Application of Data Mining Algorithms for Mammogram Classification

Milos Radovic, Marina Djokovic, Aleksandar Peulic, and Nenad Filipovic, *Member, IEEE*

Abstract—**One of the leading causes of cancer death among women is breast cancer. In our work we aim at proposing a prototype of a medical expert system (based on data mining techniques) that could significantly aid medical experts to detect breast cancer. This paper presents the CAD (computer aided diagnosis) system for the detection of normal and abnormal pattern in the breast. The proposed system consists of four major steps: the image preprocessing, the feature extraction, the feature selection and the classification process that classifies mammogram into normal (without tumor) and abnormal (with tumor) pattern. After removing noise from mammogram using the Discrete Wavelet Transformation (DWT), first is selected the region of interest (ROI). By identifying the boundary of the breast, it is possible to remove any artifact present outside the breast area, such as patient markings. Then, a total of 20 GLCM features are extracted from the ROI, which were used as inputs for classification algorithms. In order to compare the classification results, we used seven different classifiers. Normal breast images and breast image with masses (total 322 images) used as input images in this study are taken from the mini-MIAS database.**

I. Introduction

MASS diseases, such as cancer, are the leading cause of death worldwide [1]. Breast cancer is the most common cancer among women in the world. Breast cancer screening with mammography has been shown to be effective for preventing breast cancer death.

An important development that may help to improve the performance in breast cancer screening is computer aided diagnosis (CAD). It is hoped that CAD can help to decrease the number of errors. Software can help searching for suspicious signs, or could help classifying lesions in benign or malignant types. CAD system consists of several modules, such as preprocessing, segmentation and classification of pathological cases. The medical image classification procedure usually consists of three steps: (1) Texture Feature Extraction, (2) Feature Selection and (3) Classification. Texture feature have been widely used to classify normal and abnormal pattern in digital mammogram. In this paper, Co-occurrence matrix is used for texture features extraction. By using feature selection we can identify and remove irrelevant or redundant features. For classification we use seven different classifiers and compare results by using leave one out cross validation procedure.

II. Image Preprocessing

A. Image Denoising

An image is often corrupted by noise during its acquisition or transmission. The denoising process is to remove the noise while retaining and not distorting the quality of the processed image. The traditional way of image de-noising is filtering. These methods are mainly based on thresholding the Discrete Wavelet Transform (DWT) coefficients [2].

The thresholding techniques are simple non-linear techniques that eliminate all the subband coefficients that their magnitude is under a certain threshold. The type of the threshold is either hard (1) or soft (2). The reconstruction of the "clean" image, after the thresholding process, is performed with the inverse wavelet transform.

$$\text{Hard threshold:} \quad \begin{cases} y = x & if \quad |x| > T \\ y = 0 & if \quad |x| < T \end{cases} \tag{1}$$

$$\text{Soft threshold:} \quad y = sign(x)(|x| - T) \tag{2}$$

where x is the input signal, y is the signal after threshold and T is the threshold level.

For the image denoising Haar wavelet is used. The Haar wavelet's mother wavelet function ψ(t) and its scaling function φ(t) can be described as:

$$\psi(t) = \begin{cases} 1 & 0 \le t < 1/2 \\ -1 & 1/2 \le t < 1 \\ 0 & otherwise \end{cases}, \quad \varphi(t) = \begin{cases} 1 & 0 \le t < 1 \\ 0 & otherwise \end{cases} \tag{3}$$

Image denoising is executed by applying one-level wavelet decomposition, wavelet Haar, and the threshold of 50. The used thresholding technique is soft thresholding.

B. Region of Interest Extraction

In the CAD environment, one of the roles of image processing would be to detect the region of interest (ROI) for a given, specific, screening or diagnostic application. The

Manuscript received July 30th, 2013. This work was supported in part by the European Commission (Project SIFEM, ICT-600933) and the Ministry of Education, Science and Technological Development of Republic of Serbia through Grant ON174028 and III41007.

M. Radovic, is with the Research and Development Center for Bioengineering "BioIRC", Kragujevac, Serbia (phone: 381-34-301920; fax: 381-34-500089; e-mail: mradovic@kg.ac.rs).

M. Djokovic and A. Peulic, are with the University of Kragujevac, Serbia (e-mail: marina.m.djokovic@gmail.com, aleksandar.peulic@gmail.com).

