Detection of Breast Abnormalities via Fisher Linear Discriminant and Nearest Neighbor Classifier

Ikhlas Abdel-Qader[*1], Memuna Sarfraz[2], Fadi Abu-Amara[3]

[1]*Department of Electrical and Computer Engineering, Western Michigan University*
MI 49008, Tel: +12692763146, USA
ikhlas.abdelqader@wmich.edu

[2]*Eaton Corporation, Galesburg*
MI 49053, USA
memuna.sarfraz@wmich.edu

[3]*Department of Computer Engineering, College of Engineering, Al-Hussein Bin Talal University*
Ma'an, Jordan
fadiabuamara@ahu.edu.jo

***Abstract*—Early detection of breast cancer has proven to save lives. Today, mammography is the best method for early detection of breast cancer. Studies have shown that 10% - 30% of mammograms result in abnormal diagnosis. Out of these, two-thirds are false negatives (FN), which are caused by the inability of the radiologist to detect abnormalities due to various reasons such as poor image quality, image noise, or eye fatigue. This paper proposes a computer aided detection tool to assist radiologists in enhancing diagnosis outcome. The proposed abnormalities detection tool integrates Principal Component Analysis (PCA), Fisher Linear Discriminant (FLD), and Nearest Neighbor Classifier (KNN) algorithms to detect abnormalities in mammograms. The proposed method was tested using normal and abnormal mammograms from the MIAS database and resulted in 94.1% classification accuracy. The results also show that the proposed method is highly dependent on its parameters and suggestions for their selection and future work are provided.**

***Keywords*—Principal Component Analysis, Fisher Linear Discriminant, Nearest Neighbor Classifier.**

I. INTRODUCTION

Breast cancer is the most common form of cancer in women. In the USA, every three minutes a woman is diagnosed and every thirteen minutes a woman dies due to this disease. Till now, over two and a half million women have been treated from this cancer [1]. The National Cancer Institute also estimates that 12.7 percent of woman born today will be diagnosed with breast cancer at some time in their lives [1].

The only possible remedy and treatment for breast cancer is the early detection as it has considerably reduced the mortality rates in the past years [2].

The subtle differences between cancerous and normal regions in mammograms make the radiologist job of identifying abnormal regions difficult and burdensome. Computer Aided Detection (CAD) is a system used to assist radiologists through reading, analyzing, and then sorting out the mammograms as normal/abnormal.

The objective of this research paper is to develop a CAD system that confidently provides the radiologist a second reader opinion about mammographic images. The literature reported that sensitivity of mammography has been improved by 15% - 20% with a CAD system [3]. The proposed CAD system integrates PCA as a decorrelation-based module, FLD as a dimensionality reduction and feature extraction module, and KNN as a classification module.

The rest of this paper is organized as follows: section 2 presents PCA, FLD, and KNN algorithms. The proposed integrated approach is presented in Section 3. Section 4 presents the experimental results followed by the conclusions in Section 5.

II. BACKGROUND

A. Principal Component Analysis

PCA is a linear transformation and a decorrelation-based technique that maps a high dimensional space into a lower dimensional space. PCA is used as a preprocessing step to improve speed and accuracy of the classification stage while decreasing its complexity.

B. Fisher Linear Discriminant

The linear discriminant analysis (LDA) can be used to discriminate between data classes [4]. On the other hand, the Fisher linear discriminant (FLD) is the benchmark for the linear discrimination between two classes in the multidimensional space [4]. FLD was reported with attractive computational complexity

since it is only based on the first and second moments of the data distribution [4].

The LDA uses a projection matrix *W* to reshape the data set's scatter matrix in order to maximize the class separability. The matrix *W* represents the optimally discriminating features and is defined as the ratio of between class scatter to within class scatter.

The PCA algorithm transforms the data into an Eigenspace that uncorrelates the data. However, in case of a two-class problem, the two classes are not completely separable which complicates and degrades the classification phase. Therefore, the FLD algorithm can be applied after PCA resulting in a better between class scatter. Consequently, the classification stage should be improved.

