Applied Soft Computing 46 (2016) 151–161

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Application of Gabor wavelet and Locality Sensitive Discriminant Analysis for automated identification of breast cancer using digitized mammogram images

U. Raghavendra[a,*], U. Rajendra Acharya[b,c,d], Hamido Fujita[e], Anjan Gudigar[a], Jen Hong Tan[b], Shreesha Chokkadi[a]

[a] Department of Instrumentation and Control Engineering, Manipal Institute of Technology, Manipal University, Manipal 576104, India
[b] Department of Electronics and Computer Engineering, Ngee Ann Polytechnic, Clementi, 599489, Singapore
[c] Department of Biomedical Engineering, School of Science and Technology, SIM University, Clementi, 599491, Singapore
[d] Department of Biomedical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur 50603, Malaysia
[e] Iwate Prefectural University (IPU), Faculty of Software and Information Science, Iwate, Japan

ARTICLE INFO

Article history:
Received 11 January 2016
Received in revised form 9 April 2016
Accepted 27 April 2016
Available online 3 May 2016

Keywords:
Gabor wavelet
LSDA
Interclass variation
DDSM
Mammography
CAD tool

ABSTRACT

Breast cancer is one of the prime causes of death in women. Early detection may help to improve the survival rate to a great extent. Mammography is considered as one of the most reliable methods to pre-screen of breast cancer. However, reading the mammograms by radiologists is laborious, taxing, and prone to intra/inter observer variability errors. Computer Aided Diagnosis (CAD) helps to obtain fast, consistent and reliable diagnosis. This paper presents an automated classification of normal, benign and malignant breast cancer using digitized mammogram images. The proposed method used Gabor wavelet for feature extraction and Locality Sensitive Discriminant Analysis (LSDA) for data reduction. The reduced features are ranked using their *F*-values and fed to Decision Tree (DT), Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), *k*-Nearest Neighbor (*k*-NN), Naïve Bayes Classifier (NBC), Probabilistic Neural Network (PNN), Support Vector Machine (SVM), AdaBoost and Fuzzy Sugeno (FSC) classifiers one by one to select the highest performing classifier using minimum number of features. The proposed method is evaluated using 690 mammogram images taken from a benchmarked Digital Database for Screening Mammography (DDSM) dataset. Our developed method has achieved mean accuracy, sensitivity, specificity of 98.69%, 99.34% and 98.26% respectively for *k*-NN classifier using *eight* features with 10-fold cross validation. This system can be employed in hospitals and polyclinics to aid the clinicians to cross verify their manual diagnosis.

1. Introduction

Breast cancer is the second major cause of death in women. Nearly, 1.7 million new breast cancer cases are diagnosed worldwide in 2012 according to the Globocan project [1]. However, mortality rates, due to breast cancer, have been reducing due to better diagnostic facilities and effective treatments.

Breast Self-Examination (BSE) and Clinical Breast Exam (CBE) are the two methods that women can go for. However, these methods are not capable of detecting the cancer at its earliest stage. Mammography is the popular technique designed to image the breast [2]. It comprises of an X-ray systems that permits scanty application of X-ray, high contrast, and high resolution detectors. Mammography has been proven to be most effective in screening and diagnosis [3]. Generally, masses and calcifications are the two abnormalities present in the mammogram images [2]. Presence of a star-shaped structure around mass lesions characterizes the malignant tumors. Usually, spicules exhibit very weak contrast with respect to their neighboring regions on mammogram [4]. Based on the shape, masses can be classified as either benign or malignant. Benign tumors are usually in round or oval shape whereas malignant tumors are partially rounded shape with irregular outlines. Analyzing these mammogram images is a challenging task for the radiologists. Their interpretation depends on certain criteria such as experience, training etc., Sometimes; radiologists may miss breast cancer due to fatty and dense breasts and false positive judgment results a biopsy for diagnosis.

* Corresponding author.
E-mail address: raghavendra.u@manipal.edu (U. Raghavendra).

http://dx.doi.org/10.1016/j.asoc.2016.04.036
1568-4946/

Consequently, Computer-Aided Diagnosis (CAD) assists the radiologists to analyze the mammogram images. The general framework for CAD system using mammogram images consists of two major processing stages namely, feature extraction and classification. Most of the work additionally includes pre-processing stage prior to feature extraction, to enhance the quality of Region of Interest (ROI). Few of them are thresholding [5–7], region-based techniques [8–10] and edge detection techniques [11,12]. Feature extraction is a basic step to observe the nature of different classes of mammogram images. It extracts discriminative information, which can be utilized to efficiently classify into normal, benign and malignant lesions. Commonly used feature extraction techniques are Gabor filter [13,14], set of filter banks [15], Zernike moments [16] and Discrete Wavelet Transform (DWT) [17–25]. Most of the above approaches commonly use feature extraction algorithms followed by machine learning algorithms to intelligently classify the samples into normal and abnormal cases. Commonly used classification methods are artificial neural networks [7,13,22,26] *k*-nearest neighbor [27], fuzzy [20,28], Support Vector Machine (SVM) [6,27,29,30–32,33] and Linear Discriminant Analysis (LDA) [34].

