IFERP
connecting engineers... developing research

ISSN (Online) 2394-2320

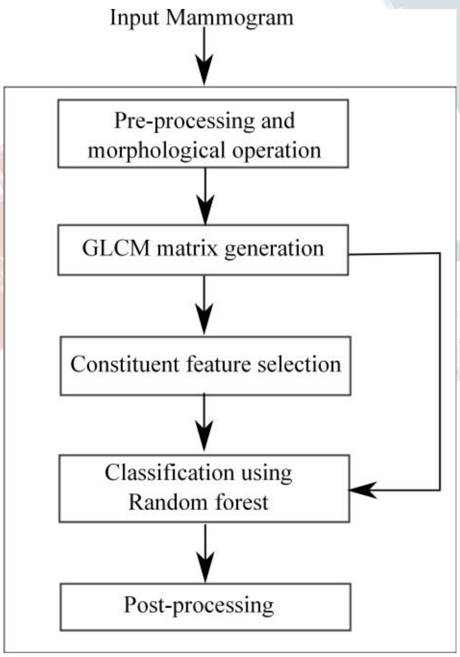
International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

Feature Vectors Generation for Mammogram Classification based on 2-D GLCM matrix

[1] Shehnaz Begum Sk, [2] T. K. Mishra
[1][2] Computer Science and Engineering
[1][2] Anil Neerukonda Institute of Science and Technology
Visakhapatnam, A.P., INDIA

Abstract - **Earlier is the diagnosis of a disease, better is the rate of recovery. So far as the fatal disease like breast cancer is concerned, it's early diagnosis may lead to improve the rate of care and thereby survival of a patient. Generally, breast cancer detection and analysis starts from capturing the Mammogram of the effected breast region. In this paper, an automated diagnosis scheme has been proposed for detecting the presence/ absence of breast cancer from such mammograms. Suitable pre-processing is applied to input mammogram images. For the feature extraction, the gray level co-occurrence matrix is framed out of the pre-processed image. The AdaBoost technique has been used for the purpose of feature selection. Classification is carried out with the help of the state-of-the-art Random-forest classifier. For the purpose of validation, the mammography image analysis society (MIAS) database has been taken into consideration. Satisfactory classification rate of 94% is achieved through the proposed scheme.**

I. INTRODUCTION


Input Mammogram
Pre-processing and morphological operation
GLCM matrix generation
Constituent feature selection
Classification using Random forest
Post-processing

Breast cancer is the most frequent cancer and constitute more than 20% of all but skin cancers in women worldwide [1]. Early detection is the key to reduce the number of cancer deaths and to improve patients' quality of lives. Mammography is considered an effective screening method for women with normal risk [2], [2]–[4]. It is not easy, however, to read a large number of mammograms accurately and consistently in a limited time. It is known that about 30% of cancers are missed on mammograms and the reported positive biopsy rates range from 12% to 46% ([5]–[9]). Even in multi-modality reading, it is important to assess images of each modality independently and thoroughly. Studies have suggested that the computer-aided detection and diagnosis (CAD) can contribute to accurate diagnosis of mammograms [10]–[15]. Computerized detection of micro-classifications on mammograms has very high accuracy. On the other hand, computerized classification of malignant and benign lesions still has some room for improvement. A number of studies investigating computerized methods for differentiating between malignant and benign masses have been proposed [16], [17]. Tan et al. [18], in their recent study, investigated a variety of different types of image features for classification of breast masses on mammograms. They found that the features related to mass shape, iso-density, and presence of fat were most frequently selected by their feature selection algorithm in a tenfold cross validation scheme. The reliability of these features depends on the accurate determination of mass contours. They also discussed the difficulty of accurate determination of speculation features due to tissue overlap.

II. PROPOSED SCHEME

In this section, the proposed work has been explored that focuses on the selection of GLCM features. The proposed scheme consists of five stages. The first stage is the input acquisition stage where details about the dataset has been presented. The second phase describes about the morphological operations performed on the acquired image to get the desired pattern. This is followed by the feature extraction and selection strategy using the AdaBoost feature selection. Finally, classification stage takes care of the proper classifier analysis for the purpose of classification. An overview of the proposed scheme is shown in Figure 1.

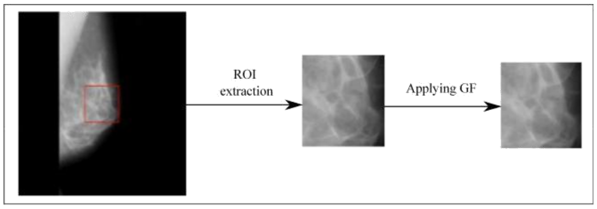


Fig. 2. Sample representation of the pre-processing.

