

Breast Cancer Detection and Classification in Early Stage from Digital Mammogram

Samir Kumar Bandyopadhyay
Professor, The Bhawanipur Education Society College, West Bengal

Abstract - Breast cancer nowadays acts as one of the most prevalent life-threatening diseases among women. Early detection of breast cancer helps to improve the prognosis of cancer and in treatment planning. Mammography is the well-known technique for detection of breast cancer. Unnecessary biopsy is time-consuming as well as increases the anxiety of patient. Computer-Aided Diagnosis (CAD) is becoming an important tool for detection and characterization of cancer and also reduces the expenditure of unnecessary biopsy. CAD plays a crucial role as second reader for detection of breast cancer in clinical practice. This paper develops a CAD model capable of locating the suspicious region. The proposed method consists of four steps: preprocessing, segmentation, feature extraction and classification. After segmentation of cancerous region, it is characterized by the hybrid extraction methods i.e. statistical features using first-order histogram and Gray Level Co-occurrence Matrix (GLCM) and Principal Component Analysis (PCA). Classification results of these two methods are compared and high accuracy obtained from GLCM according to the best angle choosing. Based on the classification result, normal and cancerous mammograms have been classified.

Keyword: Mammogram, Computer-Aided Diagnosis (CAD), Statistical features, Gray Level Co-occurrence Matrix (GLCM), Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Principal Component Analysis (PCA)

INTRODUCTION

Breast cancer is one of the major significant problems nowadays. Early detection of breast cancer improves the prognosis of cancer and helps to enhance the success rate of survival [1-4]. Different diagnosis modalities like Computed Tomography (CT), Magnetic Resonance Imaging (MRI) and Ultrasound are available but all are not suitable for breast cancer detection, the most promising technique is mammogram which is used widely by the radiologist. If the abnormalities have been found in mammogram biopsy is recommended. Biopsy is the standard clinical approach to determine the breast cancer by evaluating the suspicious area. Biopsies increase the anxiety of patients, time consuming and raise the health care costs, in that respect CAD is very useful tool which aids radiologists in detecting potential abnormalities and decrease false negative readings [5-6]. According to the report of American Institute for Cancer Research, in 2018 more than 2 million new cases of breast cancer has been diagnosed and highest rate of breast cancer has been found in the top 25 countries like Belgium and Luxembourg. It is estimated that in 2018 the mortality rate of female breast cancer is expected to be about 25.3 per 100,000 women. Every year it is counted that

1.6 million new cases are diagnosed worldwide. It is estimated that in 2016, about 246,660 women were diagnosed with breast cancer. In 2017 it is estimated that about 30% of newly diagnosed cancer in women is the breast cancer [7-10]. 81% of women with age of 50 or older breast cancer has been diagnosed and 89% of death related with this type of cancer this in this age group. In Malaysia mortality rate due to breast cancer is high compared to other main cancers. Prognoses of breast cancer among Asian women vary considerably compared to Western women due to difference in lifestyle, socioeconomic profile, and genetic background [11]. So, it is observed that, past two decades a steep increase in breast cancer in most Asian countries [12-15].

Fast and accurate treatment planning and the assessment of radiotherapy treatment efficacy accurate tumor segmentation play a vital role [16-18]. The developed method provides a powerful detection and classification tool that helps clinicians to automatically and accurately delineate lesions for better diagnosis and treatment [19-22]. This proposed abnormality detection CAD system capable of detection and classification between benign and malignant lesions by reducing the false positive rate and estimate the growth pattern of malignant and benign masses [23-27].

LITERATURE REVIEW

Researchers proposed an algorithm which is composed of background and objects texture segmentation and extraction [28-29]. Local entropy has been used to separate suspicious regions containing the masses from background parenchyma. [30-33] proposed a method of earlier detection of breast cancer, using LBP features. The background image and breast image are separated using Otsu method based on uniform LBP histogram Support vector machine (SVM) which is used to classify the normal and cancerous region. Some researchers [34-35] had used k-means for segmentation and normal and cancerous tissues have been classified using support proposed method, first detects the cancerous area and then segments the respective area [36]. Average filter and thresholding method have been used for detection of cancerous region. Some are employed Local Binary Pattern (CLBP) support vector machine (SVM) has been used for classification purpose [37].