C. Nearest Neighbor Classifier

The Nearest Neighbor is a simple yet a robust classifier where an object is assigned to the class to which the majority of the nearest neighbors belong. It is important to consider only those neighbors for which a correct classification is already known (i.e., training set). All the objects are considered to be present in the multidimensional feature space and are represented by position vectors where these vectors are obtained through calculating the distance between the object and its neighbors. The multidimensional space is divided into regions utilizing the locations and labels of the training data. An object in this space will be labeled with the class that has the majority of votes among the k-nearest neighbors.

III. PROPOSED CAD ALGORITHM

In this section, a computer aided detection algorithm of suspicious regions in mammograms is developed. PCA algorithm is used as a decorrelation-based module followed by FLD as a dimensionality reduction and a feature extraction module. Finally, a KNN classifier is used to classify the testing sub-images into normal or abnormal.

A. Sub-Images Generation

The MIAS database has a total of 119 suspicious and 203 normal mammograms. 144 sub-images are cropped and scaled into 50x50 pixels to localize the area of suspicion.

A total of 3 training sets are used. Each training set consists of 48 sub-images comprising of 24 abnormal and 24 normal sub-images. Let a training set be represented as G_{jk}= [g_1, g_2,…, g_k] where j = 2500 and k=1,2,…,48.

B. Unsupervised Learning

The training phase can be summarized as follows:

- Each sub-image in the training set is converted into a column vector *g* of dimension 2500 x 1. Then, a training matrix G_{jk} is formed by placing the sub-images as columns.
- Row-wise mean of the matrix G_{jk} is computed which results in a column vector *A* of dimension 2500 x 1.
- A matrix B_{jk} is formed by repeating the column vector *A* number of times equal to number of the sub-images.
- The deviation of each sub-image from the row-wise mean of the sub-images is calculated per $D_{jk} = G_{jk} - B_{jk.}$
- The covariance matrix of D_{jk} is computed using equation (1):

$$C_{mm} = \frac{1}{M} \sum_{k=1}^{M} D_{jk} D_{jk}^{T} \tag{1}$$

where *M* is the number of rows in *A*. The dimension of matrix *C* is 48 x 48.

- The eigenvalues λ and eigenvectors *V* of the matrix C_{mm} are computed using the PCA algorithm according to equation (2).

$$C_{mm} V = \lambda V \tag{2}$$

- The centered sub-images matrix D_{jk} is projected onto the Eigenspace per equation (3).

$$Y_{mm} = V^{T} \times D_{jk} \tag{3}$$

Two types of scatter matrices are used in this step. The first one is the within-class scatter matrix S_W representing the scatter of a single class and the second one is the between-class scatter matrix S_B representing the scatter of different classes:

$$S_W = \sum_{i=1}^{c} \sum_{y_k \in C_i} (y_k - \mu_i)(y_k - \mu_i)^T \tag{4}$$

$$S_B = \sum_{i=1}^{c} (\mu_i - \mu)(\mu_i - \mu)^T \tag{5}$$

where *C* represents number of classes, μ_i is the mean of samples in class *i* where *i*={1, 2}, and μ is the mean of all samples in the training matrix. Both S_B and S_W are of dimension 48 x 48.

- A linear transformation matrix *W* is computed as:

$$W = \frac{\det(Y^T S_B Y)}{\det(Y^T S_W Y)} = [w_1, w_2, ..., w_m] \tag{6}$$

where $\frac{\det(Y^T S_B Y)}{\det(Y^T S_W Y)}$ is the Fisher criterion that maximizes the between-class scatter while minimizes the within-class scatter. The transformation W is another projection into the Eigenspace such that

$$S_B w_i = \delta_i S_W w_i, \qquad i = 1, 2, \ldots, m \tag{7}$$

where w_i is the set of m Eigenvectors and δ_i is the set of m eigenvalues of S_B and S_W.

A number of eigenvalues are retained (N_{ev}) along with their corresponding Eigenvectors (V_{fe}) where the dimension of V_{fe} is MxN_{ev}.

