

GLCM and GLRLM based Feature Extraction Technique in Mammogram Images

K. Preetha^{1*} and Dr. S.K. Jayanthi²

¹Research Scholar, Dept. of Computer Science, Vellalar College for Women, Erode, India.

²Associate Professor & Head, Dept. of Computer Science, Vellalar College for Women, Erode, India.

*Corresponding author E-mail: kpreetha.msc@gmail.com

Abstract

A mammogram is an x-ray that allows a qualified specialist to examine the breast tissue for any suspicious areas. Mammogram helps for early diagnosis before showing symptoms of cancer. The aim of this paper is to extract the various features of pre-processed mammogram images to improve the performance of the diagnosis, which helps the radiologists in reducing the false positive predictions. Mammogram images are pre-processed using hybrid filter MAX_AVM. Shape, Intensity, Gray Level Co-occurrence Matrix and Gray Level Run-Length Matrix features that help to represent the various classes of objects are extracted and used as inputs to the classifier. The classifier helps to classify the mammogram images into a normal or abnormal pattern. Experiments were conducted on MIAS database. The result shows that the combination of GLCM and GLRLM features are efficient and achieved the maximum classification accuracy rate when compared to other features.

Keywords: Breast Cancer, Feature Extraction, GLCM, GLRLM, Mammogram, Neural Network.

1. Introduction

Nowadays, Breast cancer is one of the most common cancers in women. The healthy cells in the breast grow and change into a tumor. A tumor can be cancerous or benign. In 2017, the American Cancer Society has announced that, over 2, 50,000 new cases of invasive cancer were diagnosed each year in women. According to that data, approximately 40,000 women died of breast cancer in 2015. The American Cancer Society recommended that women with an average risk of breast cancer should undergo regular screening mammography, starting at the age of 45 to 54 years.

Because of early detection, intervention and postoperative treatment, breast cancer mortality has been decreasing. Mammography is the preferred screening examination for breast cancer. During Mammography the breast is exposed to a small dose of ionizing radiation that produces an image of the breast tissue. Mammogram shows tiny clusters of calcium called micro-calcifications. Each pixel of the mammogram corresponds to the class either normal or abnormal. The main objective of this research is to compare and analyze the different feature extraction techniques for the classification of mammogram images. Various image processing steps like Pre-processing, Feature Extraction is carried out along with Neural Network Classification.

This paper is organized as follows: Section II briefly discussed about the existing feature extraction techniques, Section III describes the methodology, Experimental results and performance evaluation are discussed in Section IV and finally conclusion is presented in Section V.

2. Literature Review

Currently, breast cancer detection is a challenging issue for women. Breast cancer is curable if it is detected at an initial stage.

A number of researches have been done on the textural analysis of mammography images by proposing different feature extraction techniques.

Manavalan Radhakrishnan and Thangavel Kuttiannan [7], have proposed a comparative analysis of feature extraction methods for the classification of prostate cancer from TRUS medical images. The extracted features like histogram, GLCM and GLRLM are analyzed separately. The support vector machine is used to classify the extracted features into normal or abnormal. The result shows that a combination of all features performed well in prostate cancer.

R.Nithya and B.Santhi [1], presented an evaluation and comparison of three different feature extraction methods like intensity histogram, GLCM, and intensity based features. These methods were used to classify the normal and abnormal patterns in mammogram images using neural networks. The result proved that GLCM features performed better when compared to other features.

Neeta V. Jog, S .R.Mahadik et al [12], stated that Gray Level Difference Matrix and Gabor feature extraction method are used as input to the classification using SVM and K-NN classifier to classify as normal or abnormal. Experimental results of the research shown that the GLDM feature extractor with SVM classifier achieved 71.83% of accuracy.

The overall literature survey says that different feature extraction techniques are applied to classify the images as normal or abnormal. Based on the study, better and efficient hybrid texture feature extraction technique has been proposed to improve the accuracy of classification using GLCM and GLRLM.

3. Proposed Methodology

The proposed system consists of three phases, namely Preprocessing, Feature extraction and Classification from the

mammogram image. The overview of the proposed methodology is depicted in Fig.1.

