International Journal of
Advances in Scientific Research and Engineering (ijasre)

E-ISSN : 2454-8006

DOI: 10.31695/IJASRE.2020.33753

Volume 6, Issue 3
March - 2020

Analysis on Results Comparison of Feature Extraction Methods for Breast Cancer Classification

Than Than Htay[1], Su Su Maung[2] and Khine Thin Zar[3]

Research Scholar [1] and Professor[2-3]

[1]Department of Computer Engineering and Information Technology

[2-3]Department of Computer Engineering and Information Technology

Yangon, Myanmar

Union of Myanmar

ABSTRACT

Feature extraction plays a vital role in image processing techniques for medical imaging. In this paper, the researchers proposed a breast cancer classification system by using image processing techniques to help the radiologists that the system can improve the mammogram screening process and increase the life of cancer patients. Our breast cancer classification system based on the combination of first-order Statistics features and second-order Gray Level Co-occurrence Matrix (GLCM) features and Support Vector Machine is used as a classifier. This system is composed of five stages. At first, preprocessing is carried out for removing noise and detail artifacts from an image, reducing the size of the image by cropping, enhancing the image to show clearly the appearance of the image. Median Filters are used to remove the artifact, noise, high-frequency components and unwanted parts in the background of the mammogram images. Secondly, Otsu segmentation is used to extract the breast region from the background image. In a third stage, enhancement are applied on segmented result images to get efficient features for a higher classification accuracy rate. First-order statistics and second-order texture GLCM features are extracted form enhanced image. Support Vector Machine is used as a classifier for the classification of abnormal and normal images. Finally, performance comparison of the first order, second order features and combination of first-order statistics and second-order GLCM features for breast cancer detection system are done with classification accuracy scores. In this system, an input MIAS database are used for our breast cancer detection system.

Key words: Breast Cancer, Enhancement, Feature Extraction, GLCM, MIAS Database.

1. INTRODUCTION

Screening and diagnosis of breast cancer in Mammogram is the common techniques for the radiologists. Mammography is processing of image using low- dose X-rays to analyze the human breast. Searching Mammogram images for sign of abnormalities of the breast cancer can cause the radiologists experts many miss judgments. A computer aided detection (CAD) become a helpful system in detection and diagnosing breast abnormalities than typical screening programs [1]. So, there is a great interest in developing methods of detecting abnormalities of breast mammogram image to help the experts and improve the screening process. This breast cancer detection system mainly focus on feature extraction for the extraction of the relevant information that characterizes each class. In this process, relevant features such as first order statistics features and second order GLCM features are extracted from objects. The classifier used feature result to classify the image. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate and classification accuracy rate with the least amount of element. MIAS database is used as an input which are the images of the database originate from a film-screen mammographic imaging process in the United Kingdom National Breast Screening Program [2]. There are five stages in that system: Image Database, preprocessing, segmentation, feature extraction and classification. The remainder part of the paper is as follow. The previous works from literature are given in Section II. The configuration of the whole system is described in Section III. Section IV describes GLCM feature extraction detail and the experiments of results are shown in section V. The last Section includes the conclusion of the work.

2. LITERATURE REVIEW

A.Qayyum [3] presented a simple breast cancer detection method for digital mammography images. In that system, they focused on the point pectoral muscle removal. GLCM feature extraction was carried out after removing the pectoral muscle and finally, Support Vector Machine (SVM) was used as classifier for classification to normal and abnormal tissues of breast images.

A system based on segmentation and classification techniques for breast tumor was developed by Dinsha[4] . In this method, preprocessing work is carried out by using Contrast Limited Adaptive Histogram Equalization (CLAHE) technique. Combination of K-means and fuzzy c-means algorithms is used for segmentation stage. Many features are extracted based on the segmented images. Finally, classification process have been done by using the SVM and Bayesian classifiers.

