International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 05, May-2015

# Mass Classification of Mammogram Images using Selected Textural Features with SVM Classifier

Nita Chayengia
Department of Electronics and Telecommunication Engineering
SVERI's COE Pandharpur, Solapur,
Maharashtra, India

Prof. Mrs. M. M. Pawar
Department of Electronics and Telecommunication Engineering
SVERI's COE Pandharpur, Solapur,
Maharashtra, India

***Abstract*—In mammography diagnosis systems, high False Negative Rate (FNR) has always been a significant problem since a false negative answer may lead to a patient's death. Development of a new Computer aided Diagnosis (CADx) system for the diagnosis of breast masses is the major objective of this paper. It aims at intensifying the performance of CADx algorithms as well as reducing the FNR by using Gray-Level Co-Occurrence Matrix (GLCM) for extraction of textural features. The input Regions of Interest (ROIs) are segmented manually and further subjected to feature extraction, selection and then classification. However from 322 MIAS database, 59 ROIs are taken into consideration for feature extraction. 19 texture features are extracted and 11 features are selected for feature classification. SVM classifier with Polynomial kernel and 80-20 train-test partition is used for classification. The Sensitivity, Specificity and Accuracy obtained by the selected features are 100%, 100%, and 100% respectively.**

***Keywords—Mammogram, ROI, Texture,SVM***

## I. INTRODUCTION

Cancer is a group of diseases that cause cells in the body to change rapidly and grow out in an uncontrolled manner. Basically cancer cells sooner or later form a lump or mass called a tumor. Breast cancer is found to be the most common cancer occurring in women all over India and accounts for 25% to 31% of all cancers in Indian women. For the year 2012, GLOBOCAN (WHO), estimated 70218 dead cases of women in India due to breast cancer. India ranks first in this regard compared to any other country in the world (second: China - 47984 deaths and third: US - 43909 deaths). In India, many non-oncology medical professionals such as General Surgeons, Gynecologists' etc. apt to treat breast cancer themselves, this leads to a lot of wrong decisions, unnecessary investigations, and painful surgeries. This directly has an effect on the outcome and longevity of the patient [8]. Breast cancer begins in the breast tissue called lobules, and also found in the ducts that connect the lobules to the nipple. Breast Cancer is commonly detected before or after symptoms are developed in a woman. Masses detected on a mammogram are basically benign or malignant. Most breast lumps turn out to be benign which are non-cancerous and are not life-threatening. But some masses turn out to be malignant; that is, they are cancerous and are life-threatening [9].

## II. LITERATURE SURVEY

Liu et.al [1] proposed a new feature selection method, known as SVM-RFE with an NMIFS filter (SRN). They achieved good accuracy rate with the SVM classifier using the selected features. These are the F-score (88%), Relief (88%), SVM-RFE (90%), SVM-RFE (mRMR) (91%), and SRN methods (93%), with a tenfold cross-validation procedure, and 91%, 89%, 92%, 92%, and 94%, respectively, with a leave-one-out (LOO) scheme. Tahmasbi et.al [2] paper is directed towards the development of a novel Computer-aided Diagnosis (CADx) system for the diagnosis of breast masses. Their objective is to intensify the performance of CADx algorithms as well as reducing the FNR by utilizing Zernike moments as descriptors of shape and margin characteristics. The designed systems yield Az ¼ 0.976, representing fair sensitivity, and Az ¼ 0.975 demonstrating fair specificity. The best achieved FNR and FPR are 0.0% and 5.5%, respectively. Chen et.al [3] proposed a rough set (RS) based supporting vector machine classifier (RS_SVM) for breast cancer diagnosis. In the proposed method (RS_SVM), RS reduction algorithm is employed as a feature selection tool to remove the redundant features and further improve the diagnostic accuracy by SVM. The proposed RS_SVM not only attains very high classification accuracy but also easily detect accurate breast diagnosis with a combination of five informative features. Zheng et.al [4] objective of research is to diagnose breast cancer based on the extracted tumor features. A hybrid of K-means and support vector machine (K-SVM) algorithms is developed to extract convenient information and identify the tumor. Based on 10-fold cross validation, the proposed methodology achieves accuracy rate of 97.38%, when tested on the Wisconsin Diagnostic Breast Cancer (WDBC) data set from the University of California – Irvine machine learning repository. Mousa et.al [5] introduced two methods based on wavelet analysis and fuzzy-neural approaches. These methods are mammography classifier based on globally processed image and on locally processed image (region of interest). This system classifies normal from abnormal masses and micro calcification. The estimation of the system is carried out on Mammography Image Analysis Society (MIAS) dataset. Guo et.al [6] proposed a novel method for breast cancer diagnosis using the feature generated by genetic programming (GP). A new feature extraction measure (modified Fisher linear discriminant analysis (MFLDA)) was developed to overcome the drawbacks of Fisher criterion. The capability of this technique is to alter information from high-dimensional feature space into one-dimensional space and automatically determine the relationship amongst data, to increase classification accuracy. Haralick et.al [7] calculated textural features based on gray tone spatial dependencies, and illustrates their application in category-identification tasks of three different kinds of image