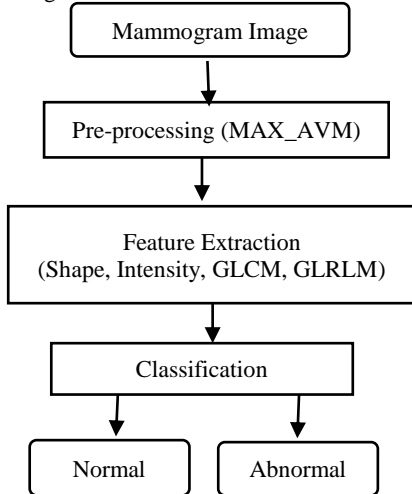


Fig 1: Diagrammatic Representation of Proposed Method

Pre-Processing

The mammogram images may contain noises such as gaussian, speckle, salt and pepper and poisson noise thus it reduces the accuracy of the cancer detection. De-noising is performed to improve the accuracy and to enhance the quality of the image. Preprocessing is done before extracting the features of the image. The image quality is enhanced by using the combination of maximum and AVM (Mean, Median and Variance) [5]. Resultant of the pre-processing is the noise free images which are used for the feature extraction.

Feature Extraction

Feature Extraction plays an important role in the detection and classification of mammogram images. It is the process by which certain features of interest within an image are detected for further processing of the mammogram image. From the preprocessed image the various types of features are extracted by using shape, Intensity and texture features.

Intensity based features are the first order statistics depends only on individual pixel values [1]. It provides the information about the gray level distribution of the mammogram image. The intensities and its variations inside the mammograms can be measured by features like Mean, Standard deviation, Entropy, Energy and Skewness [4]. The Mean is the average value gives the brightness of the image. A bright image will have low mean whereas dark image will have high mean. The mean can be defined as follows:

$$\mu = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N P(i,j)$$

Where $P(i,j)$ is the pixel value at point (i,j) of an image size $M*N$. The standard deviation is also known as square root of the variance. It shows the contrast of the gray level intensities. The low value of the standard deviation indicates the low contrast and the high value shows the high contrast of the image. This can be computed as follows:

$$\sigma = \sqrt{\frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (P(i,j) - \mu)^2}$$

Skewness shows the asymmetry of the intensity distribution. The skewness is calculated by using the formula:

$$S = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N \frac{(P(i,j) - \mu)^3}{\sigma}$$

Entropy is the statistical measure of randomness that can be used to characterize the texture of the input image. Entropy of an image is calculated as follows:

$$h = - \sum_{k=0}^{L-1} Pr_k (\log_2 Pr_k)$$

Where Pr is the probability of the K -th gray level, L is the total number of gray level.

Kurtosis is a measure the peak of the distribution of the intensity values around the mean. It can be defined as follows:

$$K = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N \left[\frac{P(i,j) - \mu}{\sigma} \right]^4 - 3$$

Energy measures the uniformity of the intensity level distribution. Energy can be calculated as follows:

$$E = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N P(i,j)^2$$

Shape feature provide the information about the shape and size of the tumor in the mammogram image. Different types of shape features like area, origin, major axis length, minor axis length, eccentricity, orientation, filled area, extreme, solidity and equiv-diameter are used to measure the properties from an enhanced image region. Area is the measure of actual number of pixels in the tumor region. Eccentricity is the measure of aspect ratio. Its ratio of length of major axis to minor axis. Values range between 0 (in case of circle) and 1 (in case of line segment). Orientation is the angle in degrees between the x-axis and major axis in an image. Extent measures the ratio of pixels in the region to pixels in the total bounding box. Solidity is the proportion of pixels present in the convex area and area, where convex area is counting the number of pixels in the region. Equiv-diameter is the diameter of the circle with the same area as the shape. Filled area is the number of pixels in the region with all holes filled in.

Texture feature is used to identify the cancer region in the mammogram image.

The texture is a combination of repeated patterns with a regular frequency. Texture models are generated using Gray-level co-occurrence matrix and Gray- level run length matrix.

Gray-level co-occurrence matrix (GLCM) textures as originally described by Haralick and others in 1973 [11].

It is the second order statistical method of examining the textures of an image that considers the spatial relationship between two pixels.

The GLCM functions characterize the textures of an image by calculating how often a pair of the pixel with gray-level or value i occur either horizontally, vertically, or diagonally to adjacent pixels with the value j (i and j represent the grey level values in the image) [7]. After creating the GLCM, several texture features are derived from the images like contrast, correlation, homogeneity and energy are calculated on the co-occurrence matrix are shown in Table 1.

Contrast returns a measure of the intensity contrast between a pixel and its neighbor over an image. Homogeneity is the measure of closeness of the distribution of elements in the GLCM. Correlation is the measure of how correlated a pixel is to its neighbor in an image.

Energy is the sum of squared elements in the GLCM. It is also known as uniformity. Figure 2 shows the generation of co-occurrence matrix for 4 gray level in the image at the direction of $\theta = 0^\circ$ and 45° .