- The matrix Y_{mm} is projected onto the Fisher linear space Z_{pq} using the Eigenvectors V_{fe} as shown in equation (8).

$$Z_{pq} = V_{fe}^T \times Y_{mm} \tag{8}$$

The dimension of Z_{pq} is N_{ev} x 48.

C. Testing Phase

The testing phase can be summarized in the following steps:

- Each test set consists of 48 sub-images: each test set consists of 24 abnormal and 24 normal sub-images resulting in a total of three test sets.
- Let the testing set be represented as $[t_1, t_2, \ldots, t_{48}]$.
- For each testing sub-image t_i, the difference between the sub-image and the mean of the training set A is computed using equation (9).

$$\gamma_{jk} = t_i^t \quad A, \quad i = 1, 2, \ldots, M \tag{9}$$

- The difference γ_{jk} is projected onto the Eigenspace Y_{mm} and the Eigenvectors space V_f as shown in equation (10).

$$P_{st} = V_f^T \times Y_{mm}^T \times \gamma_{jk} \tag{10}$$

where the dimensions of P_{st} is N_{ev} x 48.

D. Classification Phase

The classification phase can be summarized in the following steps where two classes were assumed handling the abnormal and normal sub-images.

- The Euclidean distance between the testing matrix P and each column of the Fisher linear space Z is computed.
- The nearest neighbor to the test sample is selected based on the calculated distances. Then, the test sample is assigned to the class of the nearest neighbor.

IV. EXPERIMETAL RESULTS

For the two-class problem of this work, eleven Fisher values are retained.

Table 1 shows the results of the proposed CAD algorithm with each test set consists of 24-normal and 24-abnormal sub-images. As the table indicates, the proposed CAD algorithm has classification accuracy over 91.67% in all three test sets using 144 images from MIAS database with average classification accuracy of 93.06% which indicates robustness of the proposed CAD algorithm for various cases. Tables 1 also indicates average false negative (FN) rate, an abnormal mammogram classified as normal mammogram, of 6.94% and average false positive (FP) rate, a normal mammogram classified as abnormal mammogram, of 0%.

PCA is employed globally to the training data to obtain the principal components where all the principal components are retained. PCA uncorrelates the first few principal components in the transformed data while the rest of the components are still highly correlated. On the other hand, FLD is used as a dimensionality reduction and feature extraction module. FLD is applied, which uses the basis provided by the PCA, to generate another new set of basis for the classification stage. FLD uncorrelates the data again by taking into account the different classes present in the data. This dual transformation into the Eigenspace uncorrelates the data two times thus should greatly improve the classification stage.

TABLE 1
CLASSIFICATION ACCURACY, FP, and FN RATES. EACH SET CONSISTS of 24 NORMAL and 24 ABNORMAL SUB-IMAGES

Test Set	FN	FP	Accuracy
1	6.25%	0%	93.75%
2	8.33%	0%	91.67%
3	6.25%	0%	93.75%

This CAD system uses several parameters that impact the performance and accuracy of results such as the number of selected principal components (PC), number of retained Fisher values, and number of nearest neighbors to assign.

A. Number of Selected Principal Components

Experimental results indicate that selecting all the principal components produces the highest accuracy. Thus, PCA is used in this work to decorrelate the data rather than reducing its dimensionality. Even though most of the information is contained in the first few principal components, discarding the least signifi-

cant principal components may result in loss of information depending on the application.

B. Number of Selected Fisher Values

In this work, FLD algorithm is used for the dimensionality reduction and feature extraction. The experimental results indicate that in this case, selecting 11 Fisher values achieves the highest accuracy. Thus, suggesting that retaining more than one Fishervalue improves the classification stage.

C. Number of Nearest Neighbors

In this work the nearest neighbor, which has already been classified from the training data, is used in making the decision to which class the testing sub-image belong. This value is chosen as it achieves the best results as shown in literature for the two-class problem [5].

