ANALYSIS OF GROUPS OF FEATURES FOR FINDINGS DISCERN MAMMOGR APHICTHROUGH TECHNICAL SFS

A. C. Picanço Batista*, C. L. S. Melo*, Cícero F.F.Costa Filho**

**Engenharia de Controle e Automação*
Universidade do Estado do Amazanas
Manaus, Brazil.
E-Mail: afonsopican@gmail.com

***Programa de Pós Graduação de Engenharia Elétrica*
Universidade Federal do Amazonas
Manaus,Brasil
E-Mail: cluiz@uea.edu.br

Resumen: La clasificación de los hallazgos mamográficos, microcalcificaciones y grupos de microcalcificaciones, ya sea benigno o maligno, es una tarea difícil. Esto se debe principalmente a la variabilidad de su apariencia. Función de selección apropiada es probablemente el paso más crítico de un proceso de clasificación automática. Este trabajo tuvo como objetivo identificar un conjunto de características que permite hacer la mejor clasificación automática. Los grupos con un número diferente de características se generaron utilizando la función escalar Selección - SFS. Ratio de Fisher discriminante - FDR y el área bajo la curva receptor operativo - ROC se utilizaron como medidas de distancia auxiliares. A efectos de clasificación, se emplearon diferentes arquitecturas de redes neuronales feedforward. Los resultados se evalúan mediante el método de validación cruzada utilizando mediciones de precisión, sensibilidad y especificidad.

Abstract: Classifying mammographic findings, microcalcifications and clusters of microcalcifications, as either benign or malignant, is a difficult task. This is mainly due to the variability of their appearance. Appropriate feature selection is probably the most critical step of an automatic classification process. This paper aimed to identify a set of features that allows for making the best automatic classification. Groups with different numbers of features were generated using the Scalar Feature Selection – SFS. Fisher's Discriminant Ratio - FDR and the area under Receiver Operating Curve – ROC were used as auxiliary distance measurements. For classification purposes, different architectures of feedforward neural networks were employed. The results are evaluated through the cross validation method using measurements of accuracy, sensitivity and specificity.

Keywords: Microcalcification; cluster of microcalcification; neural network; pattern recognition.

1. Introdução.

Breast cancer is the second type of cancer that affects more women around the world, is the most common among them and also a leading cause of death in Brazil [1]. The early identification is the main strategy for the control of breast cancer and mammography one of the main means used for this purpose, because its spatial resolution allows the diagnosis of millimetric nodules. [1]. Cancers observed by mammography is introduced in most instances in the form ofclusters of microcalcifications [2].

Computer systems for easy the diagnosis (CAD) have been proposed with the objective of assisting the radiologist, acting as an "alternative", assisting in the location ofabnormalities and in suggesting the diagnosis. The main modules of a CAD on a mammogram are shown in Figure 1: Caption, Detectionand Classification.

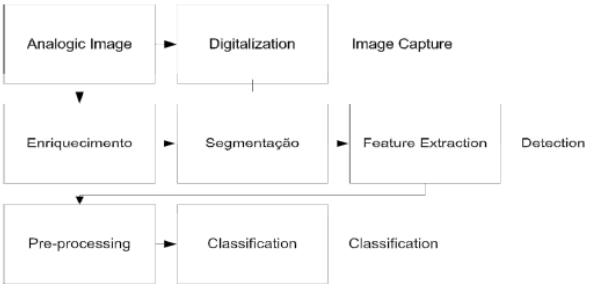

Figure 1 - Block Diagram of a Mammography CAD System.

