appropriate edge detection method.
3. Mark the foreground objects using morphological reconstruction.
4. Calculating the regional maxima and minima to obtain the good forward markers.
5. Superimpose the foreground marker image on the original image.
6. Clean the edges of the markers using edge reconstruction.
7. Compute the background markers.
8. Compute the watershed transformation of the function.

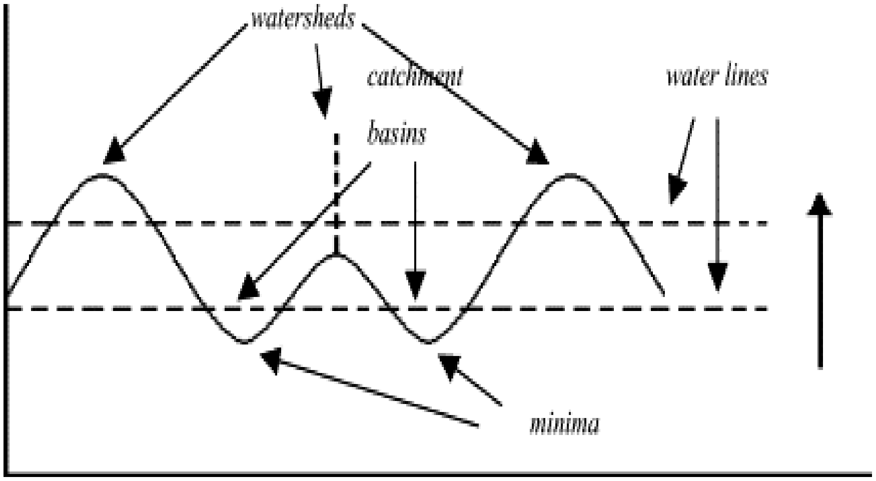


Figure 2.2.1: One-dimension Flooding Process

The Gray- scale Dilation and Erosion

A three dimensional data set is considered, the first two elements are the x and y coordinates of the pixel and the third element is gray-scale value in a gray scale image. A structuring element can be considered and applied to the gray scale image. With this concept, gray dilation can be defined as follows. Gray-scale dilation of the f by b, is defined as

$$(f \oplus b)(s,t) = \max\left\{ f(s-x, t-y) - \frac{b(x,y)}{(s-x)}, (t-y) \in D_f; (s,y) \in D_b \right\}, \quad (2.2.1)$$

Where Df and Db are the domains of f and b respectively.

Gray-scale erosion is defined as,

$$(f \Theta b)(s,t) = \max\left\{ f(s-x, t+y) - \frac{b(x,y)}{(s+x)}, (t+y) \in D_f; (s,y) \in D_b \right\}, \quad (2.2.2)$$

The Gray-scale Opening and Closing

The opening of a gray image f by a structuring element b, is defined as

$$f \circ b = (f \Theta b) \oplus b, \quad (2.2.3)$$

And closing can be defined as,

$$f \bullet b = (f \oplus b) \Theta b.$$

Figure2.2.2 (a), (b), (c) and (d) shows the foreground objects markings in mammogram images after opening and closing of image operation followed by erosion and dilation based reconstruction applied on gradient image respectively

Creating Markers

The marker controlled watershed segmentation has been shown to be robust and flexible method for segmentation of objects with closed contours, where boundaries are expressed as ridges. The marker image used for watershed segmentation is a binary image consisting of either marker points or larger marker regions, where each connected marker is placed inside an object of interest. Each initial marker has a one-to-one relationship to a specific watershed region, thus the number of markers will be equal to the final number of watershed regions. After segmentation, the boundaries of the watershed regions are arranged on the desired ridges, thus separating each object from its neighbors.

