Implementation of Practical Computer Aided Diagnosis System for Classification of Masses in Digital Mammograms

Mohamed E. Elmanna
Biomedical Engineering Department
Cairo University
Giza, Egypt
mohamed_elmanna@hotmail.com

Yasser M. Kadah
Electrical and Computer Engineering Department
King Abdulaziz University
Jeddah, Saudi Arabia
ykadah@kau.edu.sa

Abstract— **Breast cancer is the most common cancer in women worldwide. It is also the principle cause of death from cancer among women globally. Mammogram image is considered as the most reliable, low cost, and highly sensitive technique for detecting small lesions. Computer-aided diagnosis system (CAD) can be very helpful for radiologist in detection and diagnosing abnormalities earlier and faster than traditional screening programs. In this work a CAD system to distinguish between masses and normal breast tissue was proposed. We started our system by using DDSM database for mammogram images which were first preprocessed using image enhancement algorithm, then 100 regions of interest (ROIs) containing masses and normal breast tissue are extracted. Then we extracted a group of 59 texture and statistical features from the ROIs. Then we performed feature selection using Sequential Forward Selection and Sequential Floating Forward Selection. Finally we used K-Nearest Neighbor (KNN) classifier, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Support Vector Machine (SVM) classifier for classification with leave-one-out method for testing. The obtained results show acceptable sensitivity and specificity for the system.**

Index Terms— **Computer Aided Diagnosis, Mass classification, Peripheral enhancement.**

I. INTRODUCTION

Breast cancer is the most common cancer in women worldwide. It is also the principle cause of death from cancer among women globally. In Egypt, where according to latest statistics, breast cancer accounts for 37.5 percent of all cancer cases. The five year survival rate can be increased from 60% to 82% by an early diagnosis of breast cancer. So, during the last years, screening programs became essential step for women over 40 years old. Therefore, physicians have to examine a huge number of images leading to 10-30% of missed breast lesions. Computer aided tools have been shown to be powerful systems to overcome this problem, the reader's sensitivity can be increased by an average of 10% with the assistance of CAD systems.

Mammography has been successful in improving detection of cancer, particularly non-palpable breast masses and calcifications that may be malignant. There has been some

recent controversy over the benefit of mammography screening and the available evidence relating mammography screening with mortality may not be definitive. Nonetheless, a recent Institute of Medicine Report on Mammography (Committee on the Early Detection of Breast Cancer 2001) suggests that the reduction in mortality from breast cancer observed in recent years may be due to earlier detection through mammography screening [1]. However, mammography is not perfect. Detection of suspicious abnormalities is a repetitive and fatiguing task. For every thousand cases analyzed by a radiologist, only 3 to 4 are cancerous and thus an abnormality may be overlooked. As a result, radiologists fail to detect 10-30% of cancers [2]. It has been suggested that double reading i.e., independent mammogram interpretation by two radiologists, may increase the sensitivity and specificity of mammographic screening by 10% to 15 % [3]. However, the rise in costs in addition to the increased workload on the radiologists does not make double reading a cost-effective option.

By incorporating the expert knowledge of radiologists, the computer-based systems provide a second opinion in detecting abnormalities and making diagnostic decisions. Such a diagnostic procedure is called computer-aided diagnosis (CAD). A computerized system for such a purpose is called a CAD system. It has been shown that the performance of radiologists can be increased by providing them with the results of a CAD system [4]. Hence, there are strong motivations to develop a CAD system to assist radiologists in reading mammograms.

Several research groups have developed CAD programs for the detection and classification of breast abnormalities. Sahiner et al. [5] investigated the classification of regions of interest (ROI's) on mammograms as either mass or normal tissue using a convolution neural network (CNN).they employed texture feature extraction methods applied to small sub-regions inside the ROI. Receiver operating characteristic (ROC) methodology was used to evaluate the classification accuracy. Wei et al. [6] investigated the feasibility of using multiresolution texture analysis for differentiation of masses from normal breast tissue on mammograms. The wavelet transform was used to decompose regions of interest (ROIs)