N. Filipovic, is with the Faculty of Mechanical Engineering, Kragujevac, Serbia and Research and Development Center for Bioengineering "BioIRC", Kragujevac, Serbia (e-mail: fica@kg.ac.rs).

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

method for detection ROI, summarized in Fig. 1, is composed of several main steps, as described in the following sections.

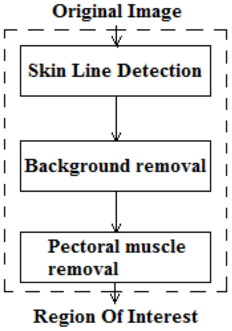

Fig. 1. Flowchart of the procedure for Region Of Interest extraction..

Skin Line detection. Identification of the breast boundary is very important. By identifying the boundary of the breast, it is possible to remove any artifact present outside the breast area, which can affect the performance of image analysis and pattern recognition techniques [4]. Following notations are used to describe algorithm for skin line detection:

I - the original mammographic image, *I(i,j)* – pixel value in the *i*-th row and *j*-th column of image *I*, *B1* and *B2* are the binary versions of original image with different threshold, *B1(i,j)* and *B2(i,j)* – pixel value in the *i*-th row and *j*-th column of image *B1* and *B2*.

```
Algorithm:
i=j
If  I(i,j)>5
        Then B1(i,j)=1
        Else B1(i,j)=0
If  I(i,j)>20
        Then B2(i,j)=1
        Else B2(i,j)=0
Skin line = B1-B2
```

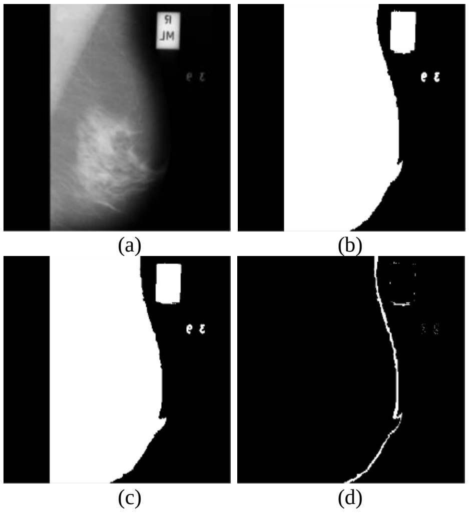

Fig. 2. (a) Original image, (b) Binary version with less threshold value, (c) Binary version with greater threshold value, (d) Detected skin line.

Background portion removal. Background removal procedure will be explained for the Left MLO (LMLO) view of mammogram (procedure for the Right MLO (RMLO) view is very similar so, it is not necessary to explain both). To remove breast background, it is necessary to create and apply a mask. Mask is created by applying algorithm which is explained below to image with detected skin line, shown in Fig. 2(d).

```
Algorithm for the mask formation:
Step 1: Start with first row.
Step 2: Scan from left to right side.
Step 3: If pixel is black then replace it with white and move to next pixel
        and repeat Step 3.
Else, when pixel is white then go to Step 4.
Step 4: Move to next pixel while pixel is white.
When pixel is black then go to Step 5.
Step 5: Replace current pixel with black and move to next pixel.
If current column is the last column go to step 6, else repeat step 5.
Step 6: Repeat step 2 to 5 for next row.
```

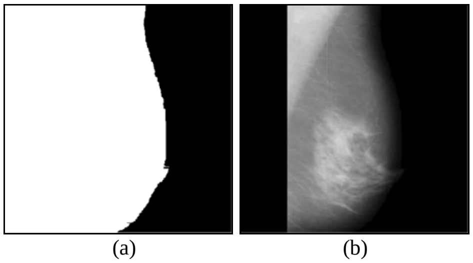

Fig. 3. (a) Mask, (b) Original mammogram with removed background.