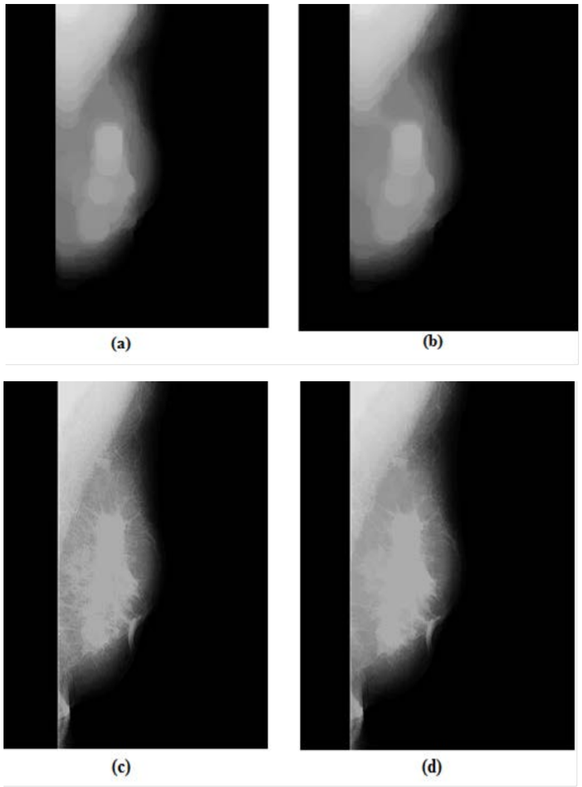


Figure 2.2.2: Foreground objects markings in mammogram image (a) Opening of image (b) Closing of image (c) Erosion based reconstruction applied on the gradient images (d) Dilation based reconstruction applied on the gradient images

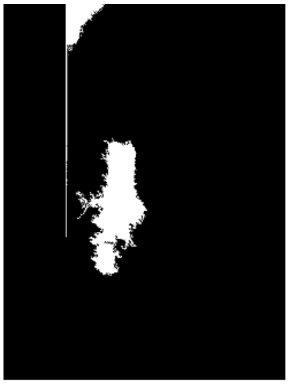


Figure 2.2.3: Background Object

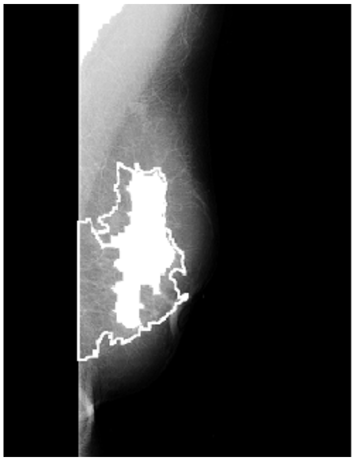


Figure 2.2.4: Watershed Segmentation on Gradient Image

The watershed algorithm is helpful to segment the objects that touch each other in an image. The watershed segmentation based on the markers and simple morphology allows a regularization of the watersheds, and is a flexible approach for further optimization parameters. Thus the results gave better identification of desired objects (tumor) than the standard algorithm. Figure 2.2.4 shows the results with markers on the mammogram images using watershed algorithm after segmentation.

C. Feature Extraction

The resources available as an image has a large set of pixel data. This has to be accurately described and hence involves feature extraction. When analysing the form the feature extraction, large number of variables are involved. Hence, the amount of memory and computational power are generally larger for the analysis. Classification algorithms are implemented to reduce the training samples into the new samples. This process called as feature extraction is a method of constructing combinations of variables. This overcomes the problem of data with sufficient accuracy.

Halarick's approach has been employed for determining texture. Gray level co-occurrence matrix texture measurements has been proposed by Haralick in the 1970s for texture feature of image. A co-occurrence matrix or co-occurrence distribution (less often co-occurrence matrix or co-occurrence distribution) is a matrix or distribution that is defined over an image to be the distribution of co-occurring values at a given offset. Mathematically, a co-occurrence matrix C is defined over an n x m image I, parameterized by an offset (Δx, Δy), as

$$C_{\Delta x,\Delta y}(i,j)=\sum_{p=1}^{n}\sum_{q=1}^{m}\begin{cases}1, & \text{if } I(p,q)=i \text{ and } I(p+\Delta x,q+\Delta y)=j\\ 0, & \text{otherwise}\end{cases} \qquad (2.3.0)$$

By constructing Grey Level Co-occurrence Matrices Dependency (GLCMs) using the directions {0°, 45°, 90°, 135°}, 15 features are extracted. These features include angular second moment, contrast, homogeneity, inverse difference moment, sum average, variance, sum entropy, entropy, difference average, difference variance, difference entropy and information measure of correlation. Following extraction of features from the four GLCMs the mean of each feature is extracted to be used as a feature to train and test the k-NN classifier. All features are normalized to zero mean.

