I J C T A, 9(28) 2016, pp. 141-147
© International Science Press

CLASSIFICATION OF NORMAL AND ABNORMAL MAMMOGRAMS BASED ON DISCRETE WAVELET TRANSFORM AND SUPPORT VECTOR MACHINE

K.Vaidehi* T.S.Subashini and R.Manivannan*****

Abstract: Nowadays computer aided design / diagnosis plays a vital role in detection of breast cancer. This paper deals with an intelligent diagnosis system based on wavelet analysis and principle component analysis. Support vector machine classifier is used to classify mammograms as either normal or abnormal. Abnormal mammograms are those which include mammograms containing masses and microcalcifications. The effectiveness of this paper is examined on MIAS (Mammogram Image Analysis Society) database using accuracy, specificity, sensitivity and Mathew's correlation co-efficient.

Keywords : Mammograms, CAD, Wavelet, SVM

1. INTRODUCTION

Globally, the breast cancer is one of the top two leading cancers. As per the reports of Indian Council of Medical Research (ICMR), in India it is number one cancer and has surpassed cervical cancer. The breast cancer mortality rate has increased over the years. Early detection of breast cancer is essential to reduce mortality and to treat adequately. Achieving this will lead to better long term survival as well as a better quality of life. Indian Cancer Society has declared 2013 as a breast cancer awareness year and is taking various initiatives to create awareness among people.

Mammography is a radiographic image of a breast that provides information about breast, which aids in the early detection and diagnosis of breast diseases among women. With digital mammography system, breast images are acquired electronically and stored directly into the computer. Digital mammography process, guidelines and advantages are vividly explained in [1]. Computer aided detection/diagnosis (CAD) can be applied easily to the digital mammogram. Basically, CAD is based on image processing and pattern classification techniques. CAD system is used to support the radiologist to make important medical decisions through physician computer interaction [2].

Wavelet techniques have proved to be indispensable in image processing, particularly when dealing with medical images such as mammograms. Due to the wide variety of signals and problems encountered in medicine, the spectrum of applications of the wavelet transform is extremely large [3]. It ranges from the analysis of the more traditional physiological signals such as electrocardiogram (ECG) to the very

* Department of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Hyderabad, India, **Email:** vainakrishna@gmail.com

** Department of Computer Science and Engineering, Annamalai University, Chidambaram – 608 001, India, **Email:** rtramsuba@gmail.com

*** Department of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Hyderabad, India, **Email:** drmanivannan@stanley.edu.in

recent imaging modalities including magnetic resonance imaging (MRI), mammograms and positron emission tomography (PET). Wavelets provide a unifying framework for decomposing images, volumes, and time series data into their elementary constituents across the scale. In this work, wavelet coefficients are extracted from the image at different scales using wavelet decomposition and the dimensions of these coefficients are reduced using principle component analysis (PCA). The PCA reduced features are then used to model the classifier. In this work, Support Vector Machine (SVM) is used for classifying mammograms as normal or abnormal.

2. LITERATURE REVIEW

A survey of the image processing and pattern analysis techniques used by the various researchers in CAD for breast cancer is presented in [2]. The authors in [4] have proposed ensemble supervised algorithm, and using Gray-level Co-occurrence Matrices (GLCM) features for classifying the mammogram data into normal and abnormal. The authors in [5] classified the mammograms either as mass or normal breast tissue using convolution neural network and obtained 90% true-positive fraction. In that, multi-resolution texture features were obtained from the region of interest for classification. The authors classified mass and non-mass breast regions on mammograms applying taxonomic indexes and SVM [6]. In which taxonomic diversity index and the taxonomic distances are used to describe the texture features and that work is carried out using DDSM database. The authors in [7] developed a CAD based CBIR for retrieving benign and malignant mass mammograms using geometrical and Zernike moment features.

3. PROPOSED METHODOLOGY

The proposed method has three major procedures: first, the digital mammograms are pre-processed to enhance and to confine the region of interest; second, wavelet decomposition is applied to extract wavelet coefficients, and PCA is used to reduce the dimension of coefficients, which are used as features for classification; third, the SVM is used to classify mammograms as normal or abnormal. This method has been tested on freely available Mini-Mias database [8]. The block diagram of the proposed method is shown in Figure 1 and it consists of three major steps: (A) Pre-processing, (B) Feature extraction, (C) Feature reduction and (D) Classification.

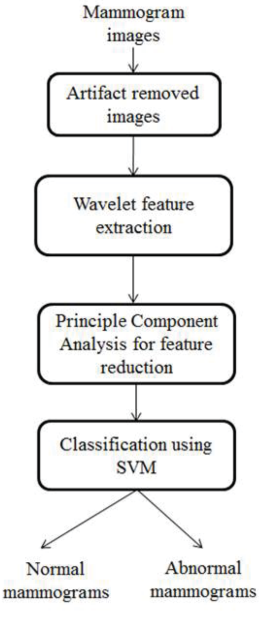

Figure 1: Block diagram of the proposed method.