Biswas[5] proposed a Computer Aided Diagnosis (CAD) system that can classify the normal and abnormal breast tissues. In this system, 2D median filter. And ROI extraction is applied for removal for noise and label artifacts . The appearance of the image enhanced by using CLAHE and second order GLCM features are extracted. KNN, SVM and ANN are classifiers for in that system.

N. Kharel [6] aimed to enhance using hybrid solution for early diagnosis of breast cancer using mammogram images. In their system, hybrid image enhancement method using CLAHE and Morphology method helps to enhance image for further computation in CAD system to early diagnosis of breast cancer using mammogram images.

Youssef [7] presented a mammograms classification method that are focused on global statistical features. In this paper, image sample are acquired with Terahertz imaging. Their work is organized as follows: preprocessing stage, feature extraction and classification processes. Statistical features such as mean, variance, Range, standard deviation and entropy were extracted. Although the mammograms can be categorized into three classes: normal, benign, and malignant, they not explained which classifier they used.

Deepa [8] introduced an adaptive approach for Computer aided Screening for abnormalities of mammograms. Their work focused on CLAHE method to enhance the contrast of intensity of the images. In segmentation stage, they used H-domes transformation firstly and applied thresholding for mapping the original gray image and new image to free from background intensity and finally proper Dilation is applied for eliminating noise. Support Vector Machine and Naïve Baye's classifiers are used and compared their results. But their result can't give the better results.

Omprakash Patel [9] introduced a paper that provides review of different popular histogram equalization techniques and experimental study based on the absolute mean brightness error (AMBE), peak signal to noise ratio (PSNR), Structure similarity index (SSI) and Entropy.

3. SYSTEM OVERVIEW

The system is carried out using MATLAB program and output results are showed with Graphical User Interface. The main steps of the system are image acquisition, preprocessing, segmentation, enhancement, feature extraction and classification. Results comparison of feature extraction are carried out with the classification accuracy rate at the last stages. In our system, MIAS's dataset are used as input images and we showed that double enhancement give the best results for our cancer detection system. This system is designed to help the radiologists for detection the abnormalities of breast cancer image accurately.Fig.2 showed the proposed system design of the early stage breast cancer detection system.

3.1. Image Database

In order to make an experiment and evaluation for the proposed system, MIAS (Mammographic Image Analysis Society) database was used. This database is an organization of research group in UK [2]. They are interested in the understanding of mammograms. This database includes left and right breast image from 61 patients. Totally, there are 322 images with three different types, namely normal, benign and malignant. These mammograms can be distinguished into three tissue types namely fatty, fatty glandular and dense glandular while these verities of abnormality can be classified as benign and malignant tumor.

DOI: 10.31695/IJASRE.2020.33753

3.2. Preprocessing

In Digital image, brightness characteristics of a pixel have an influence of background noise hence image preprocessing become necessary. The main objective of the image processing system is to improve the quality of the image and make these images ready for further processing. This important things can be made by removing the irrelevant noise and unwanted parts in the background of the mammograms. It also invloves noise removal by median filter , removing label and artifacts and cropping.

3.3. Segmentation

Image segmentation is the partition of an image into several components. The goal of segmentation is to make simpler and the representation of an image can change into something that is more meaningful and simpler to analyze [3]. The segmentation stage can give the exact location of the suspicious area of breast image for diagnosing and abnormalities classification into benign or malignant that is cancerous mass. In our system, Otsu thersholding technique is used for segmentation stage.

3.4. Image Enhancement

Image enhancement is a technique that can improve the interpretability or perception of information in images for human viewers, or provide `better' input for other automated image processing techniques. Image enhancement techniques can be distinguished into two types: they are spatial domain methods, which operate directly on pixels and frequency domain methods, which perform on the Fourier transform of an image. Histogram equalization (HE) is contrast enhancement technique in a spatial domain in image processing with the use of histogram of image. Histogram equalization usually increases the global contrast of the processing image [4]. In this research work, CLAHE method are applied on the segmented image to give excellent image quality that can give the best result for next image processing stages in our work. The result images of the segmented image from previous stage and enhanced image of CLAHE are compared with their histogram images.