Description of Features

1. Angular Second Moment (ASM): ASM is defined as a measure of homogeneity in an image.

$$\text{ASM} = \sum_{i,j\text{€} G}(i,j)^2 \qquad (2.3.1)$$

2. Contrast: Contrast shows the amount of local variation and is the opposite of homogeneity.

$$\text{Contrast} = \sum_{i,j\text{€} G}(i,j)^2\,.\,\mathrm{C(i,j)} \qquad (2.3.2)$$

3. Correlation: This measure analyses the linear dependency of gray level of neighbouring pixels.

$$\text{Correlation} = \frac{\sum_{i,j\text{€}G}\mathrm{ijC(i,j)} - m_x m_y}{S_x S_y} \qquad (2.3.4)$$

Where $m_x = \sum_i i \quad \sum_j C(i,j)$

$$m_y = \sum_j j \quad \sum_i C(i,j)$$

$$S_x{}^2 = \sum_i i^2 \sum_i C(i,j) - m_x{}^2$$

$$S_y{}^2 = \sum_j j^2 \sum_j C(i,j) - m_y{}^2$$

4. Inverse Difference Moment: It is the inverse of contrast and is a measure of the local uniformity present in an image.

$$\text{Inverse Difference Moment} = \sum\quad\sum \frac{C(i,j)}{[1+(i-j)*2]} \qquad (2.3.5)$$

5. Entropy: It is a measure of randomness in an image.

$$\text{Entropy} = \sum_{i,j\text{€}G}\mathrm{C(i,j).\log[C(i,j)]} \qquad (2.3.6)$$

D. Classification

To classify the data a simple k-NN classifier is used that works on the principle of classifying objects based on closest training examples. The k-nearest neighbour algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its k nearest neighbours (k is a positive integer). If k = 1, then the object is simply assigned to the class of its nearest neighbour.

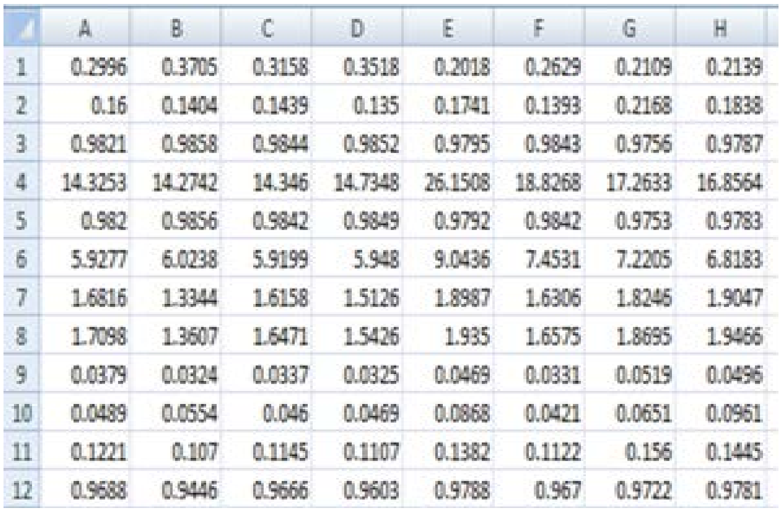
A *k*-NN classifier decides the class of an object by analyzing its k nearest neighbours within the training objects. *k*-NN classifiers are well-suited to solve the given problem because they do not have to spend additional effort for distinguishing additional classes. In the initial training phase, characteristic properties of typical image features are isolated and, based on these, a unique description of each classification category, *i.e. training class*, is created. In the subsequent testing phase, these feature-space partitions are used to classify image features.

The k-NN classifier classifies the rows of the data matrix Sample into groups, based on the grouping of the rows of Training. Sample and Training must be matrices with the same number of columns. Group is a vector whose distinct values define the grouping of the rows in Training. Each row of Training belongs to the group whose value is the corresponding entry of Group.

Table 2.4.1 and table 2.4.2 shows the sample data used in training data group. The k-nearest-neighbor classifier is based on the Euclidean distance between a test sample and the specified training samples are used. Let $\mathbf{x}_i$ be an input sample with p features $(x_{i1}, x_{i2}, \ldots, x_{ip})$ n be the total number of input samples ($i = 1, 2, \ldots, n$)with p the total number of features $(j = 1, 2, \ldots, p)$. The Euclidean distance between sample $\mathbf{x}_i$ and $\mathbf{x}_l$ ($l = 1, 2, \ldots, n$) is defined as