Pectoral muscle removal. Pectoral muscle tissue is usually denser than the rest of the breast. Therefore, pectoral muscle and a central part of the breast will be extracted by applying Local threshold operation with threshold value 155. Fig. 4(a) shows binary version of extracted pectoral muscle and a central part of the breast. Result of multiplication this binary image and original mammogram with removed background is shown in Fig. 4(b).For the LMLO type of mammogram, in order to separate pectoral muscle from the central part of image shown in Fig. 4(b), it is first necessary to remove pectoral muscle from the image. This procedure is explained below.

```
Algorithm for the central tissue extraction:
Step 1: Start with first row and n-th column, n is the first non-zero pixel
        in first row.
Step 2: Scan from left to right side.
Step 3: If pixel is non-zero then replace it with zero and move to next
        pixel and repeat Step 3.
Else, when pixel is zero then start from the next row and n-th column
      and repeat Step 3.
Stop the procedure when all rows are exhausted.
```

Extracted central tissue is shown in Fig. 4(c). Then, the pectoral muscle is isolated by subtracting image in Fig. 4(c) from the image in Fig. 4(b). Fig. 4(d) shows the isolated pectoral muscle. Finally, region of interest, mammogram without background and pectoral muscle, is obtained by subtracting image with isolated pectoral muscle from the original mammogram with removed background. Extracted ROI is shown in Fig. 4(e).

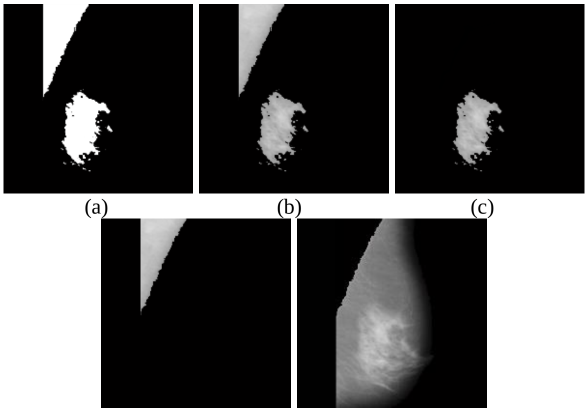

Fig. 4. (a) Binary version of mammogram after local thresholding, (a) Original mammogram after local thresholding, (c) Extracted central tissue, (d) Isolated pectoral muscle, (e) ROI.

III. Methods

The basic idea of procedure for automatic medical image classification, when applied to images, consists of three steps: (1) Texture Feature Extraction, (2) Feature Selection and (3) Classification. This procedure is shown in Fig. 5.

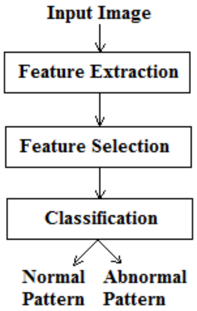

Fig. 5. Proposed methodology for mammographic image classification.

Features extraction is the important step in breast cancer detection. Some of the most commonly used texture measures are derived from the Grey Level Co-occurrence Matrix (GLCM). The gray-level co-occurrence matrix is a statistical method of examining texture that considers the spatial relationship of pixels. The GLCM functions characterize the texture of an image by calculating how often pairs of pixel with specific values and in a specified spatial relationship occur in an image, creating a GLCM, and then extracting statistical measures from this matrix. In this paper, Co-occurrence matrix is used for features extraction. GLCM is calculated in one angle (0˚) and distance value d=1. The 20 descriptors are extracted from GLCM texture measurement, features f1-f13 are features proposed by Haralick [5], Soh proposed features f14-f18 [7] and features f19 and f20 are proposed by Clausi [6].

TABLE I
20 Features Extracted from GLCM

Feature No.	Feature name
f1	Angular Second Moment (Energy)
f2	Contrast
f3	Correlation
f4	Sum of Squares: Variance
f5	Inverse Difference Moment (Homogeneity)
f6	Sum Average
f7	Sum Variance
f8	Sum Entropy
f9	Entropy
f10	Difference Variance
f11	Difference Entropy
f12	Information Measure of Correlation 1
f13	Information Measure of Correlation 2
f14	Autocorrelation
f15	Dissimilarity
f16	Cluster Shade
f17	Cluster Prominence
f18	Maximum Probability
f19	Inverse Difference Normalized
f20	Inverse Difference Moment Normalized

Calculated features are given as input to different classifiers. For classification process we used seven different classifiers, support vector machine, naive bayes classifier, k-nearest neighbor, logistic regression, decision trees, random forest and neural network, then results are compared. The goal is to create data mining model which will be able to accurately classify mammogram into normal (without masses) or abnormal (with masses).

First, we calculated accuracy of these seven data mining algorithm by using all 20 features as input. Then, we did MRMR (Minimum redundancy maximum relevance) feature selection [8] for selection of the five most relevant features. By using only these five features we were able to obtain greater classification accuracy. All models have been tested by using leave-one-out cross validation procedure.