First, it is scanning the mammographic image, with varying rates of sampling andquantization. In the enhancement phase, regions of interest (ROI) selected based on themammographic images are highlighted, seeking to expel noise and artifacts. The period of segmentation is done in order to to find suspicious áreas containingmicrocalcifications and/or clusters and the separation of the same from the bottom of the image. Soon after the step of feature extraction is carried out. Usually, the set of extracted characteristics is assigned one of the following groups: shape descriptors, such as: area, eccentricity, circularity, irregular, perimeter, density, compactness [4,5,6,7]; texture descriptors: energy, entropy, angular momentum, correlation, contrast [8,9,10,11] and

wavelet descriptors: energy, entropy [12]. The block of classification gathers the steps in of preprocessing and classificationfeatures. In the preprocessing step is drawn up the standardization of features and it is estimated the discriminating power of it, and the exclusive or a group of them. Shown in many studies in the literature, the authors perform the normalization of thesevalues for particular intervals from 0 to 1 [4,8,12,13,14,15,16] or at intervals -1 to 1 [5,17], without a concern to establish the discriminating power of them, individually or in group. This normalization is essential because the original characteristics are found in dynamic and different bands, which may negatively influence the classification step [19].

Other authors, however, are concerned with ascertain characteristics of groups with a higher discriminating power. The proportion of feature selection techniques used are diverse. In the works [6,18,23] the authors used the technique of the principal components analysis; In the works [2, 7] the authors used techniques reverse feature selection. In [2,23] the authors used techniques for direct selection of features; The k-means technique was used only in [20]. The Mahalanobis distance was used in [10] and genetic algorithm was used in [11]. In the classification stage are used supervised classifiers so as not 2/5 IV SEB-UFU 2011. Through the literature review it became clear that, in recent decades, the most widely usedmethod for supervised classification of mammographic findings was that of artificial neuralnetworks [2,4,5,6,8,9,11,12,14,15,18,20,21]. Among the unsupervised methods used is cited: k-NN [2,12,14,21] and SVM (Support Vector Machine) [12,16,23]. the main theme of this work is to improve the classification stage by identifying a set of characteristics that expose a power of discrimination, without the need to use all the features taken from lesions, and an architecture of a neural classifier direct propagation, which is order to identify malignant and benign cases. The mammographic findings were obtained in previous work [17]. The various features used were chosen through technical Scalar Feature Selection-SFS.Measures of auxiliary distance measure we used FDR (Fisher's discriminant ratio) and the area under the ROC curve (Receiver Operating Characteristics. The follow up and performance of the method was evaluated by cross-validation technique, enjoying the measures of accuracy (percentage of correct classification of the method with the histological report), sensitivity (percentage of correct classificationmethod with malignancies of the report) and specificity (percentage of correct method withbenign cases of the report).

2. Materials and Methods.

The database used contains a set of sixteen characteristics, eight of which arecharacteristics of microcalcifications, while the other half are characteristic of *clusters* of microcalcifications. These characteristics are derived from 80 samples (ROI) of images from the collectedimages of MIAS (*Mammographic Image Analysis Society*) and INCa-RJ, together with the corresponding histological analysis (biopsy).

The eight features concerning microcalcifications are: area - m1, eccentricity - m2, compactness - m3, folding - m4, contrast - m5, narrow irregularity - m6, large irregularity - m7 and guidance - m8. Related to *clusters* of microcalcifications are: perimeter - c1, area - c2, density - c3, eccentricity - c4, average distance from the center of the cluster microcalcifications - c5, orientation - c6, relative distances to the edge of the breast - c7 and pectoral muscle - c8.

The method for automatic classification of microcalcifications and *clusters* of microcalcifications used in this work can be divided into four steps:

STEP 1. Setting a ranking of the best features concerning microcalcifications and *clusters* of microcalcifications, for it using the auxiliary measures: FDR and area under the ROC curve. FDR is often used to quantify the discrimination capability of a feature [22]. This measure is calculated according to equation:

$$FDR = \frac{(\mu_1 - \mu_2)^2}{(\sigma_1^2 + \sigma_2^2} \quad (1)$$

Where: μ1 - sample mean of class ω1; μ2 - ω2 sample mean of the class; σ12 - sample variance of class ω1; σ22 - sample variance of ω2 class.

The area under the ROC curve is a two-dimensional graph where the y-axis represents the value of the rate of true positives - *tp* (cases that are correctly classified as histological report) and the X axis represents the value of the false positive rate - fp (misclassifiedcases compared by histological report).

All characteristics were normalized to a range of values between -1 to 1.