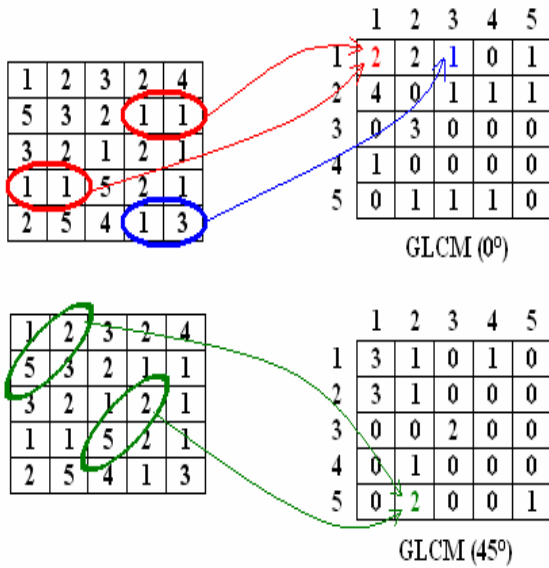


Fig 2: Generation of Gray Level Co-occurrence Matrix

Table 1: GLCM Features and Formula

FEATURES	Formula
Contrast	$\sum_{i,j=0}^{N-1} P_{ij}(i-j)^2$
Homogeneity	$\sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+(i-j)^2}$
Energy	$\sum_{i,j=0}^{N-1} (P_{ij})^2$
Correlation	$\sum_{i,j=0}^{N-1} P_{ij} \frac{(i-\mu)(j-\mu)}{\sigma^2}$ <p>Where, $\mu = \sum_{i,j=0}^{N-1} iP_{ij}$</p> <p>and $\sigma^2 = \sum_{i,j=0}^{N-1} P_{ij} (i-\mu)^2$</p>

Gray level run length method is a way of extracting higher order statistical features in the mammogram images. GLRLM is the set of continuous pixels having same gray level. The run length is the number of neighboring gray levels in particular direction. It is calculated by counting the number of times the corresponding run occurs in the image. In the gray level run length matrix $P(i, j|\theta)$, the $(i, j)^{th}$ element describes the number of runs with gray level i and length j occur in the image along angle θ [7]. The gray level run length can be defined as,

$$\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} P(i, j|\theta) \text{ and } 1 \leq N_z(\theta) \leq N_p$$

Where, N_g be the number of intensity values in the image

N_r be the number of run lengths in the image

N_p be the number of pixels in the image

$N_z(\theta)$ be the number of runs in the image along angle θ

$P(i, j|\theta)$ be the run length matrix for an arbitrary direction θ

Fig.3 Shows the two dimensional matrix of 4*4 with 4 discrete gray levels in the direction of $\theta = 0^\circ$ and 45° .

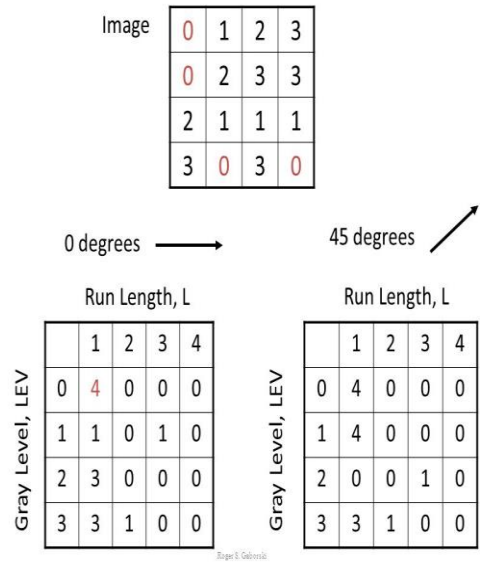


Fig3: Gray Level Run Length Matrix

The features Short Run Emphasis (SRE), Long Run Emphasis (LRE), Gray Level Non-uniformity (GLN), Run length non-uniformity (RLN), Run percentage (RP), Low Gray level Run Emphasis (LGRE), High Gray Level run Emphasis (HGRE) are used to find the accuracy of the classification. The formula is tabulated in Table II.