A. Input Acquisition and Pre-processing

The Mammographic Image Analysis Society (MIAS) database of digital mammograms (v1.21) is a leading group for providing mammoram database. This contains the original 322 images (161 pairs) at 50 micron resolution in ”Portable Gray Map” (PGM) format and associated truth data. We have referred to this database for taking the input to our proposed work. These data are benchmarked and standardized. The X-ray mammogram images may contain certain types of scratches, labels, pectoral muscles. This may lead to certain obstacles during feature extraction from it’s digital version. To address this, we have cropping process to extract the ROI (region of interest) only from the complete image. To smoothen the resultant ROI-cropped images,a Gaussian filtering (GF) has been applied to all such images. Because this GF is till date the most effective filter for smoothening images in such cases. It reduces the additive noise if any present in an image. It uses a 2D-Gaussian blur convolution. All the pixels are mapped to a new but close values. A sample representation of the pre-processing has been shown in Figure 2.

The equation for GF in 1-D and 2-D has been presented below in (1) and (2) respectively:-

$$GF_1(u) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{u^2}{2\sigma^2}} \quad (1)$$

and,

$$GF_2(u,v) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{u^2+v^2}{2\sigma^2}} \quad (2)$$

Where, u and v are the distances of pixels along horizontal and vertical axes respectively. _ refers the the standard-deviation for the corresponding Gaussian distribution.

B. Feature Extraction

In this work, we first extract the texture feature from the mammogram. This has been done on the basis of the histogram based statistical moments of the input image. This concept focuses on the distribution of the intensities through out the image surface. However, we need to represent the pixel relationship as well to strengthen our feature vector for better classification. For this, we have preferred the GLCM extraction which is the second level of the statistical feature set. The GLCM matrix corresponding to an image (img) of size (M _ N) and the offset parameters δu and δv can be given by:-

$$L_{\delta u,\delta v}(x,y) = \sum_{u=1}^{N}\sum_{v=1}^{M} \begin{cases} 1, if img(u,v) = x, \\ img(u+\delta u, v+\delta v) = y \\ 0, otherwise \end{cases} \quad (3)$$

where, x and y are the intensities for img, u and v are the spatial coordinates of img.The values for the offset (_u; _v) are depending on the direction used (angle _), and the distance value d at which the matrix is computed. If these values changes, then the values of the co-occurrence matrix will also change. In our work, we have taken the values for d in the range 1 to 30 and the value for angle to be 0o, 45o, and 90o respectively. Ninety numbers of such GLCM matrices are generated in this manner. Now, from this GLCM, several statistical features can be extracted. As such, 14 such feature sets (texture based) can be generated. However, in this work, we have considered 4 such important feature sets, namely, correlation, contrast, energy, and homogeneity. Thus, the feature set consists of 90 _ 4 = 360 feature points in total. The specific formulae for individual parameter settings are given as under:-

1) Correlation: It determine the degree of neighborhood of a pixel with respect to other pixels. Mathematically,

$$Corr = \sum_{x,y} \frac{(x-\mu x)(y-\mu y)img(x,y)}{\sigma_x \sigma_y} \qquad (4)$$

2) Contrast: It determine the degree of intensity of a pixel with respect to other pixels. Mathematically,

$$Contrast = \sum_{x,y} [x-y]^2 img(x,y) \qquad (5)$$

3) Energy: It is the measure of the texture level of a pixel with respect to other pixels. Mathematically,

$$E = \sum_{x,y} img(x,y)^2 \qquad (6)$$

4) Homogeneity: It represents the level of uniformity of the pixel distribution in am image. Mathematically,

$$Hg = \sum_{x,y} \frac{img(x,y)}{1+[x-y]} \qquad (7)$$

C. Feature Selection

For the purpose, the AdaBoost algorithm has been used. This process tells about the comparative significance of a feature point. Thereby, it selects the best feature point among a set of given features. This contributes to the increase in the classification rate of any problem. It also simultaneously tells about the inter-dependencies of the feature points. In this work, this AdaBoost has been modified with the introduction of a parameter namely score that tells about how many times a particular feature point is being selected as the best during training process. The modified version of the algorithm has been presented in Algorithm 1.