on digitized mammograms into several scales. They also used Stepwise linear discriminant analysis to select optimal features and linear discriminant classifier. Oliver et al. [7] proposed a method for reducing false positives in breast mass detection. Their approach is based on using the Two-Dimensional Principal Component Analysis (2DPCA) algorithm in order to extract features. The classifier used, is a combination of the decision tree and the k-Nearest Neighbor algorithm. They used a leave-one-out scheme and Receiver Operating Characteristics (ROC) analysis for the evaluation. Youssry et al. [8] proposed a neuro-fuzzy model for fast detection of candidate circumscribed masses in digitized mammograms. They extracted texture features from sub-image co-occurrence matrices in different orientations. Then they used the features to train neuro-fuzzy models. In spite of the success of such methods to present addition to the field, their practical implementation including all steps of preprocessing and their effect on the classification accuracy were not clear. Therefore, a complete system implementation that takes into account all the practical steps that must be performed in order to reach a diagnosis would be desirable.

In this work, we describe the detailed methodology to develop a practical CAD system for the detection and classification of masses in digital mammograms. The proposed system consists of four major steps. The first step is preprocessing step where we applied image enhancement algorithm to enhance the peripheral region of the breast and perform density compensation. The second step is the feature extraction, where we used a set of 59 features. The third step is the feature selection step, where we used two techniques: Sequential Forward Selection (SFS) and Sequential Floating Forward Selection (SFFS). The last process is the classification process where we used k-Nearest Neighbor (k-NN) classifier, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Support Vector Machine (SVM) classifiers to discriminate between normal and cancerous tissues. The entire procedure of system development is described in this paper and a block diagram of the system is presented in Fig. 1.

II. METHODOLOGY

A. Digital Mammography database

The data used in this work was taken from the University of South Florida digital database for screening mammography [9]. All images which we used are digitized using LUMISYS
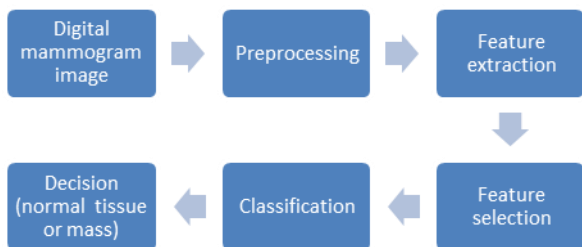
Scanner at a resolution 50 microns and at 12 bit grayscale level. Each abnormal view has a text overlay file (ground truth) which describes abnormalities present as marked by an expert radiologist. The actual abnormality location and boundary in each image are defined by a chain-code. the images of this database were down-sampled to 0.25 of the original images using Nearest-neighbor interpolation to reduce the size of the data because texture features computed with pixel size of 200 μm will have good classification accuracy according to Rangayyan study [10]. 100 ROI are extracted manually using window of size 32×32 pixels, 50 are abnormal ROI (circumscribed masses, speculated masses, ill-defined masses and architectural distortion) and 50 are normal ROI. Figure 2 shows mammogram with mass region defined by chain code.

B. Preprocessing

The preprocessing is the first step in the CAD system. Here, an image processing algorithm is used for image enhancement. We applied peripheral enhancement for mammograms in the uncompressed tissue region near the projected skin–air interface. This technique is done by Tao Wu et al [11].

In peripheral enhancement methods, the darkening due to decreased tissue thickness in the peripheral area is estimated from the mammogram and thereafter compensated for by a smoothly varying correction function. After correction, fatty tissues in the interior and peripheral regions have similar gray level values. With peripheral enhancement, the dynamic range of the mammogram greatly reduces, and as a consequence, less manual adjustments of contrast settings are required to view details close to the skin line [12]. Figure 3 shows an example for the process of peripheral enhancement performed on our system.
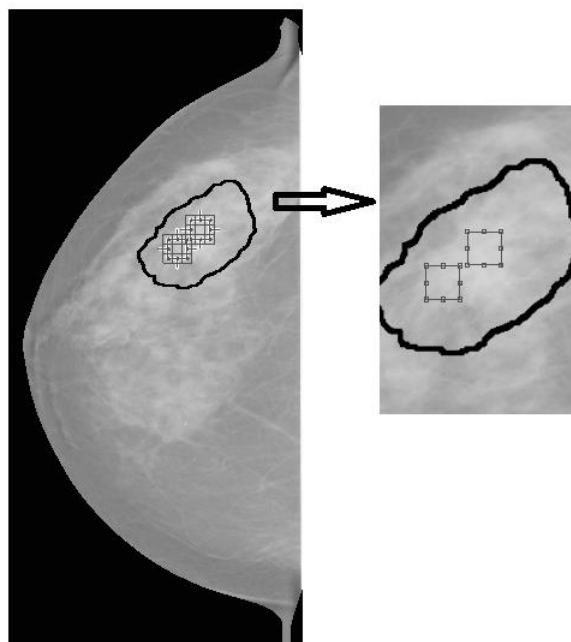