Table 2: GLRLM Features and Formula

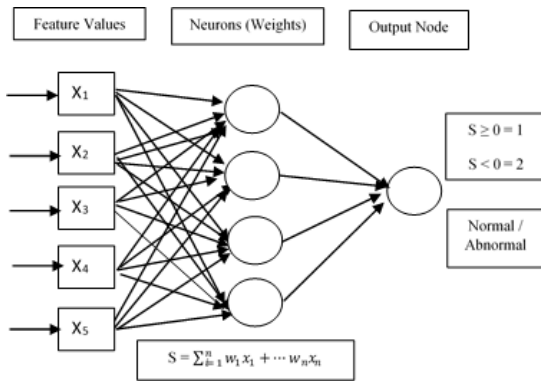
FEATURES	Formula
SRE	$\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{P(i, j \theta)}{j^2}}{N_z(\theta)}$
LRE	$\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} P(i, j \theta)j^2}{N_z(\theta)}$
GLN	$\frac{\sum_{i=1}^{N_g} (\sum_{j=1}^{N_r} P(i, j \theta))^2}{N_z(\theta)}$
RLN	$\frac{\sum_{j=1}^{N_r} (\sum_{i=1}^{N_g} P(i, j \theta))^2}{N_z(\theta)}$
RP	$\frac{N_z(\theta)}{N_p}$
LGRE	$\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{P(i, j \theta)}{i^2}}{N_z(\theta)}$
HGRE	$\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} P(i, j \theta)i^2}{N_z(\theta)}$

Once the features are extracted using shape, intensity, GLCM, GLRLM and their combinations, the feature values are given as an input to the classification.

Classification

After the features have been extracted they are input into a classifier to categorize the images into normal or abnormal classes. In classification, neural network weight optimization has been used to classify the breast cancer. The neural network is trained by adjusting the weights, to predict the correct class. The Neural Network model can be represented as:

The Neural Network based optimization of weight factor is given below:



- Extracted features, target values and number of hidden neurons are given input to the Feedforward Network
- Each of these feature values are multiplied by a connection weight, W_n and these products are summed, fed through the transfer function, to generate a result as follows:

$$\text{Sum} = W_1 X_1 + \dots + W_n X_n$$

- Output node predicts the correct class (i.e.) normal or abnormal.

From the result, combined texture features help to achieve a very good result when compared with other extracted features.

4. Experimental result analysis and discussion

The research uses the data set obtained from MIAS [3]. The set consists of 322 images that fall into one of the following classes: 67 benign, 54 malignant and 201 normal images. Shape, Intensity, GLCM and GLRLM features are extracted from the pre-processed mammogram images. The extracted feature values are passed to train the neural network to classify as normal or abnormal mammogram images. The result shows that by combining the GLCM and GLRLM features accuracy of the classification to detect breast cancer can be improved.

The performance of the various features and their combinations of the hybrid technique is an efficient feature extraction and classification of breast cancer. The performance metric values like Sensitivity (SE), Specificity (SP) and Accuracy (AC) are used to evaluate the performance of the classifier. The formulas for the above values are given in Table III. Sensitivity is a proportion of positive cases, simultaneously the specificity is a proportion of negative cases, both are well detected by the test whereas classification accuracy depends on the number of samples that are correctly classified.

Table 3: Formula for Measures

Measures	Formula
Sensitivity	$SE = TP / (TP + FN)$
Specificity	$SP = TN / (TN + FP)$
Accuracy	$AC = (TP + TN) / (TP + FP + TN + FN)$

Where, TP is the number of true positives; FP is the number of false positives; TN is the number of true negatives; FN is the number of false negatives. Confusion matrix is shown in Table IV.

Table 4: Confusion Matrix

Actual	Predicted	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

TP - Predicts abnormal as abnormal.

FP - Predicts abnormal as normal.

TN - Predicts normal as normal.

FN - Predicts normal as abnormal.

To evaluate this research, the various feature extraction methods are trained with 55% of mammogram images, 10% of Validation and tested with 35% of mammogram images (46 malignant and 185 normal). Confusion Matrix for all feature extraction methods is shown in Fig 3. The result in Fig 4 shows the confusion matrix for GLCM and GLRLM features extraction techniques. Hybridization of texture feature (GLCM, GLRLM) improves better classification rate in sensitivity, specificity and accuracy which are consistent as compared to other methods.