Algorithm 1 Modified_AdaBoost_Algorithm

```
Input the gray-level image dataset $D = \{d_1, d_2, d_3 \ldots d_{max}\}$, the feature vector $F_v$, and the limiting value for number of iterations ($R$).
Initialize: $F_v^{new} = F_v$, and initial weights, $wt^{[1]} = wt_D^{[1]} = 1$
while $r \leq max$
    Apply normalization to the weights:
        $wt_D^{[r]} = \frac{wt_d^{[r]}}{\sum_{d \epsilon D} wt_d^{[r]}}$, for all $d \epsilon D$
    Set the classifier $h(x,\theta)$ for which the $error$ is minimum, where $error$ is computed as:-
        $Error_r = 0.5 - \frac{1}{2}(\sum_{i=1}^{max} wt^{(r-1)} y_i h(x,\theta_r))$   (8)
    Select the best so far feature set that minimizes the error obtained in the above step.
    Update weights such that the total sum should account to 1.
    Increment $r$
End while
```

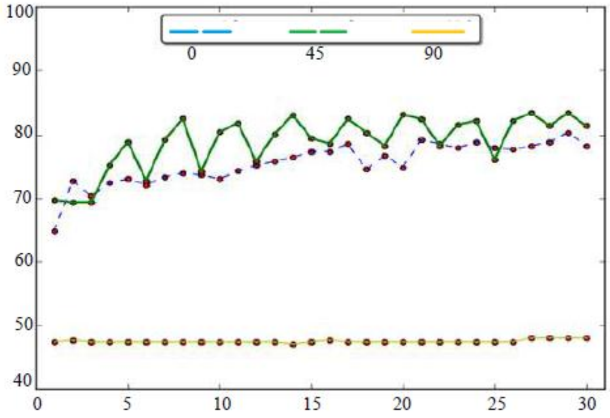


Fig. 3. Plot of accuracy for different angle values.

TABLE I

BEST NEIGHBORHOOD DISTANCE AND DIRECTION FOR PIXEL.

Angle θ	Best distance
0^o	$F_{v1}(d=21), F_{v2}(d=24), F_{v3}(d=28), F_{v4}(d=29)$
45^o	$F_{v5}(d=14), F_{v6}(d=20), F_{v7}(d=27), F_{v8}(d=29)$
90^o	$F_{v9}(d=16), F_{v10}(d=27), F_{v11}(d=28), F_{v12}(d=29)$

Initially, he algorithm 1 repeatedly selects a feature set as a weak learner. For each iteration r, it choose the feature which reduces the sum of training error (as computed) through the images corresponding to their weights. These features are selected as decision model for the purpose of classification. This decision model is similar to a decision tree having single level. The computed score is then incremented by 1. The weights are alloted to the mammogram images present in the training set. These weights are proportional to the computed error (*Errorr*). Thus, the relative impact of the images that were correctly classified by the selected feature reduces and thereby the weights of the images mis-classified by the weak learner increase. These weights can favor the training of the weak learner, for instance, decision trees can be grown that favor splitting subsets of images with enhanced weights. It also encourages the selection of features that performs well on the misclassified images by the classifier during the previous iteration and is complimentary to the previously selected feature. In this manner, AdaBoost inherently deals with the feature-correlation. The output of the proposed modified AdaBoost algorithm is a vector that is the relative feature significance of the original GLCM feature set. We set a certain threshold based on which the feature points whose values are more than that threshold are selected further. The accuracy values for all such thresholds are computed and among those, again the best threshold is selected.

D. Classifying the Mammograms

The features obtained after applying the algorithm stated in the previous section are now ready to be feed to a classifier. So far as the classification is concerned, the random forest classifier has been suggested to be an efficient classifier in the literature. The random forest algorithm developed by Brieman [] contains a set of ensemble of non-truncated decision tree classifiers which selects features points randomly at each instance. All of these decision trees generate decision vote measure for particular feature vectors based on which they split the total set of patterns. This algorithm has been suitably used for the proposed work for classifying the mammogram images into binary classes namely, normal, and abnormal.