data. He used two kinds of decision rules i.e., piecewise linear decision rule and min-max decision rule. Accuracy of test dataset is 89, 82 and 83 percent for photomicrographs, aerial photographic imagery and satellite imagery respectively.

## III. METHODOLOGY

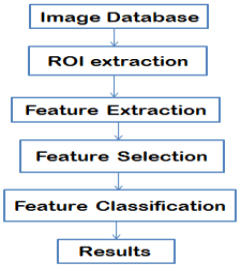



Fig.1. a) Original Image, b) ROI Image

### A. Image database

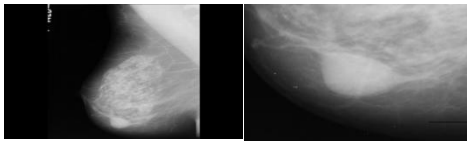
The mammogram images used in this experiment are taken from the mini mammography database of MIAS. The Mammographic Image Analysis Society, MIAS database contains all total 322 mammographic images in MLO which it contains 207 normal, 63 benign and 52 malignant cases. It has been found that images of MIAS database are basically stored in .pgm (Portable Gray Map) format. The original MIAS Database (digitized at 50 µm pixel edge) has been reduced to 200 µm pixel edge and clipped/padded so that every image is 1024 pixels×1024 pixels. All images are held as 8-bit gray level scale images with 256 different gray levels (0–255) [1]. The Mammographic Image Analysis Society has provided the image database for the purpose of research. In our experiment we have considered breast tissues which are fatty, fatty-glandular, dense-glandular and the abnormalities like, well-defined / circumscribed masses, spiculated masses and ill-defined masses as shown in Table I.

TABLE I. DISTRIBUTION OF MIAS MASSES

| Class | Benign | Malignant | Total |
|---|---|---|---|
| **Circumscribed masses** | 19+2 | 4 | 25 |
| **Spiculated masses** | 11 | 8 | 19 |
| **Ill – defined masses** | 7 | 8 | 15 |
| **Total** | **39** | **20** | **59** |

### B. Region of Interest

A region of interest (ROI) is a portion of an image that is to be filtered or performed some other operation on. ROI creates a binary mask, and this binary image have the same size as that of the image which is to be processed with pixels that define the ROI set to 1 and all other pixels set to 0. In an image we can extract one or more ROI. The regions can be defined by a range of intensities. As shown in figure 2.a) Original image having 1024x1024 pixels are cropped into 256x256 pixels in figure 2.b) keeping in mind that region of interest is not affected.



a) b)

Fig.2. a) Original Image, b) ROI Image

### C. Feature Extraction and Selection

After extracting ROI, calculate a set of features that is related to texture of the boundary and its neighbor regions. Characteristic of benign tumor are round, smooth, and well-circumscribed boundary, whereas the boundary of a malignant tumor is usually spiculated, rough, and blurry [3]. Thus, we can use a boundary analysis to classify the masses into benign or malignant. Although these features have been used in different publications, we will use only texture features in this paper to get better performance.

