An automated method for the recognition of density of mamary tissue in digital mammographic images

B. Luna-Benoso[a]
Instituto Politécnico Nacional.
Escuela Superior de Cómputo.
jmartinezp@ipn.mx
Av. Juan de Dios Batíz, esq. Miguel Othón
de Mendizábal, Av. Juan de Dios Batíz, esq. Miguel Othón de Mendizábal,
Mexico City 07738, Mexico.

J. V. Arroyo Luna[b]
Instituto Politécnico Nacional
[b]Escuela Superior de Ingeniería Química e Industrias Extractivas.
vanne_luu@hotmail.es
Av. Luis Enrique Erro s/n, Mexico city 07738, Mexico.

M.A. Maldonado-Muñoz[a]
Instituto Politécnico Nacional.
Escuela Superior de Cómputo.
mmaldonadom@ipn.mx
Av. Juan de Dios Batíz, esq. Miguel Othón de Mendizábal,
Mexico City 07738, Mexico.

Abstract: The density of breast tissue is an extremely important risk factor for the development of breast cancer. Women who have a significant amount of dense breast tissue are more likely to have cancer than those whose tissues are less dense. This paper presents an automated method for the recognition of breast tissue density. For this purpose, images obtained from the mini-MIAS data base database (Mammogram Image Analysis Society database (UK)) were used.

The process that leads to recognition is divided into three phases: 1) segmentation of the breast, 2) extraction of characteristics and 3) classification. The segmentation of the breast was carried out using methods in the spatial domain, after which 11 statistical features were extracted, and finally a neural network was applied for the classification process. The performance of the model was measured using the k-fold cross validation method and the results were compared with those obtained in other works. *(Abstract)*

Keywords: pattern recognition, segmentation of images, neural networks, breast tissue density

I. INTRODUCTION

Each year, 1,151,000 new cases are diagnosed of breast cancer all over the world [1], becoming a health problem that is increasing, being the leading cause of death among women. Early detection of cancer can reduce mortality, for this, women should perform clinical trials such as mammograms [2]. A mammogram is a breast x-ray taken in a digital image. This test can detect tumors that are too small to be detected by touch. The ability of mammography on detecting breast cancer may depend on tumor size, the density of breast tissue, and skill of the radiologist [3, 4]. The density of breast tissue is a major risk factor for developing breast cancer [5, 6, 7], making it more sensitive for those women with more dense tissue in the breast.

This paper presents the proposal of an automated method for recognition of breast tissue density. For this, we used the image bank available on mini-MIAS database [8]. This image database consists of digital mammograms classified in three types depending on the density of breast tissue presented, this can be: a) fat, b) fatty-glandular and c) dense-glandular. The method consists of three phases: 1) segmentation of the breast: for that were used the spatial domain methods [9], 2) feature extraction: after obtaining the segmented breast image, we extracted statistical features as the average, standard deviation, smoothness and entropy among others, 3) Classification: For this phase, we used neural networks models.

II. PROPOSED MODEL

Before going into details with the proposed model, it should be mentioned the definition of what is a digital image. A digital image is a two-dimensional function f(x,y) of the light intensity (brightness) at a point in space, where (x,y) coordinates are the point [9]. Since a digital image is a function f(x,y) coordinates discretized in space and in the brightness, is often represented as a two dimensional matrix $F_{ij}=(f_{ij})_{(m\times n)}$, *where m and n represent the size of the image and* $f_{ij}= f(x_i,x_j)$.

As mentioned above the method described consists of three phases: 1) segmentation of the breast, 2) features extraction and 3) classification.

A. 1) Segmentation of the breast.

For the process of segmentation of the breast, we consider the following 7 steps:

Step 1: the image is considered as originally mammography has grayscale.

Step 2: Given an image *f (x,y)* and a variable v, the binarization by thresholding used is defined as follows:

$$bin_{ij} = \begin{cases} 255, & f_{ij} \leq v \\ 0, & f_{ij} > v \end{cases}$$

For this step, we consider a threshold of 140 as a first approximation.

Step 3: Image is divided into three rectangular areas, for this purpose, consider the middle of the image horizontally, and the first 30 pixels of the image vertically as shown in the figure. 1) Once established the rectangular areas, is calculated the area enclosed in each of the rectangles defined by the binary image. It is estimated that the pectoral muscle is on the side of the image where the rectangular area enclosing the largest area.

Step 4: Extract the first connected component that is making the route from top to bottom, left to right if the estimate of the pectoral muscle was on the left, or right to left if the estimate of the pectoral muscle was on the right.