Fig. 2. Digital Mammogram with defined mass boundary. It is the case C_0001_1.RIGHT_CC in DDSM database with mass boundary defined by chain code.

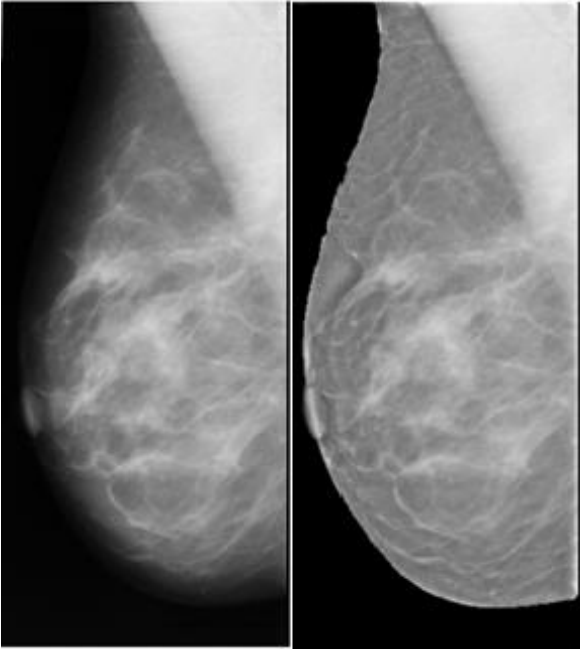Fig. 1. a schematic diagram for the CAD system.

Fig. 3. Example of a corrected mammogram from our system. On the left side the original mammogram, on the right side the mammogram after applying the peripheral enhancement algorithm.

C. Feature extraction

The feature extraction step is one of the most important factors that affect the CAD performance. Features are used to describe the character of an object. The extracted features represent a mathematical description of characteristics that are helpful for isolating the lesions or for distinguishing normal and abnormal lesions. This is an important step in most pattern-analysis tasks. an artificial system can identify suspicious area and make a final decision based on certain features of the mass. Unlike much more complicated process of a human observer to identify a mass; the machine observers make decisions with limited features.

In this work we used a set of 59 features used by A. Cao et al. [13], B Acha et al. [14], Songyang Yu et al. [15] and P Zhang et al. [16]. These features include: Energy, Entropy, Standard Deviation, Skewness, Modified Energy, Modified Entropy, Modified Standard Deviation, Modified Skewness, tail ratio, inter-distance, average of the mean slope, average of maximum slope, average height, Correlation, Contrast, Mean of gray level, Variance of gray levels, mean gradient, Variance of gradient, and Dynamic range. The previously mentioned features are measured from the ROI directly.

Other features group extracted from wavelet decomposition include: Contrast, Correlation, Energy, Homogeneity, Inverse Different Moment, Variance, Sum Average, Sum Entropy, Sum Variance, Difference Entropy and seven Invariant Moment features. This group is computed two times from two levels of the wavelet decomposition.

At the beginning the wavelet decomposition was applied on the region of interest using the wavelet Daubechues (db1), each mammogram image is decomposed up to four levels using the separable 2-D wavelet transform. These features are measured from level two and three of the wavelet decomposition. All features (except invariant moment features) are measured from gray level co-occurrence matrix

which is computed for level two and three of the wavelet decomposition.