#### 1) *Texture Features from GLCM*

The texture information of the region surrounding the mass boundary contains important information to discriminate the benign and malignant masses. Thus, we have used the texture information for mass classification. The GLCM has been widely used in several applications, including the analysis of mammographic masses [1]. The features that are extracted from the GLCM are the autocorrelation, correlation, contrast, the cluster prominence, the cluster shade, the energy, entropy, homogeneity, the maximum probability, the sum of squares, the sum of average, the sum of variance, the sum of entropy, the difference in variance, the difference in entropy, the information measure of correlation (1), information measure of correlation (2), the inverse difference normalized, and the inverse difference moment normalized [6] as shown in Table 3. In the computation of the features, we scale the gray level to 16, and four GLCMs are constructed by scanning each mass ribbon at angles of 0◦, 45◦, 90◦, and 135◦, with the pixel distances set to 1. The GLCMs are averaged before the feature extraction. Thus, we obtain 19 texture features [1]. Because the offset is often expressed as an angle, the following Table II lists the offset values that specify common angles, given the pixel distance D.

TABLE II. OFFSET VALUES

| Angle | Offset |
|---|---|
| 0 | [0 D] |
| 45 | [-D D] |
| 90 | [-D 0] |
| 135 | [-D D] |

The figure 3 illustrates the array:

offset = [0 1; -1 1; -1 0; -1 -1]

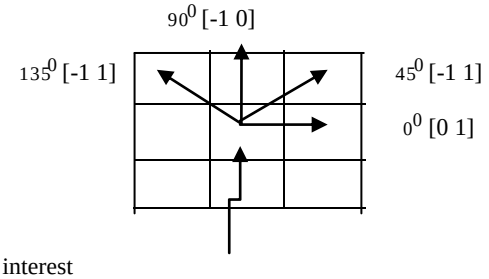



Fig.3. Gray-Level Co-occurrence Matrices (GLCM) showing offset values

A number of texture features may be extracted from the GLCM [1]. We use the following notation:

$N_g$ is the number of gray levels used.

$\mu$ is the mean value of P.

$\mu_x$ , $\mu_y$ , $\sigma_x$ and $\sigma_y$ are the means and standard deviations of $P_x$ and $P_y$ .

$P_x(i)$ is the ith entry in the marginal-probability matrix obtained by summing the rows of $P(i, j)$.

From the obtained features we have selected 11 features for feature classification based on SVM. These are the cluster prominence, cluster shade, energy, entropy, the sum of squares, the sum of average, the sum of variance, the sum of entropy, the difference in variance, the difference in entropy, and the information measure of correlation. The comparison of classification accuracy of the nineteen features done by the SVM model as shown in Table III.