Classification of six different type of breast cancer namely: CALC, CIRC, SPIC, MISC, ARCH, ASYM have been developed [38]. The proposed work is based on LBP feature and SVM classifier. An algorithm for mass detection has been used for reduction of false positive. Sun et al. presented a new multi view scheme for mammographic image analysis, that minimizing the number of false classifications. Manifold Learning method which focuses on Density Segmentation in High-Risk Mammograms has been proposed and discussed in detail the estimation of noise in applications using image processing models and discussed pixel intensity adjustments using x-ray images [39]. Some researchers proposed dynamic contrast enhancement (DCE) which is used for preprocessing of mammogram [40]. Dynamic contrast enhancement method has been used dynamic adaptive histogram equalization, Gaussian filtering and gamma correction techniques.

PROPOSED METHOD

In the study MIAS (Mammographic Image Analysis Society) dataset which is standard and publicly available has been used for our experiment. The proposed methodology has been evaluated by 70 mammograms which belong to two categories: benign and malignant.

Out of 70 mammogram images 35 mammogram images have been diagnosed as normal and 35 as malignant. The size of each image is 1024 x 1024 pixels with 240 microns resolution.

A simple approach has been proposed for detection of cancer region in mammogram. Our proposed methodology has been comprised of three sequential main steps as shown in Figure 1.

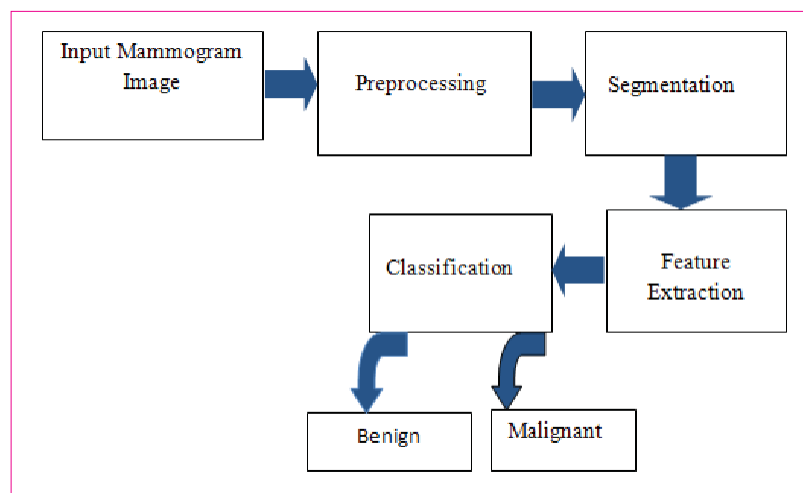


Figure 1. Block diagram of proposed method.

DESCRIPTION OF PROPOSED METHOD

Three major steps are involved for the designing of the entire system such as

(a) Preprocessing; (b) Segmentation, and (c) Feature Extraction. Detail descriptions of steps for each module are as follows:

▪ **Preprocessing.** Image preprocessing is the first and most important step as noise and unwanted distortions present in image, so to improve and enhance the image quality for human interpretation, preprocessing is required.

Procedures.

• (i) Convert the original mammogram image into gray scale image and image resizing (256 X 256) has been performed.

• (ii) Noise degrades the quality of image so noise removal is an essential step to restore and analyze the image to its original state.

• (iii) Homomorphic filter has been applied on image to enhance the contrast and compress the range of brightness.

A hybrid approach has been proposed in Homomorphic filtering process. The steps of this hybrid approach are as follows:

Step1: Input noise free image.

Step 2: Fourier Transform has been applied.

Step 3: Log Transformation has been performed on the resultant image.

Step 4: Inverse Fourier Transform has been done on the output image.

Step 5: Finally Exponential is applied to get the desired output.

The transformations are made.

Transform.

• (iv) Tophat and Bottomhat transform have been used to smooth the borders and objects of the resultant image.

• (v) Adaptive histogram equalization has been applied after combining the result of Tophat and Bottomhat transform, will

help to enhance the local contrast.

▪ **Segmentation.** After performing the preprocessing step, ROI (Region of Interest) selection is the next major step. Background and pectoral muscles are not the part of breast so these unwanted regions must be removed to increase the performance as well as to focus on the region where the probability of cancer exist. The steps of segmentation are as follows:

- Step1. Thresholding
- Step2. Morphological operations and Edge detection
- Step 3. K-Means Clustering.