3.1 Pre-processing

The images used for this work are taken from the MIAS database. Artifact removed mammogram images are taken as input images. This eliminates the need to process the background region unnecessarily. Contrast limited adaptive histogram equalization (CLAHE) method is applied for enhancing the whole image [9]. The input images are of large size 1024 x 1024 and almost 50% of the image comprises the background, the images are cropped to reduce their size to 800 x 800.

3.2 Feature Extraction

Wavelet texture measures are used to represent both normal and abnormal mammograms. The various steps involved in feature extraction are as follows:

Wavelet Decomposition

Wavelet decomposition is the first step in feature extraction. This operation returns the wavelet decomposition of the image at predefined scale, using the wavelet, Daubechies.

Wavelet transforms in two dimensions: In 2D, a two dimension scaling function $\phi(x,y)$, and three 2D wavelets, $\psi_H(x,y)$, $\psi_V(x,y)$, and $\psi_D(x,y)$ are required, as shown in equations 1 to 4.

$$\phi\,(x,y) = \phi(x)\phi(y) \tag{1}$$

$$\psi_H(x,y) = \psi(x)\psi(y) \tag{2}$$

$$\psi_V(x,y) = \psi(y)\psi(x) \tag{3}$$

$$\psi_D(x,y) = \psi(x)\phi(y) \tag{4}$$

These wavelets measure gray level variations of the images along different directions: ψ_H measures variations along columns, ψ_V corresponds to variations along rows, and ψ_D corresponds to variations along diagonals.

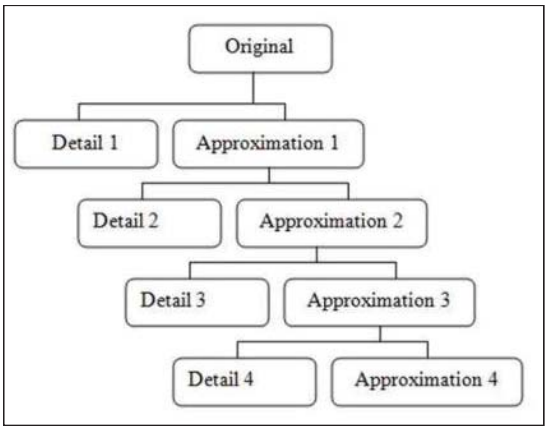

Figure 2. Wavelet decomposition

The decomposition operation returns the wavelet decomposition of the image at predefined scale, using the wavelet, Daubechies. The decomposition vector consists of horizontal detail coefficients, vertical detail coefficients and diagonal detail coefficients vectors. In this work, wavelet '*db*4' is used for the decomposition of the enhanced image.

The wavelet decomposition shown in Figure 2, starts with the original signal and fits the mother wavelet to it at the smallest scale. This produces what is called the first wavel**et "detail"** and a remainder called the **"approximation"**. Then, the time scale of the mother wavelet is doubled (called dilation/scale) and it is fit to the first approximation. This produces a second wavelet detail and the remainder is the second approximation, and the process continues until the mother wavelet has been dilated to such an extent that it covers the entire range of the signal.

Coefficient extraction

In this step, the horizontal *H*, vertical *V*, and diagonal *D* detail coefficient vectors at scale *N* are extracted by applying wavelet decomposition. These vectors are extracted at each scale excluding scale one. In this work, scale one coefficients are ignored because they contain high frequency details and noise. These details are insignificant information and will not affect the classification accuracy.

3.3 Feature Reduction

As the size of the mammogram taken is 800 x 800, the wavelet decomposition produces large number of coefficients. In this work, the numbers of coefficients are reduced by principal component analysis (PCA).

PCA involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The mathematical technique used in PCA is called Eigen analysis: Eigenvalues and Eigenvectors of a square symmetric matrix is solved with sums of squares and cross products. The Eigenvector associated with the largest Eigenvalue has the same direction as the first principal component. The Eigenvector associated with the second largest Eigenvalue determines the direction of the second principal component. The sum of the Eigenvalues equal the trace of the square matrix and the maximum number of Eigenvectors equal the number of rows (or columns) of this matrix.

In this work, PCA is applied to obtain the most relevant DWT coefficients from a large number of coefficients which are irrelevant and redundant. Implementation of PCA on the derived feature space could efficiently reduce the feature dimension without losing much information. Hence PCA is employed to reduce the dimension of the proposed feature space. These reduced features are used for modeling the SVM classifier.