TABLE III. TEXTURE FEATURES

| Feature Index | Feature | Formula |
|---|---|---|
| F1 | Autocorrelation | $F1=\sum_i\sum_j (i\times j)P(i,j)$ |
| F2 | Correlation | $F2=\frac{\sum_i\sum_j (i\times j)P(i,j)-\mu_x\mu_y}{\sigma_x\sigma_y}$ |
| F3 | Contrast | $F3=\sum_n n^2\{\sum_i\sum_j P(i,j)\}$ |
| F4 | Cluster prominence | $F4=\sum_i\sum_j (i+j-\mu_x-\mu_y)^4 P(i,j)$ |
| F5 | Cluster shade | $F5=\sum_i\sum_j (i+j-\mu_x-\mu_y)^3 P(i,j)$ |
| F6 | Energy | $F6=\sum_i\sum_j P(i,j)^2$ |
| F7 | Entropy | $F7=-\sum_i\sum_j P(i,j)\log(P(i,j))$ |
| F8 | Homogeneity | $F8=\sum_i\sum_j \frac{1}{1+(i-j)^2}P(i,j)$ |
| F9 | Maximum probability | $F9=\max_{i,j} P(i,j)$ |
| F10 | Sum of squares | $F10=\sum_i\sum_j (i-\mu)^2 P(i,j)$ |
| F11 | Sum of average | $F11=\sum_{i=2}^{2N_g} i\times P_{x+y}(i)$ |
| F12 | Sum of square / variance | $F12=\sum_i\sum_j (i-\mu)^2 P(i,j)$ |
| F13 | Sum of entropy | $F13=-\sum_{i=2}^{2N_g} P_{x+y}(i)\log\{P_{x+y}(i)\}$ |
| F14 | Difference in variance | $F14=Variance(P_{x-y})$ |
| F15 | Difference in entropy | $F15=-\sum_{i=0}^{N_g-1} P_{x-y}(i)\log\{P_{x-y}(i)\}$ |
| F16 | Information measure of correlation | $F16=\frac{HXY-HXY1}{\max\{HX,HY\}}$ |
| F17 | Inverse difference moment normalized | $F17=\sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1}\frac{p(i,j)}{1+(i-j)^2}$ |

### *D. Feature Classification*

To create this kind of feature-based classification, we should have some information of what features make noble analysis of class membership for the classes we are trying to differentiate [4]. The classification method with supervised learning involves two steps:

1. Training – this is where we discover what features are useful for classification by looking at pre-classified examples.

2. Testing – this is where we look at new examples and assign them to classes based on the features we have learned during training.

Features are trained and tested before passing through SVM classifier as shown in Figure 4.

During the process as shown in Figure 4, we tried to utilize an equal number of images taken from the MIAS database. Fifty nine ROIs were used for the experiments; among them, 39 were benign and 20 were malignant. In the process 12 masses were used as testing samples and 47 masses were used as the training samples. The SVM classifier was used as this classifier has six different kernel values here.

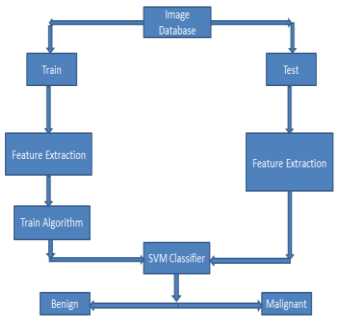



Fig.4. Feature Classification

#### *1) SVM Classifier*

The Support vector machines (SVM) is originally developed by Boser et al. (1992) and Vapnik (1995) is based on the Vapnik–Chervonenkis (VC) theory and structural risk minimization (SRM) principle [7]. This technique tries to find the tradeoff between reducing the training set error and increasing the margin, in the direction to achieve the best results. Performance of the classifier is measured by the true positive rate (TPR), the true negative rate (TNR), and the accuracy (ACC). The numbers of true positive in a classifier is represented as TP, false positive as FP, true negative as TN, and false negative number as FN. In order to assess SVM classifier prediction performance, we calculate the sensitivity, specificity and classification accuracy respectively as shown in Table IVand the equations are as given below:

$$TPR = \frac{TP}{TP + FN} \quad (1)$$

$$TNR = \frac{TN}{TN + FP} \quad (2)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

TABLE IV. TEXTURE FEATURES

| Metrics | Formula | Description |
|---|---|---|
| Sensitivity | $TPR = \frac{TP}{TP + FN} \times 100\%$ | Percentage of abnormalities correctly detected / classified as abnormalities |
| Specificity | $TNR = \frac{TN}{TN + FP} \times 100\%$ | Percentage of normal structures correctly detected / classified as normal |
| Accuracy | $ACC = \frac{TP + TN}{TP + FN + TN + FP} \times 100\%$ | Percentage of abnormalities and normal structures correctly detected / classified |