▪ **Thresholding.** Thresholding is the simplest method where pixels are partitioned depending on their intensity value. Each pixel compares with fixed constant value that is called threshold. The intensity values and the threshold values are compared to detect which value is higher, if the intensity value is greater than threshold value then each pixel is replaced with white pixel and it is considered to be 'foreground', it is black pixel if the pixel value less than that constant value and it is considered as 'background' [Pavel Kral P. et. al. (2016)]. Thresholding has been used to eliminate the unnecessary details and remove the small objects.

▪ **Morphological Operation.** Erosion and dilation have been used as morphological operation.

• **Erosion.** Eliminating the small details and to highlight the holes and gaps of different regions erosion has been used.

• **Dilation.** Dilation has been used to smooth the object boundary, close the holes and gaps and also it expands the size of object.

▪ **Edge Detection.** The 'sobel' edge detection method has been used find the edges. Two types of threshold values have been used to detect the strong and weak edges. A low edge sensitivity has been identified by high threshold value and low threshold value is used for high edge sensitivity.

▪ **K-Means.** K-Means is one of the most popular cluster analysis methods which uses partitioning techniques. This clustering technique classifies or group different objects into k number of groups depending on attributes and features.

In this clustering algorithm each k cluster k centre must select randomly, and K value is fixed in advance. In a cluster, pixels having minimum distance attributes are grouped together. Euclidean distance calculates the distance between each data object and cluster centroid. Depending on the pixel of a cluster, a new centre has to calculate for each cluster. Pixels movement from one cluster to another cluster depending on the change of centroid in every step. This process will continue until no pixels will move from one cluster to another.

▪ **Feature Extraction.** Feature extraction is the most important decision-making process for detection of abnormalities of mammogram. Extracted features are very useful to differentiate the masses and normal breast tissue. First order and second order statistical based features have been extracted for benign and malignant tumor of mammogram images. The statistical features are defined by the following equations shown in Tables 1 & 2.

Table 1. First Order features.

Moment	Definition	Formulae
Mean	The mean is the average value of all pixels in an image.	$\mu = \sum_{i=1}^{G-1} ip(i)$
Standard Deviation	It is the measurement of the average contrast.	$\sigma = \sqrt{\sum_{i=0}^{G-1} (1 - \mu)^2(p(i))}$
Variance	It determines the intensity variation around the mean.	$\sigma^2 = \sum_{i=1}^{G-1} (1 - \mu)^2(p(i))$
Kurtosis	Kurtosis is the fatness of the histogram.	$\sigma^{-4} \sum_{i=1}^{G-1} (1 - \mu)^4(p(i))$

Table 2. GLCM (Gray Level Co-Occurrence Matrix) features.

Moment	Definition	Formulae
Contrast	It measures the sudden change of intensity value in image.	$\sum_{i,j} i - j ^2 p(i, j)$
Correlation	It measures how the reference pixel correlates with its neighbor over an image.	$\sum_{i,j} \left(\frac{(i - \mu_i)(j - \mu_j)p(i, j)}{\sigma_i \sigma_j} \right)$
Energy	Angular second moment is also known as Energy which represents the summation of squared elements in the GLCM.	$\sum_{i,j} p_{i,j}^2$
Homogeneity	Homogeneity describes the distribution closeness of the element in the GLCM-to-GLCM diagonal.	$\sum_{i,j} \frac{p(i, j)}{i + i - j }$

Classification. Image classification plays a key role in various application domain including biomedical imaging, biometry, robot navigation, remote sensing etc. Classification of benign and malignant from mammogram image is one of the most challenging tasks. In our proposed method, two classification method GLCM and Principal component Analysis (PCA) has been used for classification purpose and their performance has been analyzed.

RESULTS AND DISCUSSION

The different stages of images of proposed module are as follows:

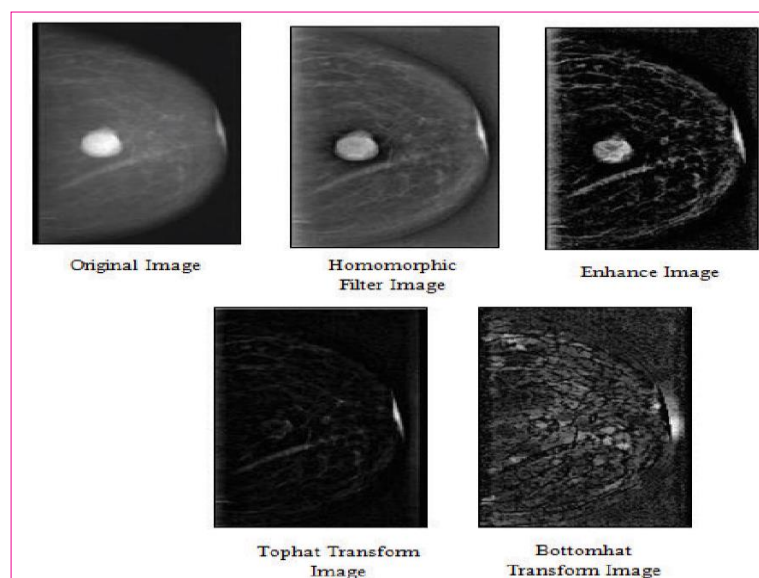


Figure 2. Preprocess images of benign mass.