Totally 322 images, taken from mini-MIAS database, have been used for classification. Seven different data mining algorithms have been used to model relationship between GLCM features and mammogram class (normal or abnormal). Models have been tested by using leave-one-out cross validation procedure. In order to compare classification results of a different data mining algorithms we calculated accuracy:

$$AC = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

where, TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives and FN is the number of false negatives.

IV. Results

Table 2 summarizes classification results (accuracy and ROC values) when using all 20 features as input.

TABLE II
CLASSIFICATION RESULTS WHEN USING ALL 20 GLCM FEATURES

	ACCURACY	**ROC**
Naive Bayes	70.6667	0.766
Logistic regression	74	0.825
SVM	72	0.72
KNN	68	0.73
C4.5	74	0.791
Random forest	70.6667	0.783
MLP	**76**	0.788

Table 2 shows that multilayer perceptron gave the best accuracy (76%). It has 11 neurons in a single hidden layer and sigmoid activation functions in all neurons. Learning was performed using the backpropagation algorithm with momentum (momentum constant 0.2) [9]. The stopping criterion was defined as a maximum number or learning epochs (1000).

In order to improve classification accuracy we extracted 5 most relevant features by using MRMR algorithm – *Sum Average*, *Contrast*, *Sum of Squares: Variance*, *Cluster Prominence* and *Autocorrelation*. This algorithm tends to select features which are most relevant to the class and have the least correlation between themself. By using only top 5 selected features we were able to obtain even greater accuracy (Table 3). Table 3 shows that the greatest accuracy (79.33%) is achieved with C4.5 decision trees algorithm [10]. The attractiveness of decision trees algorithm is due to the fact that, in contrast to neural networks and some other data mining algorithms, decision trees represent rules, which can readily be expressed so that humans can understand them.

TABLE III
CLASSIFICATION RESULTS WHEN USING 5 SELECTED GLCM FEATURES

	ACCURACY	**ROC**
Naive Bayes	73.33	0.829
Logistic regression	74	0.833
SVM	72	0.72
KNN	74.67	0.834
C4.5	**79.33**	0.811
Random forest	73.33	0.795
MLP	69.33	0.713

V. CONCLUSION

This paper shows that advanced techniques of image processing and data mining are useful in computer aided diagnosis. The methods like one presented in this paper could assist the radiologist and improve the accuracy of detection. Classification is done based on textural descriptors obtained from features extraction process. Results show that few of proposed data mining algorithms are able to deal with the problem of mammogram classification. This approach has potential for further development because of its simplicity that will motivate real-time breast cancer diagnosis in providing a second opinion to radiologists.

REFERENCES

[1] B. Novakovic, J. Jovicic, N. Milic, F. Jusupovic, M. Grujicic, and D. Djuric, “Nutrition care process in cancer,” *HealthMED Journal*, vol. 4, no. 2, pp. 427-433, 2010.

[2] J.N. Ellinas, T. Mandadelis, A. Tzortzis, and L. Aslanoglou, “Image de-noising using wavelets,” *T.E.I. of Piraeus Applied Research Review*, vol. 9, no. 1, pp. 97-109, 2004.

[3] R. Rangarajan, R. Venkataramanan, and S. Shah, *Image Denoising Using Wavelets*. Technical report, College of Engineering, University of Michigan, 2002.

[4] J.S. Suri, and A. Farag, *Deformable models: biomedical and clinical applications*. New York, USA: Springer Science+Business Media, LLC, 2007.

[5] R.M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” *IEEE Transactions on systems, man and cybernetics*, vol. 3, no. 6, pp. 610-621, 1973.

[6] D.A. Clausi, “An analysis of co-occurrence texture statistics as a function of grey level quantization,” *Canadian Journal of Remote Sensing*, vol. 28, no. 1, pp. 45–62, 2002.

[7] L.K. Soh, and C. Tsatsoulis, “Texture Analysis of SAR Sea Ice Imagery Using Gray Level Co-Occurrence Matrices,” *IEEE Transactions on geoscience and remote sensing*, vol. 37, no. 2, pp. 780-795, 1999.

[8] C. Ding, and H. Peng, “Minimum redundancy feature selection from microarray gene expression data,” *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 2, pp. 185-205, 2005.

[9] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Prentice Hall, New Jersey, USA, 1999.

[10] R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.