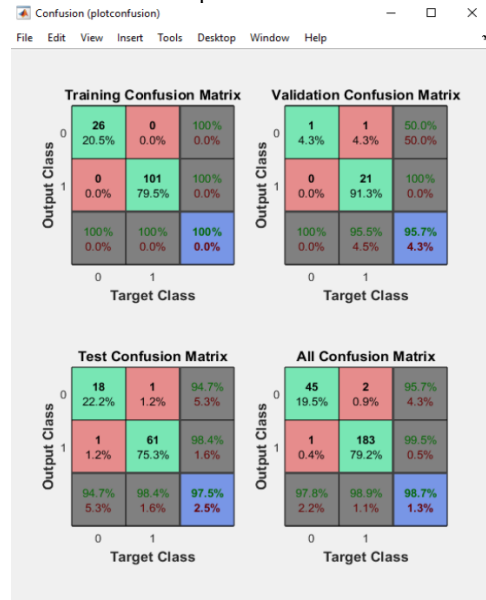


Fig 3: Confusion Matrix for All Feature Extraction Methods

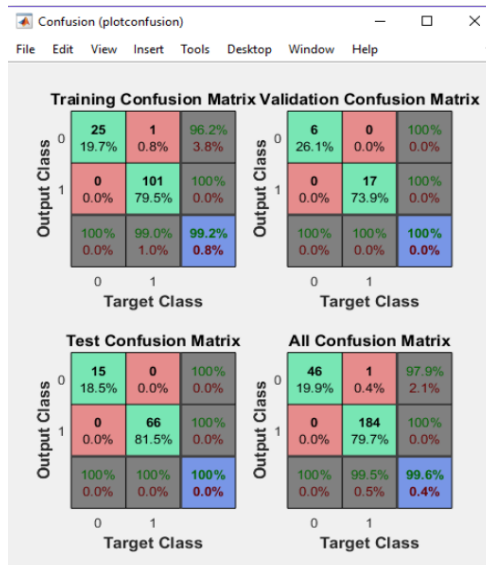


Fig 4: Confusion Matrix for GLCM and GLRLM Feature Extraction Methods

5. Conclusion and future scope

In this research, different feature extraction techniques are combined to provide more reliable and consistent performance accuracy. Neural Network is used to classify the mammogram image as normal or abnormal for early diagnosis of breast cancer. Hybridization of GLCM and GLRLM features produce high accuracy when compared to other features. From the experimental results, it is proved that 99.6% accuracy is achieved by GLCM and GLRLM feature extraction methods.

References

- [1] Nithya R & Santhi B, "Comparative Study on Feature Extraction Method for Breast Cancer Classification", *Journal of Theoretical and Applied Information Technology*, Vol.33, No.2, (2011).
- [2] A.M. Abdalla, S. Dress & N. Zaki, "Detection of Masses in Digital Mammogram using Second order Statistics and Artificial Neural Network", *International Journal of Computer Science and Information Technology (IJCSIT)*, Vol.3, No.3, (2011).
- [3] Suckling J, Parker J, Dance D, Astley S, Hutt I & Boggis, C, "The mammographic images analysis society digital mammogram database", *Exerpta Medical International Congress Series*, Vol. 1069, (1994), pp.375–378.
- [4] Monika Sharma RB & Dubey Sujata SKG, "Feature Extraction of Mammograms", *International Journal of Advanced Computer Research, (IJACSA)*, Vol.2 No.3, (2012).
- [5] Preetha K & Jayanthi SK, "Breast Cancer Detection and Classification using Artificial Neural Network with Particle Swarm Optimization", *International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)*, Vol.2, (2016).
- [6] Chidambaranathan S, "Breast Cancer Diagnosis Based on Feature Extraction by Hybrid of K-Means and Extreme Learning Machine Algorithms", *ARPN journal of engineering and Applied Sciences*, Vol.11, No.7, (2016).
- [7] Manavalan R & Thangavel K, "Comparitive Analysis of Feature Extraction methods for the Classification of Prostate Cancer from TRUS Medical Images", *International Journal of Computer Science Issues*, Vol.9, No. 1, (2012).
- [8] Ireaneus Anna Rejani Y & Thamarai selvi S, "Early Detection of Breast Cancer using SVM Classifier Technique", *International Journal on computer Science and Engineering*, Vol.1, No.3, (2009).
- [9] Adeyemo T, "Feature Extraction techniques for Mass detection in Digital mammogram (Review)", *Journal of Scientific Research and Reports*, (2012).
- [10] Anupa Maria S, "Textural Features based Breast Cancer Detection: A Survey", *Journal of Emerging trends in Computing and information Sciences*, Vol.3, No.9, 2012.
- [11] Haralick RM, Shanmugam K & Dinstein I, "Textural Features of Image Classification", *IEEE transactions on Systems, Man and cybermetrics*. Vol.3, No. 6, (1973), 610-621.
- [12] Neeta VJ & Mahadik SR, "Implementation of Segmentation and Classification Techniques for Mammogram Images", *International Journal of Innovative Research in Science Engineering and technology*, Vol.4, No. 2, (2015).