Step 5: We consider two vertical lines of pixels, one to the height of the end of the pectoral muscle and the other the top half as shown in the figure. Performing a scan on the vertical lines from top to bottom so that if there is an area of black pixels followed by a zone of white pixels, and then another area of black pixels, means that the image, in addition to segmented pectoral muscle, has segmented part of the breast, in this case is considered again step 2 with a threshold equal to the value plus an increase of 10. In case there is a black pixel area followed by only an area of white pixels means that has been segmented only the pectoral muscle.

Step 6: Calculate the difference between the original image of the mammogram and the negative of the image obtained in step 5, to obtain the image of the mammogram with the pectoral muscle removed.

Step 7: Binarize the image with a threshold of 70 and then the extraction of the connected component with the largest area.

Figure 1 shows the result of applying the steps from two images of digital mammography database mini-MIAS. In step 5 corresponding to the figure a) shows that it has segmented the pectoral muscle with a section of the breast, in this case, re-perform step 2 with a threshold equal to the number with an increased value of 10. In the column for step 5* shows the result of applying step 1 to 5 in a satisfactory manner. Figure 1 also shows the segmented image of the breast.

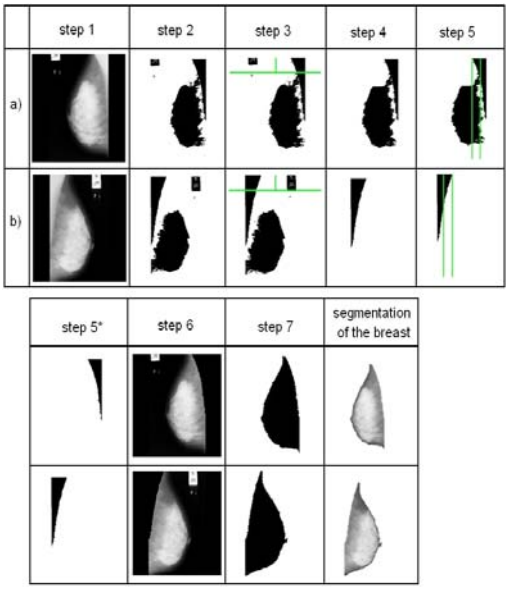

Figure 1. Process of segmentation of the breast

2) Extraction of features.

Once the segmentation of the breast obtaining images as shown in Figure 1 is proceeded to the feature extraction. A total of eleven statistical features were extracted from the image as used in [10, 11]. The respective formulas of each feature are shown in Table 1, where N represents the total number of pixels, L is the total number of gray levels, I(f_ij) is the gray level value of pixel (i,j) in the image f(x,y), P(j) is the probability that the value j of the gray level occurring in the image f(x,y), T(i) is the number of pixels with gray value i in the image f(x,y), P(I(f_ij)) is the probability that the gray level I(f_ij) occurs in the image f(x,y) and P(f_ij)=T(I(f_ij))/N.

Feature	***Expression***
Average	$\mu = \frac{\sum_{ij} f_{ij}}{N}$
Standard deviation	$\sigma = \sqrt{\frac{\sum_{ij}(f_{ij}-\mu)^2}{N}}$
Smoothness	$R = 1 - \frac{1}{(1+\sigma^2)}$
Skewness	$\frac{\sum_{ij}(f_{ij}-\mu)^3}{N\sigma^3}$
Kurtosis	$\frac{\sum_{ij}(f_{ij}-\mu)^4}{(N-1)\sigma^4}$
Uniformity	$\sum_{i=0}^{L-1} P(i)^2$

Average histogram	$AH_g = \frac{1}{L}\sum_{i=0}^{L-1} T(i)$
Skew modified	$MSK = \frac{1}{\sigma^3}\sum_{ij}(f_{ij}-\mu)^3 P(f_{ij})$
Modified standard deviation	$\sigma_m = \sqrt{\sum_{ij}(f_{ij}-\mu)^2 P(f_{ij})}$
Entropy	$Etp = -\sum_{j=0}^{L-1} P(j)\log_2[P(j)]$
Entropy modified	$\sum_{ij} P(f_{ij})\log_2[P(I(f_{ij}))]$

Table 1: Statistical features considered.

Table 2 shows the characteristics extracted from some images of the mini-MIAS database.