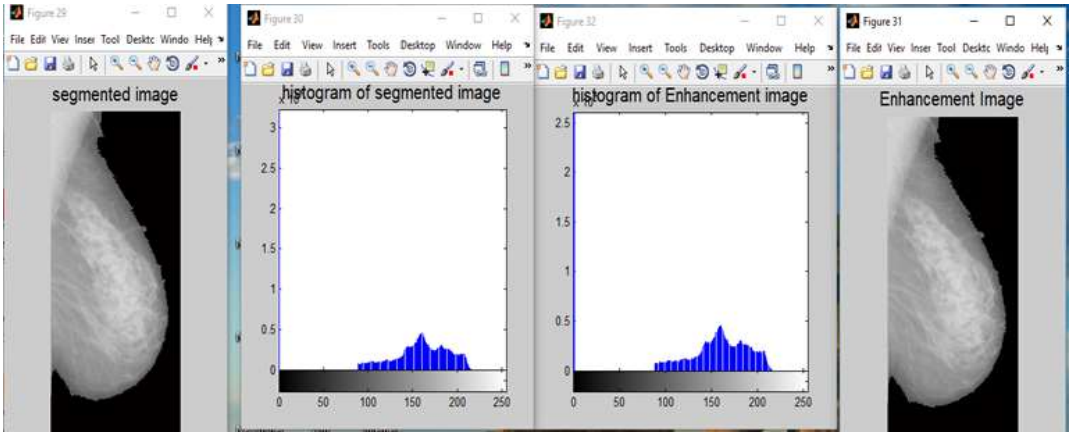

Figure.1 the result of segmented image and enhanced image with their histogram images.

3.5. Feature extraction

In this paper, first order statistics and second order texture GLCM are formulated to extract the statistical texture features. Only three first order features are mean, standard deviation, skewness, kurtosis and five second order namely energy, entropy, homogeneity, correlation and contrast are computed. Energy is the measurement of image smoothness. Contrast measures the local variation's level which gives the lower values for low contrast image and high values for high contrast image. Homogeneity measures the similarity of pixels. Diagonal gray level co-occurrence matrix gives homogeneity of 1. Entropy gives the measure of randomness. For Correlation, this feature measures how a pixel is correlated to its neighborhood. Feature values range from -1 to 1, these extremes indicating perfect negative and positive correlation respectively [10].

3.6. Classification

After getting the features of the images from the feature extraction stage, classification of breast image is performed to classify the breast images into normal and abnormal image. Support Vector Machine is used as classifier for early stage detection system. Features or attributes are values measured from a mass of a mammography image. As we described in the previous section, we have selected 4 statistical features and 5 texture features for the classification of mammography image to normal or abnormal.

DOI: 10.31695/IJASRE.2020.33753

Among the mammography images, two major groups (Normal (1) and Abnormal (-1)) were selected for this research. Hence, we compute the feature values of each mammography images.

In our case, we have two groups and 322 images which is the total number of dataset. In the training process the group values were provided because we use supervised learning method. The feature results getting form training stage are saved in the feature database. In order to test the classification accuracy of the system, testing dataset which are not in the training dataset will be used that are randomly depend on the state of the default random stream.