3.4 Classification

Pattern classification techniques are often used to evaluate the effectiveness of features in normal versus abnormal discrimination. It is expected that the better the discrimination capabilities, the better the feature will serve in the objective representation of normal and abnormal images in a database. In this work, SVM classifiers are used to discriminate normal and abnormal mammograms.

4. PERFORMANCE MEASURES

Supervised Machine Learning (ML) has several ways of evaluating the performance of learning algorithms and the classifiers they produce. Measures of the quality of classification are built from a confusion matrix which records correctly and incorrectly recognized examples for each class. Table 1 presents a confusion matrix for binary classification, where *tp* stands for true positive, *fp* for false positive, *fn* for false negative, and *tn* for true negative counts. The various performance measures which are used to assess the classifiers performance are

Table 1.
Confusion matrix for binary classification.

Class / Recognized	*As Positive*	*As Negative*
Positive	*tp*	*Fn*
Negative	*fp*	*tn*

Accuracy assesses the overall effectiveness of the algorithm, by showing the probability of the true value of the class label.

It is given by

$$\text{accuracy} = (tp + tn)/(tp + fp + fn + tn)$$

Sensitivity and specificity are statistical measures of the performance of a binary classification test. **Sensitivity** measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of cancerous patients who are identified as having the condition). **Specificity** measures the proportion of negatives which are correctly identified (e.g. the percentage of healthy people who are identified as not having the condition). A theoretical, optimal prediction can achieve 100% sensitivity (i.e. predict all people from the cancerous group as cancerous) and 100% specificity (i.e. do not predict anyone from the healthy group as cancerous).

$$\text{sensitivity} = tp \,/\, (tp + fn)$$

$$\text{specificity} = tn \,/\, (fp + tn)$$

The **Positive Predictive Value (PPV)** is the fraction of the predicted benign class (positive) which is correct. The **Negative Predictive Value (NPV)** stands for the fraction of the malignant (negative) predictions which are correct.

$$\text{PPV} = tp \,/\, (tp + fp)$$

$$\text{NPV} = tn \,/\, (tn + fn)$$

Mathews Correlation Co-efficient (MCC) is calculated to get a better picture of the performance of the classifier, when the number of samples in the two classes is unbalanced. When compared to accuracy, MCC is used in cases where the number of samples in each of the classes differs considerably.

$$\text{MCC} = \{(tp*tn)(fp*fn)\} \,/\, \{\text{sqrt}\{(tp + fp)*(tp + fn)*(tn + fp)*(tn + fn)\}\}$$

5. EXPERIMENTAL RESULTS AND DISCUSSION

The images are cropped to the size 800 x 800 to remove the artifacts and unwanted background regions. 209 normal and 79 abnormal mammogram images are used to train and test the classifier. Initially, the mammograms are pre-processed to remove the artifacts, and the results are shown in Figure 3. CLAHE method is used to enhance the mammogram and the result is given Figure 4.

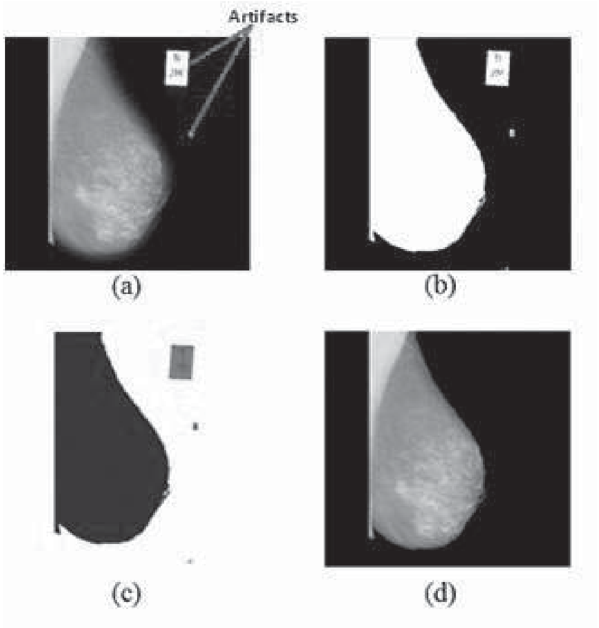

Figure 3. (a) Original input image (b) Binarized image (c) Connected component labeled image (d) artifact removed image.