The final feature group is Contrast, Correlation, Energy, Homogeneity and Entropy. These features are calculated from the gray-level co-occurrence matrix. The co-occurrence matrix is taken in the east direction at pixel spacing of 1.

D. Feature selection

Feature selection is an important part of any classification scheme. The success of a classification scheme largely depends on the features selected and the extent of their role in the model. Only a few features may be useful or 'optimal' while most may contain irrelevant or redundant information that may result in the degradation of the classifier's performance.

In this work we used sequential forward selection (SFS) and Sequential floating forward selection (SFFS) for feature selection. A Matlab toolbox for pattern recognition (PRTools4 [17]) will be used to perform the feature selection process. Sequential forward selection (SFS, or the method of set addition) introduced by Whitney [18] which is a bottom-up search procedure that adds new features to a feature set one at a time until the final feature set is reached. Suppose we have a set of d1 features, Xd1. For each of the features ξ_j not yet selected (i.e. in $\chi - X_{d1}$) the criterion function $J_j = J(X_{d1} + \xi_j)$ is evaluated. The feature that yields the maximum value of J_j is chosen as the one that is added to the set X_{d1}. Thus, at each stage, the variable is chosen that, when added to the current set, maximizes the selection criterion. The feature set is initialized to the null set. When the best improvement makes the feature set worse, or when the maximum allowable number of features is reached, the algorithm terminates. The main disadvantage of the method is the nesting effect. This means that a feature that is included in some step of the iterative process cannot be excluded in a later step. Thus, the results are sub-optimal [19]. Alternatively, the Sequential Floating Forward Selection (SFFS) method was introduced by Pudil et al. [20] to deal with the nesting problem and is likely to provide better results. In this work the evaluation function or the criterion function used for SFS and SFFS are 1-Nearest Neighbor leave-one-out classification performance.

E. Classification

Classification is the process of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. The classification process is divided into the learning phase and the testing phase. In the learning phase, known data are given and the feature parameters are calculated by the processing which precedes classification. Separately, the data on a candidate region which has already been decided as a tumor or as normal are given, and the classifier is trained. In the testing phase, unknown data are given and the classification is performed using the classifier after learning. We used Four Classifiers for the CAD system, The Voting K-Nearest Neighbor (K-NN) Classifier, the Linear Discriminant Analysis (LDA) classifier, the Quadratic Discriminant Analysis (QDA) classifier, and the support vector machine (SVM) classifier. We also used A Matlab toolbox for pattern recognition

(PRTools4 [17]) to perform the classification for LDA and QDA classifiers.

1) k-nearest neighbor (KNN)

k-nearest neighbor (K-NN) classifier distinguishes unknown patterns based on the similarity to known samples. The K-NN algorithm computes the distances from an unknown patterns to every sample and select the K-nearest samples as the base for classification. The unknown pattern is assigned to the class containing the most samples among the K-nearest samples [21]. The Majority rule used to decide how to classify the sample is nearest point tie-break and the distance metric used is Euclidean distance.

2) Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a classification method originally developed in 1936 by R. A. Fisher. It is simple, mathematically robust and often produces models whose accuracy is as good as more complex methods. LDA is used to find the linear combination of features which best separate two or more classes of objects or events. LDA assumes that the different classes have the same covariance matrix.

3) Quadratic Discriminant Analysis (QDA)

Quadratic Discriminant Analysis, aims to find the quadratic combination of features. It is more general than linear discriminant analysis. Unlike LDA, QDA does not make the assumption that the different classes have the same covariance matrix. Instead, QDA makes the assumption that each class k has its own covariance matrix.

4) Support Vector Machines (SVM)

Support vector machines represent an extension to nonlinear models of the generalized portrait algorithm developed by Vapnik and Lerner. The SVM algorithm is based on the statistical learning theory which describes the properties of learning machines that allow them to give reliable predictions. SVM performs an implicit mapping of data into a higher dimensional feature space, where linear algebra and geometry can be used to separate data [22]. For our specific problem, Value of the box constraint C=1, and SVM with a linear kernel is used to classify between the different classes.