Multiresolution analysis of mammogram is one of the popular techniques as they highlight abnormalities in the extracted texture features. The original mammogram image is decomposed into several sub-bands that can preserve information regarding both high and low frequencies. Wavelet transform is one among these techniques. Variety of multiresolution based techniques is introduced in the recent years [17–25]. Wei et al. [35] have used multi resolution texture analysis to differentiate normal and abnormal tissues and achieved an Area Under Curve (AUC) of 0.96. Another approach uses textural information of the mammogram image as a statistical measure of a pixel in a region. This information is essential to discriminate normal from benign and malignant patterns. The most popular method of texture analysis is Gray Level Co-occurrence Matrix (GLCM) to compute second order statistical measures of an image [36–38]. Chan et al. [39] have proposed Spatial Gray Level Dependence matrices based approach with genetic algorithm as a feature selection technique and achieved AUC of 0.89. In Ref. [40], GLCM is optimized using Hybrid Particle Swarm Optimization and Genetic algorithm and achieved the classification accuracy of 94%. Recently, in order to increase the overall efficiency, a classifier is accompanied with the novel feature weight adjustments technique [3]. The detailed review of the existing classification techniques is presented in Ref. [2].

Most of the studies include constrained sample size during the experiment [3,5,6,7,41] and the method may not be extended to large samples as majority of them concentrate on local appearances [3,5,7,41]. Many techniques use morphological operations [42] and various methods for image enhancement. But, deciding a proper structuring element is difficult [43]. Also a single texture feature is not sufficient to model benign and malignant masses due to slight structural difference. Moreover, analysis of mammogram images with a particular orientation and frequency may not be able to extract significant information about the abnormalities.

The bank of Gabor filters is one of the efficient tools, which can extract the directional textural features [44] and these directional textural features represent the structural properties of masses in mammograms at different orientations and frequencies. However, dimensionality of these extracted features is high due to various orientations and frequencies [44]. In general, dimensionality reduction techniques may degrade overall performance and every extracted feature may not contribute in enhancing classification accuracy [2,45]. Thus, there is a need for efficient dimensionality reduction technique with feature selection scheme for better accuracy. This paper proposes a novel way of addressing the above issues using Gabor filters coupled with Locality Sensitive Discriminant Analysis (LSDA). Fig. 1 describes the proposed method with a detailed block diagram. The local appearance of the breasts is characterized by Gabor wavelet features. LSDA is used to reduce the obtained dimensionality of the original data space in order to get most discriminative features among the different classes. Then the features are ranked according to their *F*-value. Finally, these features are fed to different set of classifiers such as Decision Tree (DT), Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), *k*-Nearest Neighbor (*k*-NN), Naïve Bayes Classifier (NBC), Probabilistic Neural Network (PNN), Support Vector Machine (SVM), AdaBoost and Fuzzy Sugeno (FSC) to obtain the best classifier.

This manuscript is divided into six different sections with Introduction in Section 1. Section 2 briefly describes the details of data acquisition and techniques used for the feature extraction and selection. Section 3 presents different classifiers used. Section 4 gives the obtained results and the discussions are given in Section 5. Finally, conclusion of the paper is presented in Section 6.

2. Materials and method

2.1. Data acquisition

In this work, we have used 690 mammogram images taken from Digital Database for Screening Mammography (DDSM) [46,47]. This database contains approximately 2620 mammogram images in terms of three classes i.e. normal, benign and malignant in 43 volumes. In this work, we have used 230 normal, 230 benign and 230 malignant mammogram images and the images are resized to 1360×794 using bicubic interpolation [48]. Since the aspect ratio (ratio of the width to the height of the image) of the resized image is kept the same as that of the original image, the quality of the resized image is comparable to that of the original image.

2.2. Pre-processing

In order to achieve the reliable and acceptable accuracy, pre-processing is performed on the mammogram images [49]. The main aim of this stage is to eliminate the undesirable information, such as margins, labels, and some regions of tissues (pectoral muscle). In order to achieve this, images are first smoothed using average filter of size 9×9 and is binarized using global thresholding technique (with threshold $\mathrm{Th}_1 = 0.1$) and labels are removed. The obtained binary mask is projected on the original image and is smoothed again by applying average filter of size 39×39. Further, smoothed image is subjected to global thresholding with a threshold (Th_2) of 0.5. In order to remove the pectoral muscle, the region having an aspect ratio of greater than 1.7 is removed. These parameters value is selected using stimulation through which the selected values could provide the best accuracy in terms of computation. Also these parameters are fixed for the entire dataset. Fig. 2 shows the steps involved in the elimination of pectoral muscle.

2.3. Feature enhancement and extraction

Feature extraction is an essential stage for interpretation and analysis of digital images. In this work, input images are subjected to morphological filtering, namely top-hat and bottom-hat [50]. The obtained images using these two operators' I_{T} and I_{B} respectively are described as

$$I_{\mathrm{T}} = I_{\mathrm{P}} - I_{\mathrm{O}}, I_{\mathrm{O}} = I_{\mathrm{D}} o I_{\mathrm{E}} \tag{1}$$

$$I_{\mathrm{B}} = I_{\mathrm{C}} - I_{\mathrm{P}}, I_{\mathrm{C}} = I_{\mathrm{E}} o I_{\mathrm{D}} \tag{2}$$