E. Performance Analysis

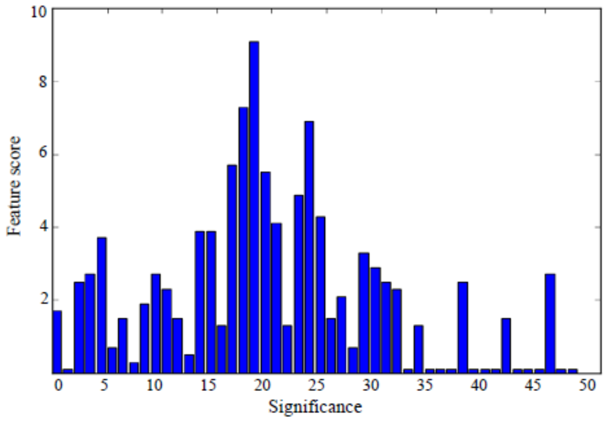


Fig. 4. Plot of feature versus their significance measure.

As mentioned earlier, the MIAS dataset has been used for validating the experimental proof of the proposed work. The original MIAS database was digitalized at 50 micron pixel edge which has been reduced to 200 micron pixel edge which makes every image of 1024 x 1024 pixels. This includes truth-markings on those locations where abnormality may be present. Some of the images consist of more than one abnormalities. Therefore, we get a total of 330 images, out of which 207 are normal, 69 are benign and 54 are malignant. So, there are 63% of normal data, 16% of malignant data and 21% of benign data. All images are 8-bit gray level images and they are in portable gray map (.pgm) format. The dataset is divided in 75:25 composition as the training and test dataset respectively.

After applying the pre-processing and morphological operations to the images, the feature vectors are extracted and selected thereby. This is followed by a proper experimental verification of the proposed scheme that how efficient the feature vectors are performing. Nevertheless to say that the feature sets here includes 4 features derived from constituent GLCM. Those are contrast, energy, homogeneity and correlation of gray level values for a particular pixel distance d and angle of direction θ The value for *d* and θ have been considered to be in the range of [1, 30] and [0_o; 45_o; 90_o]. In this manner, for each of the angle θ, we get 30 numbers of GLCM matrices.

The best four matrices are selected through individual calculations of the miss-classification errors in predicting the feature matrix. Performance of the GLCM features is dependent upon the relation between pixels, neighborhood and angle values. Finally, a total of 12 matrices and their feature set are obtained. Each of the vectors $F_{v1}; F_{v2}; : : : ; F_{v12}$ are the GLCM matrices containing four features contrast, energy, homogeneity and correlation. Therefore total size of the features become $3 \times 4 \times 4 = 48$. Now, for the classification task, the discrimination analysis of random forest algorithm is carried out using these 48 features, extracted from 12 different GLCM matrices. A sample is shown in Table I. The number of decision trees are fixed to be 100. Based on the confusion matrix obtained from this classifier, the best so far achieved accuracy, sensitivity and specificity are 94%, 90.56% and 86.20% respectively. The AdaBoost feature selection method, as described earlier is executed to select the best scoring features from the set of 48 feature points based on their significance. The significance of features are demonstrated in Figure ??. Some of the best scoring features are Energy, Homogeneity features of GLCM matrix of F_{v8} and Correlation of GLCM matrix of F_{v8} as shown in the Figure 4.

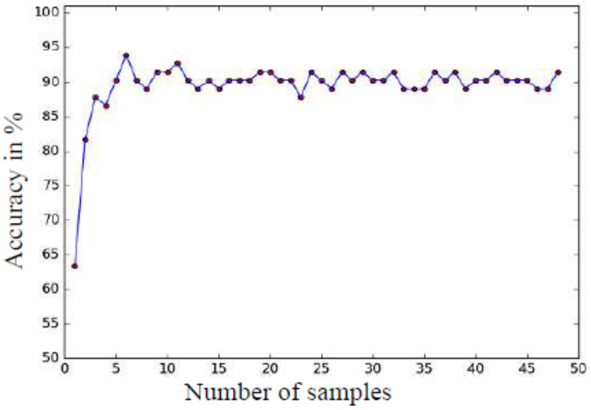


Fig. 5. Plot of rate of accuracy for the proposed scheme.

An analysis is made on varying the number of best features. Among these, the highest rate of accuracy we get is 94% which is obtained when number of best features is six. These are the Energy, Homogeneity, Contrast, Correlation features of GLCM matrix of Fv8 and Correlation, Homogeneity of GLCM matrix of Fv7. Based on these informative features, we trained the random forest classifier using 75 % of database images and test on the remaining 25 % images. The overall accuracy plot of the proposed scheme is shown in Figure 5.