III. RESULTS & DISCUSSION

We used a set of 100 mammograms for classification stage. Fifty of the ROIs are known to be masses while the remaining are known to be normal tissues. We measured quantitatively the detection performance of the classifiers by computing the sensitivity and specificity of the data and drawing Receiver Operating Characteristic (ROC) curves.

Sensitivity: Measures how well the algorithm can identify abnormal samples.

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

Specificity: Measures how well the algorithm identifies normal samples.

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

True Positive (TP): account of all samples which are correctly called by the algorithm as being abnormal.
True Negative (TN): account of all samples which are correctly called by the algorithm as being normal.

False Positive (FP): account of all samples which are incorrectly called by the algorithm as being abnormal while they are normal.
False Negative (FN): account of all samples which are incorrectly called by the algorithm as being normal while they are abnormal.

In the feature selection stage, 14 features are selected using sequential forward selection, and 17 features are selected using sequential floating forward selection. Table 1 shows the selected features ranked according to selection. The results of K-nearest neighbor classifier (K-NN) , linear discriminant analysis (LDA) ,Quadratic discriminant analysis (QDA) and Support Vector Machine classifier (SVM) is presented in table 2 for feature selection using Sequential forward Selection (SFS) and presented in table 3 for feature selection using Sequential floating forward Selection (SFFS). Also Figure 4 shows Receiver operating characteristic (ROC) curves for the classifiers. Results show that for the training step, the K-NN classifier with K= 1 is better than other Classifiers in all feature selection techniques (sensitivity = 1 , specificity = 1), Then K-NN classifier with K=3 in all feature selection strategies give the second best result (sensitivity = 0.96 , specificity = 0.98). For the testing, the KNN classifier (k=1) using SFFS gives the best result (sensitivity = 0.94, specificity = 0.98), then KNN classifier (k=1) using SFS is the second one (sensitivity = 0.96, specificity = 0.94), then KNN classifier using SFFS gives (sensitivity = 0.88, specificity = 0.94). For the testing set, in KNN classifier, (k=1) has the best result (accuracy= 0.95 for SFS and accuracy=0.96 for SFFS), then k=3 gives better results than K=5, 7 (accuracy=0.90 for SFS and accuracy=0.91 for SFFS). For the testing set, SVM classifier using SFFS gives better result (accuracy=0.89) than LDA, QDA, and KNN (K=5,7). For the testing set, when we compare between LDA and QDA classifiers we can see that QDA using SFS gives the best result (accuracy=0.88) , then LDA using SFFS gives (accuracy=0.87). KNN classifier using (k=1) is the superior as a result of using 1-neareset neighbor classifier for the evaluation function of SFS and SFFS.

TABLE I. FEATURES SELECTED BY FEATURE SELECTION STAGE USING SFS AND SFFS.

SFS features	SFFS features
1. Mean of gray level	1. Mean of gray level
2. Entropy (from ROI directly)	2. Correlation (from level 2)
3. Invariant moment ($\emptyset_4$ from level 3)	3. Variance of gradient
4. Correlation (from level 2)	4. Entropy (from ROI directly)
5. Modified energy	5. Average height
6. Modified skewness	6. Sum average (from level 2)
7. Modified standard deviation	7. Invariant moment ($\emptyset_5$ from level 2)
8. Energy (from ROI directly)	8. Correlation (from GLCM)
9. Variance of gradient	9. Invariant moment ($\emptyset_6$ from level 3)
10. Skewness	10. Correlation (from level 3)
11. Homogeneity (from level 2)	11. Standard deviation
12. Invariant moment ($\emptyset_6$ from level 2)	12. Inverse different moment (from level 3)
13. Energy (from GLCM)	13. Variance (from level 3)
14. Modified energy	14. Sum entropy (from level 2)
	15. Contrast (from level 2)
	16. Contrast (from level 3)
	17. Dynamic range

TABLE II. Classification results using Sequential forward Selection (SFS) in terms of sensitivity and specificity.