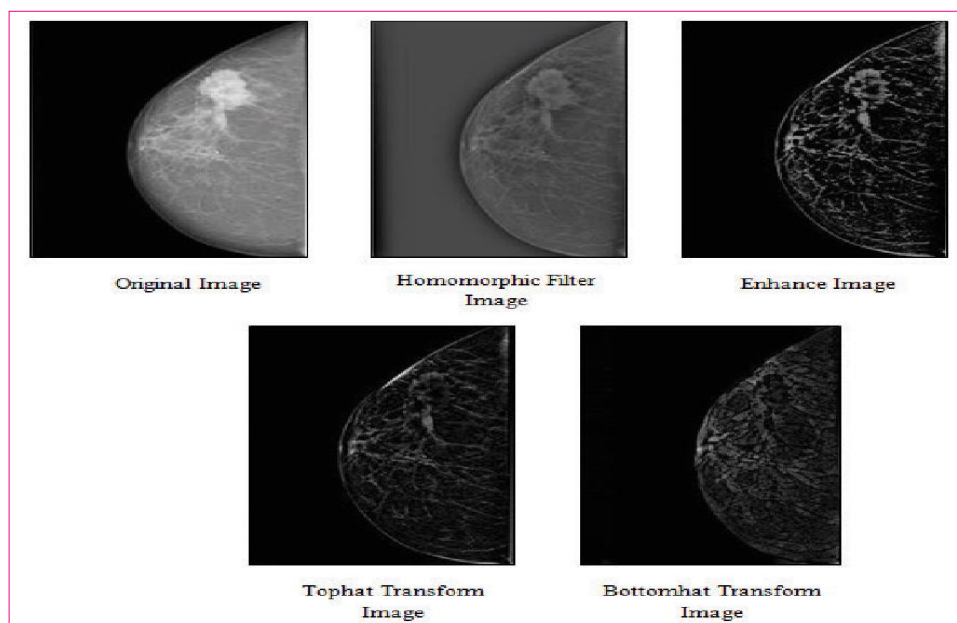


Figure 3. Preprocess images of malignant mass.

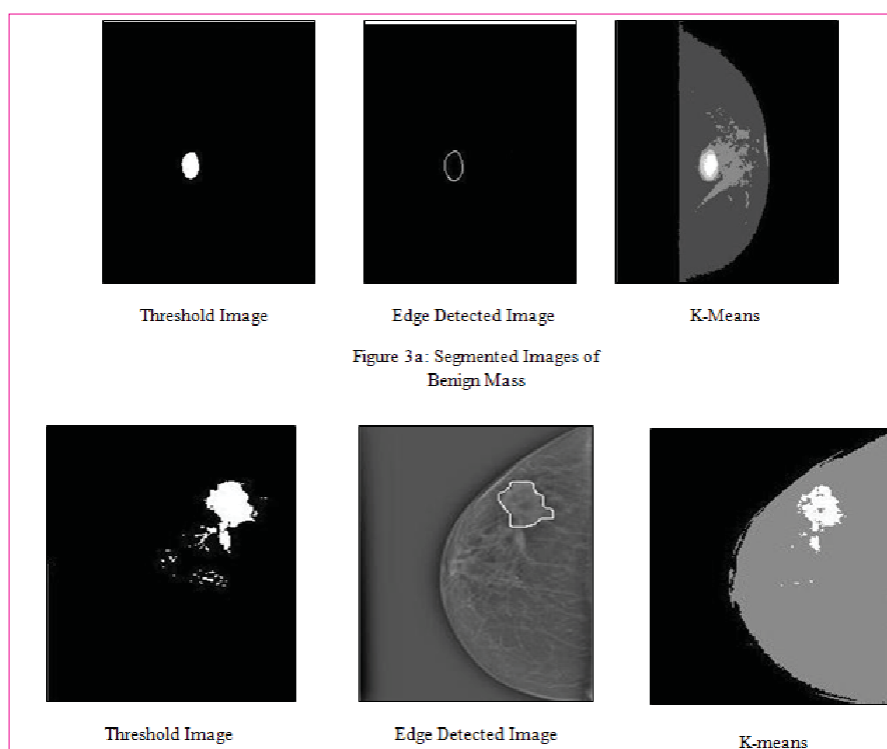


Figure 4. Segmented images of malignant mass.