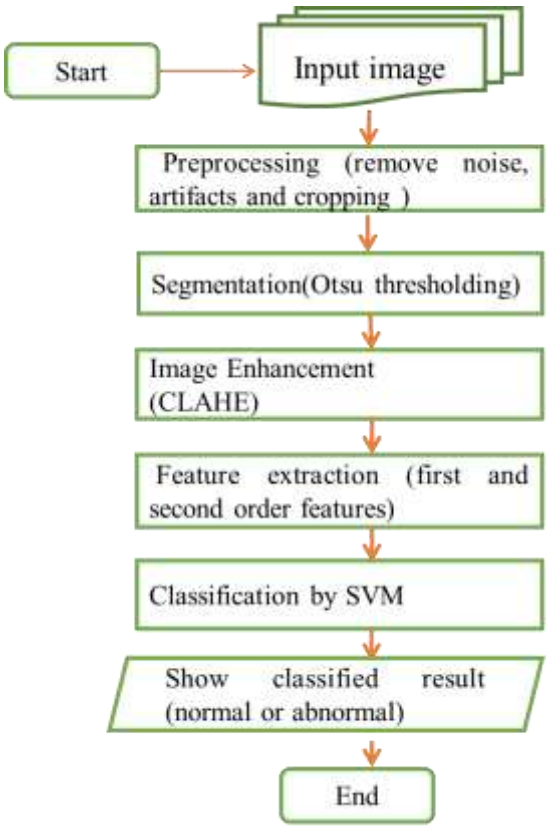

Figure 2. Flow diagram of breast cancer classification system

4. FEATURE EXTRACTION

Feature extraction is a method that can capture visual content of images for indexing and retrieval. Feature extraction gives a piece of information that is necessary for performing the computational task related to a particular application [11]. Two types of texture feature are first order texture feature and second order texture feature. In the first order, features are derived by using the individual pixel and not need to consider the neighbor relationship. For the second order features, GLCM compute the statistical features from the pixel of the neighbor relationship [12].

4.1 First Order Features

The gray levels of image region can be represented with random variable (I). The first-order histogram P(I) is defined as:

$$P(I) = \frac{number\ of\ pixels\ with\ gray\ level\ (I)}{total\ number\ of\ pixels\ in\ the\ region} \quad \text{----------} \quad (1)$$

Based on the definition of P (I), the Mean m_1 and Central Moments μ_k of I are calculated by;

$$m_1 = E(I^1) = \sum_{I=0}^{N_g-1} I^1 P(I) \quad \text{----------} \quad (2)$$

$$\mu_k = E[I - E(I^k)] = \sum_{I=0}^{N_g-1} (I - m_1)^k P(I), \text{ k=2, 3, 4} \quad \text{----------} \quad (3)$$

where N_g is the number of possible gray levels .

DOI: 10.31695/IJASRE.2020.33753

The histogram width can be measured by the variance. Standard Deviation is µ1/2 and can be used in edge sharpening, as intensity level get changes at the edge of image by large value. Skewness can be used to measure of the degree of histogram asymmetry around the Mean. In our system, mean, standard deviation, skewness, Kurtosis are used.

4.2 Second Order Statistics or GLCM

Gray level co-occurrence matrix (GLCM) is one of the most well-known texture analysis methods. It also called second order statistic. The statistical features can be computed by GLCM from intensities of the image [13].

R. M. Haralick, K. Shanmugam and I. Dinstein [14] derived 14 statistical features. In our system, we extracted five of those, energy, homogeneity, correlation, contrast, and entropy from mammogram image. In all the equations, P(i, j) is the normalized value of GLCM. Each entry (i, j) is the number of occurrences of the pair of gray levels i and j which are a distance in original image. N_g is the number of gray levels in the image.

From Co-occurrence matrix, features can be calculated which quantifies coarseness, smoothness and texture related information that have high discriminatory power. Among them [14], Energy, Homogeneity, Correlation, Contrast, and Entropy are few such measures which are given by:

$$Energy = \sum_{i,j=0}^{N_g-1} p(i,j)^2 \text{----------} \quad (4)$$

$$Homogeneity = \sum_{i,j=0}^{N_g-1} \frac{1}{1+(i,-j)^2} P(i,j) \text{ ----------} \quad (5)$$

$$Correlation = \frac{\sum_{i,j=0}^{N_g-1}(i-\mu_i)(j-\mu_j)P(i,j)}{\sigma_i \sigma_j} \text{ ----------} \quad (6)$$

$$Contrast = \sum_{i,j=0}^{N_g-1}(i,j)^2 P(i,j)] \text{ ----------} \quad (7)$$

$$Entropy = \sum_{i,j=0}^{N_g-1} P(i,j)[\ln P(i,j)] \text{----------} \quad (8)$$