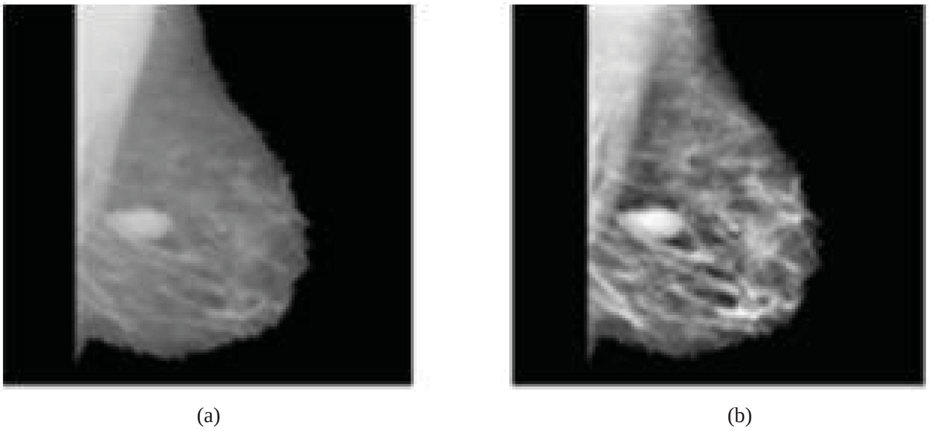

Figure 4. (a) Artifactless input image (b) CLAHE enhanced image

Wavelet decomposition is then applied to the enhanced image and the horizontal, vertical and diagonal co-efficients for levels two to five are extracted. In this work, wavelet 'db4' is used. To reduce the number of coefficients, PCA is then applied to obtain the significant first fifteen principal components. Finally, the reduced features are modeled with SVM classifier. Reduced features from level 2 to 5 are used to train and test the classifiers. All the features are first normalized between -1 and +1 for the classifier to have a common range to work with.

Performance measures of SVM in %.

Performance measures	*Level 2*	*Level 3*	*Level 4*
Accuracy	93.40	87.15	75.35
Sensitivity	92.34	86.60	76.56
Specificity	96.20	88.61	72.15
PPV	98.40	95.26	87.91
NPV	82.61	71.43	53.77
MCC	84.73	70.82	45.06

Leave one out procedure has been adopted in testing the performance of the SVM classifier. The SVM with polynomial kernel is trained to provide a value of 0 for normal mammograms and 1 for abnormal mammograms. The classifier output for the test data is compared with the original class attribute for identifying true positives, true negatives, false positives and false negatives. Table 2 shows the classification result obtained using PCA applied wavelet features.

6. CONCLUSIONS

In this work, the mammogram is first diagnosed as normal or abnormal using PCA reduced wavelet features and diagnosis is done using SVM. The classification achieves the best performance with features extracted from level 2, because mass and microcalcification are represented as high frequency information which is obtained in the highest wavelet decomposition levels. The level 2 achieves best accuracy of 93.40% and sensitivity of 92.34%. After the fourth level the performance degrades gradually. This work shows, CLAHE coupled with wavelet features and SVM classifier is very effective for automatic classification of normal and abnormal classes in digital mammograms.

7. ACKNOWLEDGMENT

The authors would like to thank Correspondent and Principal of Stanley College of Engineering and Technology for Women, Hyderabad for their valuable help and support. Their help is greatly acknowledged with appreciation.

References

1. Mark P.Bowes, "Digital Mammography: Process, Guidelines and potential advantages," *eRadimaging*, (2012) URL: www.eradimaging.com/site/article.cfm.
2. Rangaraj M Rangayyan, Biomedical image analysis, CRC press, 2004.
3. Unser, Michael, and Akram Aldroubi. " A review of wavelets in biomedical applications", Proceedings of the IEEE 84.4 (1996): 626-638.
4. Banaem, hossein Yousefi, Alireza Mehri Dehnavi, and Makhtum Shahnazi. "Ensemble supervised classification method using the regions of interest and grey level co-occurrence matrices features for mammograms data", Iranian Journal of Radiology, 12(3), 2015.
5. Wei, Datong, Heang-Ping Chan, Mark A, Helvei, "Classification of mass and normal breast tissue on digital mammograms" multiresolution texture analysis", Medical Physics, 22.9 (1995): 1501-1513.
6. De Oliveira, F.S.S. de Carvalho Filho, A.O., Silva, A.C., de Paiva, A.C., & Gattass, M. "Classification of breast regions as mass and non-mass based on digital mammograms using taxonomic indexed and SVM", Computers in biology and medicine, 57, (2015): 42-53.
7. K.Vaidehi, T.S.Subashini, "Content Based Benign and Malignant Mass Mammograms Retrieval", Internaltional journal of Applied Engineering Research, 21(9), (2014). ISSN : 0973-4562.
8. J.Suckling, J.Parker, D.Dance et al., "The Mammogram Image Analysis Society Digital Mammogram Database," Exerpta Medica, *International Congress Series*, vol.1069, pp.375-378, 1994.
9. Vaidehi, K., and T. S. Subashini. "Automatic Characterization of Benign and Malignant Masses in Mammography." *Procedia Computer Science* 46 (2015): 1762-1769.