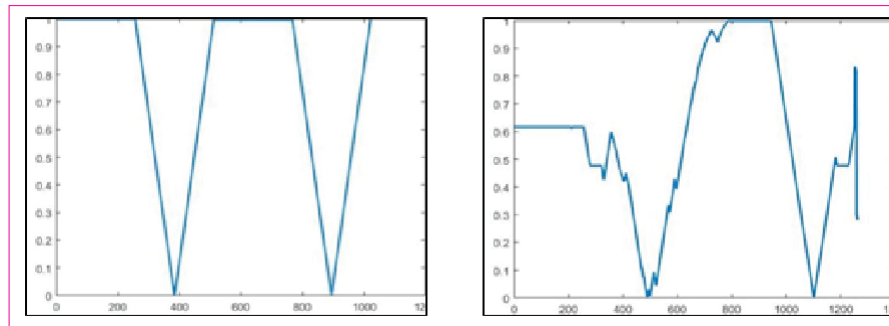


Figure 5. Boundary signature for benign and malignant.

Tables 3 and 4 represent the two different types of statistical based features first order and GLCM features (second order) for detection of abnormalities from mammogram. First order texture features such as, standard deviation, mean, variance, and kurtosis have been calculated from the image. Variance depends on the mean value and the mean value is low if the mass is benign, mean shows the brightness part in the image. Kurtosis shows a lower value for the malignant part.

Standard deviation shows the lower value for malignant tumor and higher value for benign tumor.

Table 3. Statistical features.

Types	Standard Deviation	Mean	Variance	Kutosis
Benign	0.2592	0.3290	0.0672	21.56
Malignant	0.5973	0.4658	0.8660	10.01

Table 4 represents the GLCM features. GLCM texture features like autocorrelation, energy, correlation, contrast and homogeneity are calculated from the image. In our proposed method the distance $d=1$ has been used for each angle: 0° , 45° , 90° , 135° and calculated for both benign and malignant mass. This table represents the variation extracted features with respect to different neighborhood degrees.

Table 4. GLCM features.

Types	Angle	Autocorrelation	Energy	Correlation	Contrast	Homogeneity
Benign	0°	445748	0.0924	0.9432	0.5345	0.8344
	45°	442799	0.0928	0.9456	0.5362	0.8356
	90°	446291	0.1086	0.9485	0.5375	0.8378
	135°	442932	1.4716	0.9498	0.5382	0.8388
Malignant	0°	485644	0.1051	1.9412	0.5625	0.7695
	45°	482896	0.1092	1.9522	0.5639	0.7786
	90°	485292	0.1264	1.9536	0.5655	0.7798
	135°	485496	0.1469	1.9556	0.5696	0.7878

Signature. A distance has been calculated as a function of angle for benign and malignant mass from the centroid to boundary. Figure 6 shows the signature of boundary for benign and malignant mass.

Table 5 shows the detection rate of our proposed method (First Order statistics & GLCM) and PCA. Table 6 evaluates the performance of the different degree of GLCM classification method with neighborhood distance 1 with PCA.

Table 5. Comparison of First Order statistics & GLCM, PCA.

Methods	Detection Rate (%)
Proposed First Order Statistics & GLCM	98
PCA	92

Table 6. Detail performance evaluation of GLCM method with respect to different angles with PCA.

Degrees,°	Recognition Rate, %
0	79.52
45	42.26
90	78.36
135	71.75
PCA	94.86

The accuracy rate of GLCM method is significantly high compared to PCM. The computational time of GLCM is less that provides a better performance by enhancing the speed and effectiveness of algorithm. PCA method is usually popular for face recognition purpose. In PCA image dimension can be reduced due to change of gray level and that is the main pitfall of PCA. Regardless, PCA is also very efficient method, in our test out of 70 images it unable to detect only 4 images. But its computation time is very high and test speed is not also satisfactory.

Figure 5 represents the time complexity of our proposed method and PCA and Figure 6 show the accuracy of the two classification methods. Time complexity of the two classification methods is shown in figure 7.