IMAGE	AVERAGE	STANDAR DEVIATION	SMOOTHNESS	SKEWNESS	KURTOSIS	UNIFORMITY	AVERAGE HISTOGRAM	SKEW MODIFIED	MODIFIED ESTÁNDAR DESVIATION	ENTROPY	ENTROPY MODIFIED
mdb001	140.206467	44.290692	0.999490	-0.038118	1.686019	0.007462	760.886292	-476.6246	1792.818481	7.129341	12260.208984
mdb007	141.454224	25.672352	0.998485	-0.796881	3.242217	0.012782	1172.109863	-251.6581	1089.602661	6.561669	24019.404297
mdb57	142.772308	22.640242	0.998053	-0.920047	3.946599	0.014793	1268.192139	-52.25266	1002.554199	6.393721	28890.464844
mdb145	132.114624	34.612823	0.999166	0.493009	2.272666	0.009728	1759.074463	-326.8648	1923.309448	6.944399	30298.964844
mdb220	136.796997	33.405602	0.999105	-0.711143	3.893436	0.009459	1317.439209	-1548.9267	1567.274902	6.881667	45683.531250
mdb310	136.902664	20.349438	0.997591	-2.816161	15.245199	0.025686	1690.333374	-4094.945	1109.188721	5.657229	77624.945312

Table 2. Characteristics extracted from the mini-Mias

B. 3. Classification

For the classification process, the multi-layer perceptron neural network model was used with forward and backward connections using the backpropagation learning algorithm. This has a configuration of a layer of input neurons, a layer hidden in the middle, and a layer of output neurons. Each junction between one neuron and another is associated with a weight. We define w_{ij} as the weight that feeds neuron j from neuron i and we have that the value of neuron j, x_j is the sum of w_{ij} xi for all i.

To train the network using the backpropagation algorithm, this procedure is performed by calculating the value of all neurons. Their weights are initialized randomly. Once the output layer is reached, the results obtained are compared with those obtained by obtaining a network error measure for the evaluated data. This error is retropropagated by adjusting all weights. This procedure is performed once for each data in a training set and at the end it is iterated to ensure that the error is becoming smaller.

III. EXPERIMENTS AND RESULTS

[1] In this paper we use the bank of images available from mini-MIAS database [8]. The images are classified into three types, according to the breast tissue, this can be: a) fat, b) fatty-glandular and c) dense-glandular. The repertoire of images comprises a total of 322 images, each of size 1024 x 1024, of which 106 correspond to fat, 104 to fatty tissue and 112 to glandular density-glandular tissue.

Given the data bank, the first step was the digital image processing to obtain the segmentation of the breast and the 11 statistical characteristics were extracted. The set was partitioned into a training set and a test set. Later, different tests were performed using sets from 40 and up to 200 patterns to test the neural network, validating the results using the cross validation method, which consists of partitioning the total set into k blocks of equal size, then we use k-1 blocks as a training set and a block as test set, we repeat the process for each of the blocks, in this way each block at some point it is a test set, and the rest of the training.

Figure 2 shows the experimental results using 40, 60, 100, 150 and 200 images of the mini-Mias database. Obtaining a percentage of yield of between 83.5% and 94.75%.

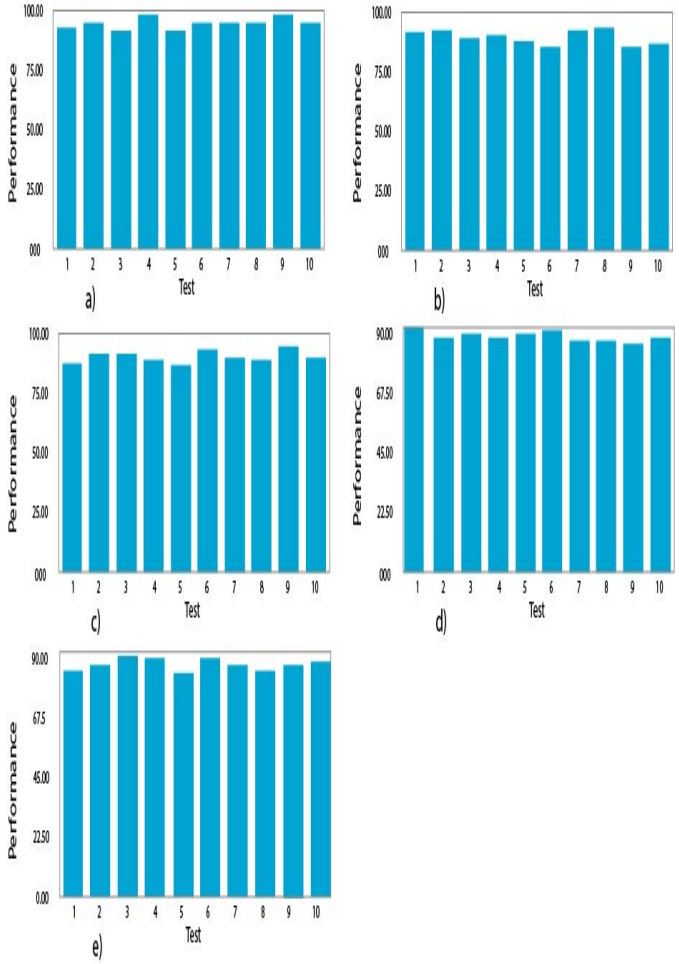

Figure 2. Experimental data using: a) 40, b) 60, c) 100, d) 150 and e) 200 images of the mini-Mias database.