In this equation μi and μj are the means and σi and σj the standard deviations.

$$\mu_i = \sum_{i,j=0}^{N_g-1} i\ [\ P(i,j)] \text{ ----------} \quad (9)$$

$$\mu_j = \sum_{i,j=0}^{N_g-1} j\ [\ P(i,j)] \text{ ----------} \quad (10)$$

$$\sigma_i = \sqrt{\sum_{i,j=0}^{N_g-1} P(i,j)(i-\mu_i)^2} \text{ ----------} \quad (11)$$

$$\sigma_j = \sqrt{\sum_{i,j=0}^{N_g-1} P(i,j)(i-\mu_j)^2} \text{ ----------} \quad (12)$$

All these features can provide a high classification rate to distinguish two different kinds of images [15]. Combination of first order and second order statistical features are used for our system.

5. EXPERIMENTAL RESULT

Experiments for proposed system are performed on the Mini-MIAS database, which is freely available. In this database contain 322 breast images at resolution of 1024*1024 pixels. At the initial preprocessing stage, the median filter was applied to mammogram image for making removal of the noise. After removing the unwanted small portion of the binary image, the resulting binary mask contains a breast region only. Then the mask is multiplied element wise with noise suppressed mammogram to extract the breast region in mammogram. Finally, cropping is done to reduce the size of the image. After preprocessing the image, Otsu segmentation stage is performed to extract the breast region from the background image. To improve the image's quality, we performed the image enhancement again on segmented images by using Contrast Limited Adaptive Histogram

DOI: 10.31695/IJASRE.2020.33753

Equalization method. At the feature extraction stage, we extract the first order features and GLCM features, totally 9 features from enhancement results.

After the random selection of training and testing dataset, we need to train the classifier with 80% training sample content that represents members of all of the groups. It is very important to find good training samples because the quality of the training sample has direct impact on the quality of classification. The second parameter to train the classifier is a training label specification, which is a sequence of training and testing label elements. After training the model, the performance of the classifier is measured using 20% of test dataset for accuracy, sensitivity, specificity based on the following performance measuring parameters.

Table 1. Confusion matrix that are used for the result of classification

Actuals group		
	abnormal	normal
abnormal	TP	FN
normal	FP	TN

$$\text{Accuracy} = \frac{TP+TN}{N} \times 100\% \quad \text{----------} \quad (13)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \times 100\% \quad \text{----------} \quad (14)$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \times 100\% \quad \text{----------} \quad (15)$$

- Accuracy - overall correctness of classifier
- Specificity – true negative rate
- Sensitivity –true positive rate
- Higher value the of both Specificity and Sensitivity, better the performance of the system

Where: - FN = False Negative,
FP = False Positive
TN = True Negative,
TP = True Positive

Finally we compared the feature results by using Support Vector Machine as a classifier and compare their classification results. In our system, combination of enhanced first order and second order GLCM feature give the better result than the original first order features and second order features according to the confusion matrix that are shown in Table 1.