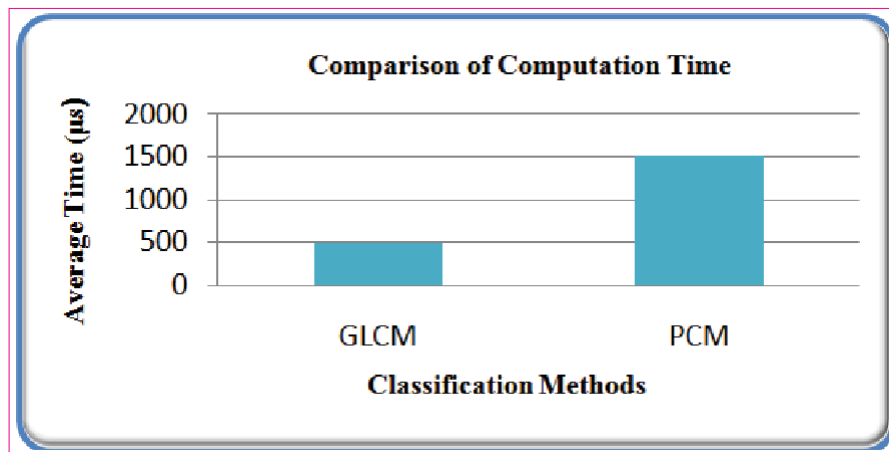


Figure 6. Time complexity of the two classification methods.

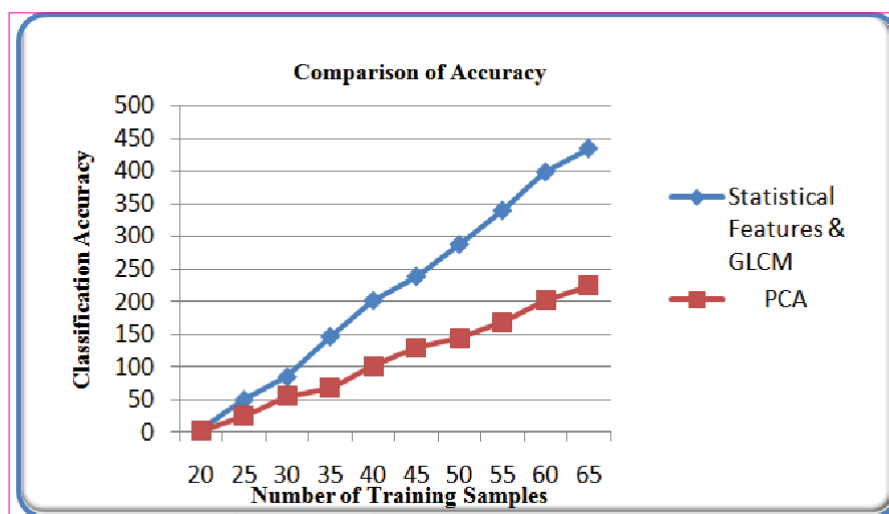


Figure 7. Time complexity of the two classification methods.

CONCLUSION AND FUTURE DIRECTION

This paper enabled to develop an algorithm for diagnosis of breast cancer which exhibits a promising performance. The aim of this paper was to enhance the accuracy of this proposed system for breast cancer detection and classification. Statistical and GLCM both feature extraction methods have been used to increase the efficiency and accuracy of the proposed breast cancer detection system. In GLCM, four different angles with distance =1 have been used. Two parameters:

(a) accuracy and (b) time complexity have been used to evaluate the performance of the proposed algorithm. In our proposed method, recognition rate is very high compared to PCA. Though PCA is an efficient method but its computation time is more than GLCM.