Table 3 shows the comparison of the results with other proposed models.

Classifier used	Number of patterns used	Number of patterns used for this work	Success percentage	Proposed solution	Reference	
DAG-SVM	180 Patterns	150	77.57%	85.73%	Zwiggelear	[56]
Bayesian networks	40 Patterns	40	80%	94.75%	Miller and Astley	[57]
SVM	40 Patterns	40	95.44%	94.75%	T.S. Subashini	[59]
Statistical classifiers with genetic algorithms	67 Patterns	60	85.00%	90.33%	P. Zhang and B. Verma	[60]

Table 3. Comparison of results with other works

In this paper we present an automated method for the recognition for density of breast tissue in digital mammography images. The process was divided into three phases. In the first phase was carried out the segmentation of the breast digital mammography images. The second phase was the extraction of statistical features of the segmented image of the breast.

In the third phase it includes the classification process, where the neural network model with backpropagation was used. Table 3 shows the results of the different experiments, where it was observed that a yield of 94.75% was obtained considering 40 images, and 89.5% considering 200 images of the mini-Mias bank.

ACKNOWLEDGMENT

The authors wish to thank the Instituto Politécnico Nacional (Secretaría Académica, Secretaría de Investigación y Posgrado, ESCOM, ESIQIE, COFAA, EDD) for his financial support for the development of this work.

REFERENCES

[1] Gutiérrez I. Z., Schneider F. J., ¿Sabemos qué causa el cáncer de mama? Influencia actual de los diferentes factores de riesgo. Review Article Progresos de Obstetricia y Ginecología, Volume 52, Issue 10, October 2009, pp. 595-608.

[2] Scott-Pretorius E., Solomon-Jeffrey A., Mamografía de cribado., Radiología Secretos (Segunda edición), Madrid, Elsevier 2006, pp. 39-43.

[3] Hendrick E. R., Impact of ACRIN DMIST Results on the Technologist and Mammography Practice Original., Research Article Seminars in Breast Disease, Volume 8, Issue 4, December 2005, pp. 178-184.

[4] Basset L. W., Hoyt A. C., Oshiro T., Digital Mammography: Clinical Image Evaluation Review Article Radiologic Clinics of North America., Volume 48, Issue 5, September 2010, pp. 903-915.

[5] Wong C. S., Lim G. H., Gao F., Jakes R. W., Offman J., Chia K. S., Duffy S. W. Mammographic density and its interaction with other breast cancer risk factors in an Asian population. British Journal of Cancer, 104 (5), pp. 871-874, Mar 2011, doi:10.1038/sj.bjc.6606085.

[6] Stone J., Ding J., Warren R., Stephen W. H., John L., Using mammographic density to predict breast cancer risk: dense area or percentage dense area., Breast Cancer Research, 12 (6), p.R97, Nov 2010 doi:10.1186/bcr2778.

[7] Colin C., Prince V., Valette P. J., Can mammographic assessments lead to consider density as a risk factor for breast cancer?, Review Article European Journal of Radiology, In Press, Corrected Proof, Available online 4 February 2010.

[8] Suckling J., Parker J., Dance D., Astley S., Hutt I., Boggis C., Ricketts I., Stamatakis E., Cerneaz N., Kok S., Taylor P., Betal D., and Savage J., The Mammographic Image Analysis Society Digital Mammogram Database. Excerpta Medica. International Congress Series, vol. 1069, 1994, pp. 375-378.

[9] González R., Woods J. Digital image processing, 3rd edn. Prentice Hall, New Jersey, 2008.

[10] Zhang P., Verma B., Kumar K., Neural vs. statistical classifier in conjunction with genetic algorithm based feature selection., Pattern Recognition Letters, Volume 26, Issue 7, 15 May 2005, pp. 909-919.

[11] Subashini T. S., Ramalingam V., Palanivel S., Automated assessment of breast tissue density in digital mammograms., Original Research Article Computer Vision and Image Understanding, Volume 114, Issue 1, January 2010, pp 33-43.