D. Implemented Algorithms

Tables 2 and 3 show results of the proposed CAD algorithm against PCA and FLD algorithms for various testing data. The average accuracies of PCA, FLD, and PCA-FLD are 78.47%, 47.92%, and 93.06% where all the principal components are retained and eleven Fisher values are retained. These results indicate that the proposed PCA-FLD algorithm outperforms PCA and FLD algorithms for all the testing sets. The improvements of PCA-FLD algorithm over PCA and FLD algorithms are 18.59% and 94.2% which indicates that performance of the FLD algorithm can be greatly improved if data preprocessed by PCA.

V. CONCLUSIONS

A tool that can be part of a CAD system has been developed and implemented in this paper. The framework is based on integrating PCA, FLD, and KNN classifier. The performance of the proposed tool is compared against the individual performance of PCA and FLD. Extensive simulations using 144 sub-images were performed. The results indicate that combining PCA and FLD algorithms improves PCA algorithm accuracy of 18.59% and FLD algorithm accuracy of 93.06% in all testing sets.

The ability of the proposed framework to correctly classify mammograms depends upon various factors including the proper cropping of images, number of retained principal components, number of retained Fisher values, and number of nearest neighbors taken into consideration. The framework implementation resulted in the highest accuracy when all the principal components and few Fisher values were retained, and only one neighbor is considered. Results also indicate that PCA reduces the computational complexity for the between-class and within-class scatter matrices.

Future work should include testing the CAD algorithm on other mammogram databases. Other biological features can be integrated within the framework to help automating the parameters selection process. The proposed algorithm can be further enhanced by modeling the problem as a multiclass problem through including three classes: normal, malignant, and benign. Malignant regions have well defined boundaries whereas benign regions do not have such a characteristic. This fact can be utilized to improve the classification phase.

REFERENCES

[1] (2010) The National Cancer Institute website. [Online]. Available: www.cancer.gov/cancertopics/types/breast

[2] C.H. Lee, "Screening mammography: proven benefit, continued controversy," Radiologic clinics of North America, vol. 40, pp. 395-407, 2002.

[3] M.L. Giger, Krassemeijer, and S.G. Armato, "Computer-aided diagnosis in medical imaging," IEEE Transactions on Medical Imaging, vol. 20, pp. 1205-1208, 2001.

[4] T. Cooke, "Two variations on fisher linear discriminant for pattern recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24 (2), pp. 268-273, Feb. 2002.

[5] W. Zhao, R. Chellappa, and A. Krishnaswamy, "Discriminant analysis of principal components for face recognition", in Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition, April 1998, 14-16, pp. 336-341.

TABLE 2
COMPARISON BETWEEN PCA and PCA-FLD ALGORITHMS for TEST SETS 1-3

PC	Test Set No.1		Test Set No.2		Test Set No.3	
	PCA	PCA-FLD	PCA	PCA-FLD	PCA	PCA-FLD
11	60.41%	79.16%	55.16%	60.41%	37.5%	41.67%
20	58.33%	68.75%	22.91%	43.75%	39.58%	41.67%
30	43.75%	72.91%	52.08%	54.16%	43.75%	47.91%
40	45.83%	89.58%	58.33%	64.58%	56.25%	56.25%
48	**77.08%**	**93.75%**	**66.67%**	**91.67%**	**91.67%**	**93.75%**

TABLE 3
COMPARISON BETWEEN FLD and PCA-FLD ALGORITHMS for TEST SETS 1-3

Fisher Values	**Test Set No.1**		**Test Set No.2**		**Test Set No.3**	
	FLD	**PCA-FLD**	**FLD**	**PCA-FLD**	**FLD**	**PCA-FLD**
1	56.25%	89.58%	52.08%	91.67%	54.16%	54.16%
4	60.41%	89.58%	56.25%	91.67%	52.08%	89.58%
9	43.75%	89.58%	50%	91.67%	56.25%	93.75%
11	**41.67%**	**93.75%**	**50%**	**91.67%**	**52.08%**	**93.75%**
15	43.75%	91.67%	50%	91.67%	52.08%	72.91%