LITERATURE CITED

- [1] Abbasa, Q., M.E. Celebic and I.F. Garce. 2013. Breast mass segmentation using region-based and edge-based methods in a 4-stage multiscale system. *Biomedical Signal Processing and Control* 8: 204-214.
- [2] Abdallah, Y. 2011. Application of Analysis Approach in Noise Estimation, Using Image Processing Program. Lambert Publishing Press GmbH & Co. KG 2011, pp. 123-125.
- [3] Abdallah, Y. and R. Yousef. 2015. Augmentation of X-rays images using pixel intensity values adjustments. *International Journal of Science and Research Vol.4*, Pp.2425-2430.
- [4] Alamr, S.S., NV Kalyankar, SD Khamitkar, Image segmentation by using threshold technique, *Computer Science*, Vol.2, No.5, 2010, Pp.83-86.
- [5] Balakumaran, T., I. L. A. Vennila, and C.G. Shankar, Detection of microcalcification in mammograms using wavelet transform and fuzzy shell clustering, *International Journal of Computer Science and Information Technology*, vol. 7, no. 1, pp. 121–125, 2010.
- [6] Choi, J.H., Eun-Surk Yi, Ji-Youn Kim and Byung Mun Lee, A Novel Approach to Perform Analysis and Prediction on Breast Cancer Dataset using R, *International Journal of Grid and Distributed Computing*, SERSC Australia, vol.11, no.2, February (2018), pp. 41-54.
- [7] de Oliveira Martins, L., G. Braz, Jr., A.C. Silva, A.C. de Paiva and M. Gattass. 2009. Detection of masses in digital mammogram using K-means and support vector machine. *Electronic Letters on Computer Vision and Image Analysis Vol.8*, No.2, Pp.39-50.
- [8] Duffy, S.W., L. Tabar, and R. A. Smith, The mammographic screening trials: commentary on the recent work by Olsen and Gotzsche, *CA Cancer Journal of Clinic*, vol. 52, no. 2, pp. 68–71, 2002.
- [9] Feng, C., S. Zhang, D. Zhao and C. Li. 2016. Simultaneous extraction of endocardial and epicardial contours of the left ventricle by distance regularized level sets. *Medical Physics Vol.43*, Pp. 2741.
- [10] Guzman-Cabrera, R., J.R. Guzman-Supulveda, M. Torres-Cisneros, D.A. May-Arrijoa,
- [11] J. Ruiz-Pinales, O.G. Ibarra-Manzano, G. Avina-Cervantes, A. Donzalez Parada, Digital image processing technique for breast cancer detection, *International Journal of Thermophilic*, 2013, Vol.34, Pp.1519-1531.
- [12] Hao, X., Y. Shen and S.-r. Xia. 2012. Automatic mass segmentation on mammograms combining random walks and active contour, *Journal of Zhejiang University-Science Computers, (Computer & Electron) Vol.13*, No.9, Pp. 635-648.
- [13] Jain, M., S.K. Singh and K. Saxena. 2017. An efficient indexing for content-based image retrieval based on number of clusters using clustering technique. *International Journal of Artificial Intelligence and Applications for Smart Devices Vol.5*, No.1, Pp.1-10.
- [14] Lee, M.-C., Jui-Fang Chang and Jung-Fang Chen, Fuzzy Preference Relations in Group Decision Making Problems Based on Ordered Weighted Averaging Operators, *International Journal of Artificial Intelligence and Applications for Smart Devices*. Vol. 2. No. 1. May. 2014, GV Press, Pp: 11-22.
- [15] Lee, S.-R. 2017. Performance Measurement of Educational Management Information System to Prevent Colorectal Cancer, *International Journal of Big Data Security Intelligence*. Vol. 4. No. 2. Dec. 2017, pp:1-6.
- [16] Li, S, Yanping L, Kaitao Y, Shaozi L. ECG analysis using multiple instances learning for
- [17] Myocardial infarction detection. *IEEE Transactions on Biomedical Engineering* 2012.
- [18] Liu, J., X. Zhuang, L. Wu, D. An, J. Xu, T. Peters and L. Gu. 2017. Myocardium segmentation from de MRI using multicomponent Gaussian mixture model and coupled level set. *IEEE Transactions of Biomedical Engineering Vol.1*, Pp.1-10.
- [19] Liu, Y., C.Li, S. Guo, Y. Song and Y. Zhao. 2014. A novel level set method for segmentation of left and right ventricles from cardiac MR images. *Conference Proceedings IEEE Engineering Medicine Biology Society Vol.47*, Pp.19-22.
- [20] Lu, X, Yang R, Xie Q, Ou S, Zha Y, Wang D. Nonrigid registration with corresponding points constraint for automatic segmentation of cardiac DSCT images. *Biomed Eng Online* 2017; 16: 39.
- [21] Na, S., Liu Xumin, Guan Yong, Research on k-means Clustering Algorithm, *Third International Symposium on Intelligent Information Technology and Security Informatics*, IEEE 2010, Pp. 63-67.
- [22] Naresh, S. and S. Vani Kumari. 2015. Breast cancer detection using local binary patterns, *International Journal of Computer Applications Vol.123*, No.16, Pp.6-9.
- [23] Oliver, A., Xavier Llado, False Positive Reduction in Mammographic Mass Detection Using Local Binary Patterns, *MICCAI2007, Part I, LNCS 4791*, 2007, pp.286-293.
- [24] Pavel Kral, P. and L. Lenc. 2016. LBP features for breast cancer detection, *ICIP2016*, Pp. 2643-2647.
- [25] Penaflor-Espinosa, M.J.B. 2017. The type, extent of use, and perceived effects of complementary and alternative medicine among breast cancer survivors. *International Journal of Advanced Science and Technology Vol.108*, Pp.57-70.
- [26] Ray, A., Indra Kanta Maitra and Debnath Bhattacharyya, Detection of Cervical Cancer at an Early Stage Using Hybrid Segmentation Techniques from PAP Smear Images, *International Journal of Advanced Science and Technology*, SERSC Australia, vol. 112, March (2018), Pp. 23-32.
- [27] Ribeiro, P.B., Roseli A.F. Romero, Patricia R. Oliveira, Homero Schiabel, Luciana B. Verçosa, Automatic segmentation of breast masses using enhanced ICA mixture model, *Neurocomputing Vol.120*, 2013, Pp.61–71, Elsevier.
- [28] Sahamijoo, A., Farzin Piltan, Sareh Mohammadi Jaberri and Nasri b Sulaiman, Prevent the Risk of Lung Cancer Progression Based on Fuel Ratio Optimization, *International Journal of u - and e - Service, Science and Technology*, SERSC Australia, Vol. 8, No.2, February (2015), Pp. 45-60.
- [29] Sakthi, A. and M. Rajaram. 2016. Density based multiclass support vector machine using iot driven service-oriented architecture for predicting cervical cancer. *International Journal of u - and e - Service, Science and Technology Vol.9*, No.11, Pp.195-216.
- [30] Sharma, P. 2017. Prediction of heart disease using 2-Tier SVM data mining algorithm. *International Journal of Advanced Research in Big Data Management System Vol.1*, No.1, Pp.11-24.
- [31] Sharma, P. and A K Mishra, Analysis and Classification of Plant microRNAs using Machine Learning Approach, *Asia-Pacific Journal of Neural*