## IV. RESULTS

For the purpose of classification, it is anticipated that the linear separability of the mapped samples is enhanced in the kernel feature space so that applying traditional linear algorithms in this space could result in better performance compared to those obtained in the original input space. Inappropriate selection of kernel can give worse classification performance than that of the linear one. Therefore, selecting a proper kernel with good class separability plays a vital role in kernel-based classification algorithms. Here the Polynomial function performs well to do the task. A feature extraction method for finding the most significant features is proposed and implemented to detect and classify breast cancer in mammogram masses. The method is based on a constructing the database samples in 50-50%, 70-30%, 80-20% training-testing partition. The classification accuracy rate achieved by the proposed method using 50-50, 70-30 and 80-20 train-test partition is 79.31%, 94.44% and 100% respectively for benign versus malignant masses as shown in Table V.

TABLE V. TEXTURE FEATURES

| Metrics | Classification Accuracy | | |
|---|---|---|---|
| | 50-50 % train–test partition | 70-30 % train–test partition | 80-20 % train–test partition |
| Sensitivity (%) | 83.33 | 83.33 | 100 |
| Specificity (%) | 78.26 | 91.67 | 100 |
| Accuracy (%) | 79.31 | 94.44 | 100 |

As shown in Table VI sensitivity, specificity and classification accuracy for 50-50, 70-30 and 80-20 train-test increases respectively. We trust the proposed system can be very helpful in assisting the physicians to make the truthful diagnosis on the patients.

TABLE VI. TEXTURE FEATURES

| Training-testing partition (%) | Number of samples in set | | Train dataset Accuracy (%) | Test dataset Accuracy (%) |
|---|---|---|---|---|
| | Training set | Testing set | | |
| 50-50 | 30 | 29 | 100 | 79.31 |
| 70-30 | 41 | 18 | 100 | 94.44 |
| 80-20 | 47 | 12 | 100 | 100 |

## V. RESULTS

In this paper, we have studied and presented the results of the classification of breast masses with a data set of 322 images. For ROI extraction original image having 1024x1024 pixels are cropped into 256x256 pixels and we have considered 39 benign and 20 malignant masses of circumscribed, spiculated and ill-defined masses. After the ROI extraction, each mass was represented with19 texture features. Before classification, feature selection was performed with 59 ROIs. The SVM classifier using a Polynomial kernel is employed to perform the classification tasks on 59 ROIs. With the SVM classifier, Sensitivity, Specificity and Accuracy achieved in this paper is 100%, 100%, and 100% respectively.

## REFERENCES

[1] Amir Tahmasbi, Fatemeh Saki, Shahriar B.Shokouhi, "Classification of benign and malignant masses based on Zernike moments" Computers in Biology and Medicine 41 (2011) 726–735.

[2] Hui-Ling Chen, Bo Yang, Jie Liu, Da-You Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis" Expert Systems with Applications 38 (2011) 9014–9022.

[3] Bichen Zheng, Sang Won Yoon, Sarah S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms" Expert Systems with Applications 41 (2014) 1476–1482.

[4] Rafayah Mousa, Qutaishat Munib*, Abdallah Moussa, "Breast cancer diagnosis system based on wavelet analysis and fuzzy-neural" Expert Systems with Applications 28 (2005) 713–723Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[5] Hong Guo, Asoke K. Nandi, "Breast cancer diagnosis using genetic programming generated feature" Pattern Recognition 39 (2006) 980 – 987.

[6] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features of Image Classification", IEEE Transactions on Systems, Man and Cybernetics, vol. SMC-3, no. 6, Nov. 1973.

[7] Breast Cancer India : http://www.breastcancerindia.net/

[8] Breast cancer statistics : http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/breast-cancer-statistics.