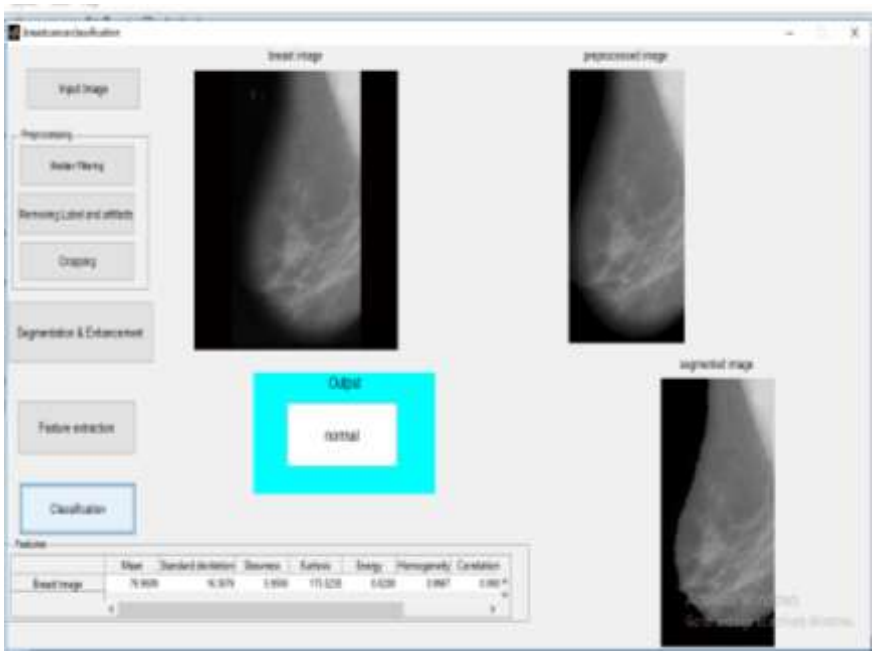

Figure 3. Output result with GUI for the breast cancer classification system

DOI: 10.31695/IJASRE.2020.33753

Table 2. Confusion matrix that are used for the result of proposed system

Actuals group		
	abnormal	normal
abnormal	26	3
normal	2	33

Table 3. Confusion matrix that are used for the result of first order features based system

Actuals group		
	abnormal	normal
abnormal	23	6
normal	15	20

Table 4. Confusion matrix that are used for the second order features based system

Actuals group		
	abnormal	normal
abnormal	17	12
normal	6	29

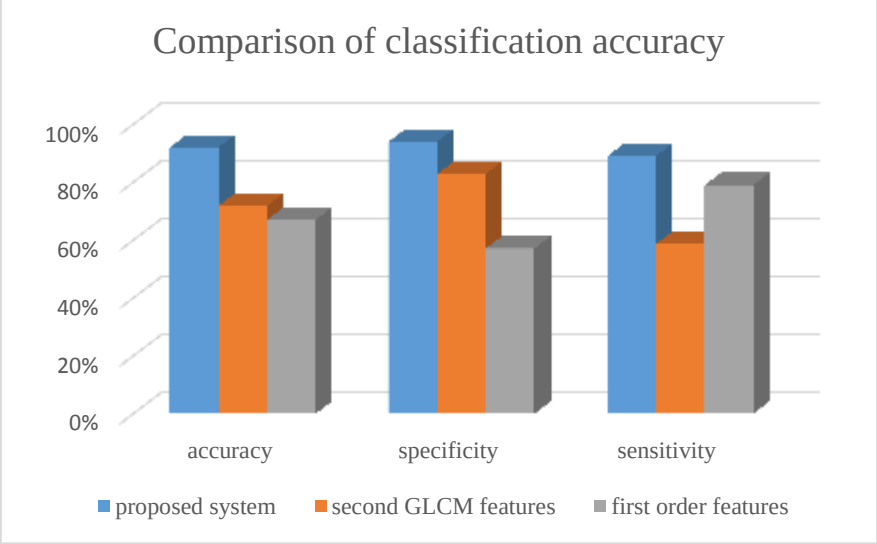

Figure 4. The chart showing the comparison of classification accuracy result of feature extraction methods

6. DISCUSSION AND CONCLUSION

Preprocessing stage helps to removes any unwanted noise in the image and to improve the quality of images. Enhancement after segmenting the breast ROI can improve the result for instructing effectively the feature extraction. In figure.3, combination of first order and second order GLCM features give the better result than each by comparing with other research works. Classification accuracy rate is higher for enhanced features than the original features. Enhancement can solve the challenging problem of detection in abnormalities of low contrast mammogram image. In our system, although adding one stage such as enhancement after segmentation stage on original breast cancer detection system, this system longs a few seconds than original system. This system will be classified breast cancer and improve the accuracy rate than other system so that the breast cancer detection and classification can be a useful application in medical imaging field. As a future work, we will apply our breast cancer detection on other image dataset and train with other classification methods such as KNN, Naive Bays and will compare their classification accuracy rate.