	Sequential Forward Selection				
	Train		Test		AUC
classifier	Sensitivity	Specificity	Sensitivity	Specificity	
KNN(k=1)	1	1	0.96	0.94	0.9482
KNN(k=3)	0.96	0.98	0.92	0.88	0.9338
KNN(k=5)	0.96	0.92	0.92	0.84	0.9053
KNN(k=7)	0.96	0.88	0.9	0.84	0.8891
LDA	0.92	0.88	0.88	0.84	0.8753
QDA	0.86	1	0.86	0.9	0.8759
SVM	0.92	0.9	0.9	0.86	0.9111

TABLE III. Classification results using Sequential Floating Forward Selection (SFFS) in terms of sensitivity and specificity.

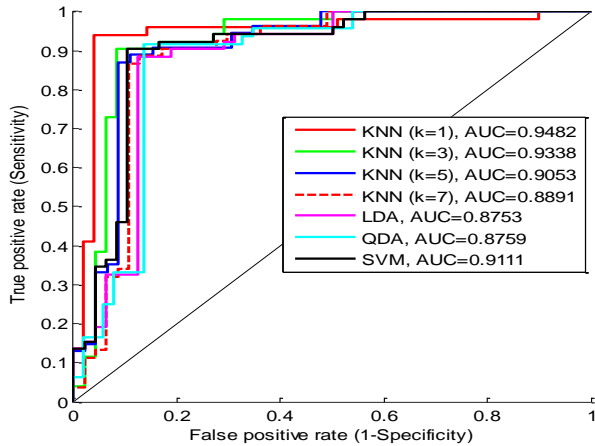|	Sequential Floating Forward Selection				
	Train		Test		AUC
classifier	Sensitivity	Specificity	Sensitivity	Specificity	
KNN(k=1)	1	1	0.94	0.98	0.9536
KNN(k=3)	0.96	0.98	0.88	0.94	0.9093
KNN(k=5)	0.94	0.94	0.84	0.9	0.8640
KNN(k=7)	0.9	0.94	0.82	0.92	0.8574
LDA	0.96	0.92	0.92	0.82	0.8971
QDA	0.9	1	0.86	0.82	0.8493
SVM	0.96	0.92	0.92	0.86	0.9421

Fig. 4. ROC curves for the classifiers performance. Results for the classifiers using SFS in feature selection.
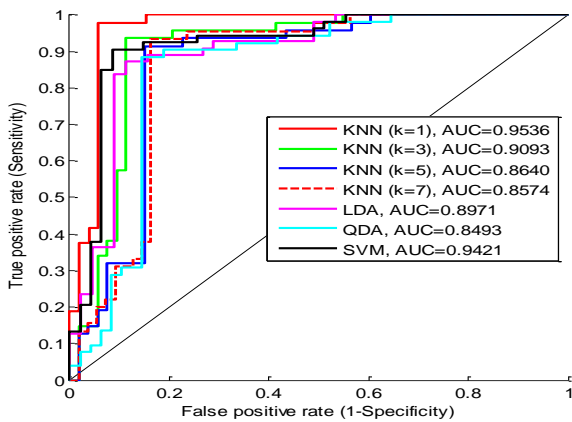

Fig. 5. ROC curves for the classifiers performance. Results for the classifiers using SFFS in feature selection.

IV. CONCLUSIONS

In this work, a practical CAD system for detection and classification of masses is presented. The proposed system consists of four major steps. The first step is preprocessing step where we applied image enhancement algorithm to enhance the peripheral region of the breast and perform density compensation. The second step is the feature extraction, where we used a set of 59 texture and statistical features. The third step is the feature selection step, where we used two techniques: Sequential Forward Selection (SFS) and Sequential Floating Forward Selection (SFFS). The last process is the classification process where we used k-Nearest Neighbor (k-NN) classifier, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Support Vector Machine (SVM) classifiers to discriminate between normal and cancerous tissues. The entire procedure of system development is described in this paper and the results show very promising potential for practical application where the KNN classifier (k=1) using SFFS for feature selection provides excellent sensitivity = 0.94 while maintaining a specificity of 0.98. by comparing between the feature selection techniques, the results show that SFFS technique can choose the most powerful features better that SFS. In the classification step, KNN classifier gives the best performance among all other classifiers.