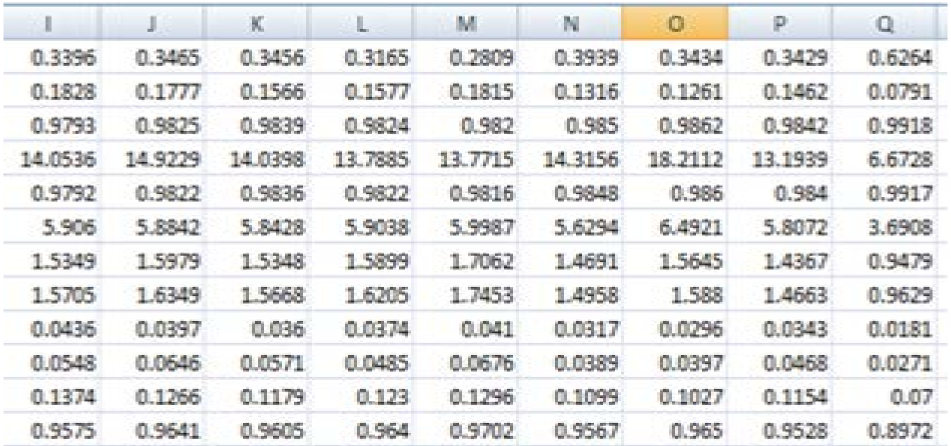
$$d(\mathrm{x}_i, \mathrm{x}_l) = \sqrt{(x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2 + \cdots + (x_{ip} - x_{lp})^2}. \text{ --- (2.4.1)}$$

Table 2.4.1: A Sample Data Used in Training Data Group



	A	B	C	D	E	F	G	H
1	0.2996	0.3705	0.3158	0.3518	0.2018	0.2629	0.2109	0.2139
2	0.16	0.1404	0.1439	0.135	0.1741	0.1393	0.2168	0.1838
3	0.9821	0.9858	0.9844	0.9852	0.9795	0.9843	0.9756	0.9787
4	14.3253	14.2742	14.346	14.7348	26.1508	18.8268	17.2633	16.8564
5	0.982	0.9856	0.9842	0.9849	0.9792	0.9842	0.9753	0.9783
6	5.9277	6.0238	5.9199	5.948	9.0436	7.4531	7.2205	6.8183
7	1.6816	1.3344	1.6158	1.5126	1.8987	1.6306	1.8246	1.9047
8	1.7098	1.3607	1.6471	1.5426	1.935	1.6575	1.8695	1.9466
9	0.0379	0.0324	0.0337	0.0325	0.0469	0.0331	0.0519	0.0496
10	0.0489	0.0554	0.046	0.0469	0.0868	0.0421	0.0651	0.0961
11	0.1221	0.107	0.1145	0.1107	0.1382	0.1122	0.156	0.1445
12	0.9688	0.9446	0.9666	0.9603	0.9788	0.967	0.9722	0.9781

Table 2.4.2: A Sample Data Used in Training Data Group

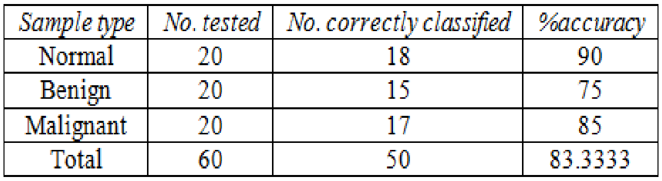


I	J	K	L	M	N	O	P	Q
0.3396	0.3465	0.3456	0.3165	0.2809	0.3939	0.3434	0.3429	0.6264
0.1828	0.1777	0.1566	0.1577	0.1815	0.1316	0.1261	0.1462	0.0791
0.9793	0.9825	0.9839	0.9824	0.982	0.985	0.9862	0.9842	0.9918
14.0536	14.9229	14.0398	13.7885	13.7715	14.3156	18.2112	13.1939	6.6728
0.9792	0.9822	0.9836	0.9822	0.9816	0.9848	0.986	0.984	0.9917
5.906	5.8842	5.8428	5.9038	5.9987	5.6294	6.4921	5.8072	3.6908
1.5349	1.5979	1.5348	1.5899	1.7062	1.4691	1.5645	1.4367	0.9479
1.5705	1.6349	1.5668	1.6205	1.7453	1.4958	1.588	1.4663	0.9629
0.0436	0.0397	0.036	0.0374	0.041	0.0317	0.0296	0.0343	0.0181
0.0548	0.0646	0.0571	0.0485	0.0676	0.0389	0.0397	0.0468	0.0271
0.1374	0.1266	0.1179	0.123	0.1296	0.1099	0.1027	0.1154	0.07
0.9575	0.9641	0.9605	0.964	0.9702	0.9567	0.965	0.9528	0.8972

III.RESULTS

The results obtained are shown below in Table 3. Normal, Benign and Malignant Mammogram images of 20 each, were used for testing. Thus, 60 images were used in total. This is tabulated as shown in Table 3.