DOI: 10.31695/IJASRE.2020.33753

REFERENCES

[1] L. Tabar, P. Dean. , "Mammography and breast cancer; the new era," *International Journal of Gynecology and Obsterics*, vol. 28, Issuse 3, September 2003, pp. 319–326.

[2] J Suckling et al. "The Mammographic Image Analysis Society Digital Mammogram Database" Exerpta Medica. International Congress Series 1069, pp375-378 ,1994.

[3] A.Qayyum and A.Basit, "Automatic breast segmentation and cancer detection via SVM in mammograms'',in *Proc.IEEE,*2016.

[4] D. Dinsha, "Breast tumor segmentation and classification using SVM and Bayesian from thermogram images,"*Unique Journal of Engineeringand Advanced Sciences*, vol. 2, 2014, pp: 147-151.

[5] R.Biswas, A.nath and S.Roy, "Mammogram classification using Gray-Level Coccurrence Matrix for diagnosis of breast cancer''in *Proc. International Conference on Micro-Electronics and Telecommunications Engineering*, 2016,pp161-166.

[6] N. Kharel, A. Alsadoon, P.W.C. Prasad, and A.Elchouemi, "Early diagnosis of breast cancer using contrast limited adaptive histogram equalization (CLAHE) and Morphology methods", in *Proc. 8th International Conference on Information and Communication Systems (ICICS)*, IEEE: 120-124,2017.

[7] M.A. Alolfe, A.M. Youssef, Y.M. Kadah, and A.S. Mohamed, 2008. "Computer-aided diagnostic system based on wavelet analysis for microcalcification detection in digital mammograms", Proceedings of Cairo International Biomedical Engineering Conference, IEEE : 1-5.

[8] A.K. Deepa, S. Niyas and M. Sasikumar, "An adaptive approach for computer aided screening of mammograms and classification of abnormalities'', in *Proc. International Conference on Communication and Network Technologies (ICCNT),*2014,pp 169-173.

[9] O. Patel, Y.P. Maravi, and S. Sharma, 2013. "A comparative study of histogram equalization based image enhancement techniques for brightness preservation and contrast enhancement", arXiv preprint arXiv: 1311.4033

[10] K.N.Nyein Hlaing and Anilkumar K.G, "Myanmar paper currency recognition using GLCM and k-NN'', in *Proc.Second Asian Conference on Defence Technology (ACDT),*IEEE,2016

[11] N. Ponraj, Poongodi and M.Mercy, "Texture analysis of mammogram for the detection of breast cancer using LBP and LGP: A Comparison'', in *Proc. IEEE Eighth International Conference on Advanced Computing (ICoAC),*2016,pp-182-185.

[12] G.Kumar, P.K. Bhatia, "A detailed review of feature extraction in image processing system'' in *Proc. International Conference on Advanced Computing & Communation Technologies*,2014, pp.5–12.

[13] V. Kumar, P. Gupta, "Importance of statistical measures in digital image processing," *International Journal of Emerging Technology and Advanced Engineering*, ISBN: 2250-2459, vol. 2(8), August 2012.

[14] R. M. Haralick, K. Shanmugam, & I. Dinstein, "Textural features for image classification,'' in *Proc. IEEE Transactions on Systems, Man and Cybernetics*, vol. 3, pp. 610-621, November 1973.

[15] H.D. Cheng, J. Shan, W. Ju, Y. Guo, L. Zhang., "Automated breast cancer detection and classification using ultrasound images: A survey". *Pattern recognition*, Vol.43, No.1:299-317, 2010.

DOI: 10.31695/IJASRE.2020.33753