REFERENCES

[1] Technology evaluation center, Computer-Aided Detection (CAD) in Mammography, Assessment Program Volume 17, No. 17 December 2002.

[2] M. Sampat, M. Markey, A. Bovik, "Computer-aided detection and diagnosis in mammography", in Handbook of Image and Video Processing(ed. Bovik), 2nd edition 2005, pgs. 1195-1217.

[3] C. Vyborny, M. Giger, and R. Nishikawa, "Computer-aided detection and diagnosis of breast cancer", Radiologic Clinics of North America 38(4): 725-740, 2000.

[4] S. Yu, L. Guan, "A Cad System For The Automatic Detection Of Clustered Microcalcifications In Digitized Mammogram Films", IEEE Transactions On Medical Imaging, Vol. 19, No. 2, February 2000.

[5] B. Sahiner et al. , "Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images", Medical Imaging, IEEE Transactions on, vol. 15, pp. 598-610, 1996.

[6] D. Wei et al. , "Classification of mass and normal breast tissue on digital mammograms: multiresolution texture analysis", Medical Physics., vol. 22, pp. 1501-13, 1995.

[7] A. Oliver, X. Lladó, J. Martí, R. Martí, and J. Freixenet, "False positive reduction in breast mass detection using two-dimensional PCA", In: Lect. Not. in Comp. Sc. , vol. 4478, pp. 154–161, 2007.

[8] N. Youssry, F. Abou-Chadi, and A. El-Sayad, "Early detection of masses in digitized mammograms using texture features and neuro-fuzzy model", 4th Annual IEEE Conf on Information Technology Applications in Biomedicine, 2003.

[9] The Digital Database for Screening Mammography [online]. Available http://marathon.csee.usf.edu/Mammography/Database.html

[10] R. Rangayyan, T. Nguyen, F. Ayres, A.Nandi, " the effect of pixel resolution on texture features of breast masses in mammograms", *Journal of Digital Imaging*, Springer, 2010 October; 23(5): 547–553.

[11] T. Wu, R. Moore, and D. Kopans, "Multi-threshold peripheral equalization method and apparatus for digital mammography and breast tomosynthesis", US Patent 7,764,820, Google Patents, 2010.

[12] N. Karssemeijer, P. Snoeren, "Image Processing", Digital Mammography, Springer, pp 69-83, 2010.

[13] A. Cao, Q. Song, X. Yang, and Z. Wang, "mammographic mass detection by robust learning algorithms", Recent advances in breast imaging, mammography, and computer-aided diagnosis of breast cancer, JS Suri, RM Rangayyan, 2006.

[14] B. Acha, C. Serrano, R. Rangayyan, and J. Desautels, "detection of microcalcifications in mammograms", Recent advances in breast imaging, mammography, and computer-aided diagnosis of breast cancer, JS Suri, RM Rangayyan, 2006.

[15] S. Yu, L. Guan, "A CAD system for the automatic detection of clustered microcalcifications in digitized mammogram films", IEEE Transactions on Medical Imaging, 2000.

[16] P. Zhang, B. Verma, and K. Kumar, "Neural vs. statistical classifier in conjunction with genetic algorithm based feature selection", Elsevier Pattern Recognition Letters, 2005.

[17] R. Duin et al., "A matlab toolbox for pattern recognition", Delft University of Technology, 2004.

[18] A. Whitney, "A Direct Method of Nonparametric Measurement Selection", IEEE Transactions in Computers, 1100—1103, 1971.

[19] A. Webb, "Statistical Pattern Recognition", Second Edition.

[20] P. Pudil, J. Novovicova, J. Kittler, "Floating Search Methods in Feature Selection", Pattern Recognition Letters 15,1119—1125, 1994.

[21] A. Omara, A. Mohamed, A. Youssef, and Y. Kadah, "Computer Aided Diagnosis in Digital Mammography", the third Cairo International Biomedical Engineering Conference, CIBEC '06, 2006.

[22] C. Miquel, "Texture Descriptors applied to Digital Mammography", Msc. thesis, de University of Girona, Spain, 2009.