Table 3: Classification Result as Normal, Benign and Malignant



Sample type	*No. tested*	*No. correctly classified*	*%accuracy*
Normal	20	18	90
Benign	20	15	75
Malignant	20	17	85
Total	60	50	83.3333

IV.CONCLUSION

The demonstration of the feasibility of the lesion's detection in the mammogram images were implemented in the proposed work. Differentiating between normal, benign and malignant mammograms were differentiated in the diagnosis of mammograms by the techniques of computer vision. From the results obtained, the following conclusions can be drawn.

- It is found that difficulty in marking the exact area of the lesion since neighboring normal tissue was also included.
- The sensitiveness of the system is more for differentiating as normal mammograms when benign and malignant mammograms are compared. Accuracy of 90% is obtained for the classification of normal mammograms.
- Expert radiologist gives the assessment of the breast as ground truth and hence it can be called as a subjective inter observer result.
- Finally, by extracting additional features, the classification of normal and abnormal breasts can be improved and thus improvement in the sensitivity of the classification algorithm.
- 83.33% is the overall accuracy rate obtained from the proposed work for classifying the mammograms.

- As future work, the CAD system can be developed by a more sensitive detector and a classifier in a system.

REFERENCES

[1] NHS breast screening program review 1999. NHS Breast Screening Program 1999.

[2] R. Highnam and M. Brady, Mammographic Image Analysis, Kluwer Academic Publishers, 1999.

[3] N.F. Boyd, J.W. Byng, R.A. Jong, E.K. Fishell, L.E. Little, A.B.Miller, G.A. Lockwood, D.L. Tritchler and M.J. Yaffe, "Quantitative classification of mammographic densities and breast cancer risk: Results from the canadian national breast screening study", Journal of the National Cancer Institute, Vol. 87, No. 9, Pp. 670–675, 1995.

[4] H.S. Sheshadri and A. Kandaswamy, "Detection of Breast Cancer Tumor based on Morphological Watershed Algorithm", ICGST, International Journal on Graphic, Vision and Image Processing, Vol. 5, 2005.

[5] R.C. Gonzalez and R.E. Woods, Digital Image Processing, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 2001

[6] I. Tomek, "An experiment with the nearest neighbor rule", IEEE Transactions on Information Theory, Vol. 6, No.6, Pp. 448–452, 1976.

[7] D.L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data", IEEE Transactions on Systems, Man and Cybernetics, Vol. 2, No.3, Pp. 408–421, 1972.

[8] Thomas B. Fomby, K-Nearest Neighbors Algorithm: Prediction and Classification, Department of Economics, Southern Methodist University, Dallas, 2008.

[9] D.G. Terrell and D.W. Scott, "Variable kernel density estimation", Annals of Statistics, Vol. 20, Pp. 1236–1265, 1992.

[10] P. Mills, "Efficient statistical classification of satellite measurements", International Journal of Remote Sensing, Vol. 32, No. 21, Pp. 6109-6132, 2011.

[11] F. Nigsch, A. Bender, B. van Buuren, J. Tissen, E. Nigsch and J.B.O. Mitchell, "Melting Point Prediction Employing k-nearest Neighbor Algorithms and Genetic Parameter Optimization", Journal of Chemical Information and Modeling, Vol. 46, No. 6, Pp. 2412–2422, 2006.

[12] T.M. Cover and P.E. Hart, "Nearest neighbor pattern classification", IEEE Transactions on Information Theory, Vol. 13, No. 1, Pp. 21-27, 1967.

[13] David Bremner, Erik Demaine, Jeff Erickson, John Iacono, Stefan Langerman, Pat Morin and Godfried Toussaint, "Output-sensitive algorithms for computing nearest-neighbor decision boundaries", Discrete and Computational Geometry, Vol. 33, No. 4, Pp. 593-604, 2005.