III. CONCLUSION

In this paper, an efficient technique has been proposed for the classification of normal and abnormal categories of mammogram images. It is inferred that, due to challenging properties of MIAS database where visual appearances of all the mammograms are much close to each other, classification performance of this work is significantly satisfactory. Based on the exhaustive experiments conducted on GLCM matrices, for finding the best pixel distance and angles. It is concluded that proposed 48 GLCM features based on 3 different angles (0, 45, and 90) from four mentioned pixel distances, classified digital mammograms with 94% accuracy, 90.56% sensitivity and 86.40 % specificity using random forest classifier. This work can be further used suitably for other databases as well and the generic property thus obtained should be analyzed.

REFERENCES

[1] A. C. Society, Global Cancer Facts & Figures.

[2] S. D. N. D. A. G. O. G. L. Tabar, G. Fagerberg, “Update of the Swedish two-county program of mammographic screening for breast cancer,” Radiol. Clin. N. Am., vol. 02, no. 30, pp. 187–210, 1992.

[3] P. S. L. V. R. R. S. Shapiro, W. Venet, “Selection, follow-up and analysis in the health insurance plan study: a randomized trial with breast cancer screening,” J. Natl. Cancer Inst. Monogr., vol. 03, no. 67, pp. 65–74, 1985.

[4] B. C. S. W. L.L. Humphrey, M. Helfand, “Breast cancer screening: a summary of the evidence for the u.s. preventive services task force,” Ann. Intern. Med., vol. 04, no. 137, pp. 347–367, 2002.

[5] D. S. G. W. F.M. Hall, J.M. Storella, “Nonpalpable breast lesions: Recommendations for biopsy based on suspicion of carcinoma at mammography,” Radiology, vol. 167, no. 167, pp. 353–358, 1988.

[6] G. W. P. J. S. E. H. D.A. Hall, C.A. Hulka, “Positive predictive value of breast biopsy performed as a result of mammography: there is no abrupt change at age 50 years,” Radiology, vol. 200, pp. 357–360, 1996.

[7] R. B.-B. B. G. J. L. R. D. R. R. S.-B. B. Y. E.A. Sickles, D.L. Migioretti, “Performance benchmarks for diagnostic mammography,” Radiology, vol. 235, pp. 775–790, 2005.

[8] A. K.-L. H. G. A. R. S. J. H. S. D. Gur, L.P. Wallace, “Trends in recall, biopsy, and positive biopsy rates for screening mammography in an academic practice,” Radiology, vol. 235, pp. 396–401, 2005.

[9] L. A.-E. S. C. L. B. M. G. P. C. K. K. D. B. D. W. W. B. R. B.-B. R.D. Rosenberg, B.C. Yankaskas, “Performance benchmarks for screening mammography,” Radiology, vol. 241, pp. 55–66, 2006.

[10] M. U. T.W. Freer, “Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center,” Radiology, vol. 220, pp. 781–786, 2001.

[11] D. I. R.L. Birdwell, P. Bandodkar, “Computer-aided detection with screening mammography in a university hospital setting,” Radiology, vol. 236, pp. 451–457, 2005.

[12] J. R. T.E. Cupples, J.E. Cunningham, “Impact of computer-aided detection in a regional screening mammography program,” AJR, vol. 185, pp. 944–950, 2005.

[13] M. R.-T. W. D. A. C. P. J. S. N. S. S.-G. H.P. Chan, B. Sahiner, “Improvement of radiologists characterization of mammographic masses by using coputer-aided diagnosis an roc study,” Radiology, vol. 212, pp. 817–827, 1999.

[14] C. V.-C. M. Z. Huo, M.L. Giger, “Breast cancer: effectiveness of computer-aided diagnosis observer study with independent database of mammograms,” Radiology, vol. 224, pp. 560–568, 2002.

[15] R. S.-C. M. M. G. K. D. Y. Jiang, R.M. Nishikawa, “Improving breast cancer diagnosis with computer-aided diagnosis,” Acad. Radiol., vol. 06, pp. 22–32, 1999.

[16] J. L. D. R.M. Rangayyan, F.J. Ayres, “A review of computer-aided diagnosis of breast cancer: toward the detection of subtle signs,” J. Frankl. Inst., vol. 344, pp. 312–348, 2007.

[17] A. H. M. Elter, “Cadx of mammographic masses and clustered microcalcifications: a review,” Med. Phys., vol. 36, pp. 2052–2068, 2009.

[18] B. Z. M. Tan, J. Pu, “Optimization of breast mass classification using sequential forward floating selection (sffs) and a support vector machine (svm) model,” Int. J CARS, vol. 09, pp. 1005–1020, 2014.