- Networks and Its Applications. Vol. 1.No. 2. Sep. 2017, GVPress. Pp:1-6.
- [32] Singh, A.K. and B. Gupta. 2015. A novel approach for breast cancer detection and segmentation in a mammogram, Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015), Procedia Computer Science Vol.54, Pp.676-682.
- [33] Srinivasa Rao, D. and V.Sucharita. 2016. Analysis of different utility mining methodologies in transactional databases. International Journal of Big Data Security Intelligence Vol.3, No.1, Pp.1-8.
- [34] Strange, H., Erika Denton, Minnie Kibiro, and Reyer Zwiggelaar, Manifold Learning for Density Segmentation in High-Risk Mammograms, IbPRIA 2013, Pp. 245–252, 2013, Springer-Verlag Berlin Heidelberg 2013.
- [35] Sun, L., L. Li, W. Xu, W. Liu, J. Zhang, and G. Shao. 2010. A novel classification scheme for breast masses based on multi-view information fusion, 4th International Conference on Bioinformatics and Biomedical Engineering (ICBBE), 2015, Pp. 1–4.
- [36] Tabar, L. B.Vitak, H. H. Chen, M. F. Yen, S. W. Duffy, and R. A. Smith, Beyond randomized controlled trials: organized mammographic screening substantially reduces breast carcinoma mortality, Cancer, Vol. 91, No. 9, Pp. 1724–31, 2001.
- [37] Torres Pereira, E., S.Pimentel Eleuterio and J. Marques de Carvalho. 2014, Local binary patterns applied to breast cancer classification in mammographies. RITA Volume 21, No.2, Pp. 1-15.
- [38] Tosteson, A.N., D.G. Fryback, C. S. Hammond, L. G. Hanna, M. R. Grove, M. Brown,
- [39] Q. Wang, K. Lindfors, and E. D. Pisano, Consequences of false-positive screening mammograms, JAMA internal medicine, Vol. 174, No. 6, Pp. 954–961, 2014.
- [40] Yu, J., Rui, Yong; Tang, Yuan Yan; Tao, Dacheng, 2014. High-order distance-based multiview stochastic learning in image classification. IEEE Transactions on Cybernetics Vol.44, No.12, Pp.2431-2442.