STEP 2. Such as the characteristics of each mammographic finding are eight, was chosen in the formation of four groups of features with 5, 6, 7 and 8 features, which have higherdiscriminating power for the set of features of microcalcifications and for the set of features regarding the clusters of microcalcifications, totaling eight groups formed.

For forming these groups was used the SFS technique (Scalar Feature Selection). In general, the selection of features, is used as a criterion a measure of distance betweenclasses – C(k). In this work we used the auxiliary measures mentioned in Step 1 The formation of groups of features follows the following procedure:

- It is calculated C (k) for all features and organized features in a vector in descendingorder of C (k);

- Selects to 1st feature that corresponds to the first element of the array of ordered away;

- Compute the correlation coefficient between the selected feature and all other featuresby applying:

$$P_{ij} = \frac{\sum_{n-1}^{n} x_{ni} x_{nj}}{\sqrt{\sum_{n-1}^{n} x_{ni}^2 \sum_{n-1}^{n} x_{nj}^2}} \quad (2)$$

Where: *xni* - *n* sample feature i; *xnj* - n sample feature *j*; N – total number of samples of a feature.
C2 is chosen characteristic for which:

$$i_2 = \arg max_j \left\{ \alpha_1 C(j) - \alpha_2 \left| \rho_{1_1 j} \right| \right\} \text{ for any } j \neq i_1 \quad (3)$$

Where: *α1* and *α2* are constant Gaussian (0 <α <1) which prioritize the importance of the terms C (j); *ρij* is the correlation coefficient among traits *i* and *j*; C (j) is the distance measure feature *j*.
– For the choice of the k-th feature applies:

$$i_k = \arg max_j \left\{ \alpha_1 C(j) - \frac{\alpha_2}{k-1} \sum_{r=1}^{k-1} \left| \rho_{1_r j} \right| \right\} \text{ for any } j \neq i_r \quad (4)$$

The Table 1 and 2 show the characteristic groups of microcalcifications and clusters of microcalcifications using as distance measure the FDR and area under the ROC curve, respectively. The characteristics of these groups were applied as input variables of the neural network direct propagation for classification of microcalcifications and *clusters* of microcalcifications in benign and malignant cases.

Table 1 - Groups resulting from the application of the measure FDR

GROUP	CHARACTERISTICS OF MICROCALCIFICATIONS	CHARACTERISTIC OF *CLUSTERS* OF MICROCALCIFICATIONS
GROUP 1 - containing the 5th best characteristics	m_6, m_8, m_1, m_3, m_2	c_7, c_1, c_8, c_5, c_6
GROUP 2 – containing the 6th best characteristics	$m_6, m_8, m_1, m_3, m_2, m_7$	$c_7, c_1, c_8, c_5, c_6, c_2$
GROUP 3 – containing the 7th best characteristics	$m_6, m_8, m_1, m_3, m_2, m_7, m_4$	$c_7, c_1, c_8, c_5, c_6, c_2, c_4$
GRUPO 4 – containing all of the characteristics	$m_6, m_8, m_1, m_3, m_2, m_7, m_4, m_5$	$c_7, c_1, c_8, c_5, c_6, c_2, c_4, c_3$

Table 2 - Groups resulting from the application of the measure area under the ROC curve

GROUP	CHARACTERISTICS OF MICROCALCIFICATIONS	CHARACTERISTIC OF *CLUSTERS* OF MICROCALCIFICATIONS
GROUP 1 - containing the 5th best characteristics	m_3, m_4, m_1, m_6, m_2	c_3, c_7, c_6, c_8, c_1
GROUP 2 – containing the 6th best characteristics	$m_3, m_4, m_1, m_6, m_2, m_7$	$c_3, c_7, c_6, c_8, c_1, c_5$
GROUP 3 – containing the 7th best characteristics	$m_3, m_4, m_1, m_6, m_2, m_7, m_8$	$c_3, c_7, c_6, c_8, c_1, c_5, c_2$
GRUPO 4 – containing all of the characteristics	$m_3, m_4, m_1, m_6, m_2, m_7, m_8, m_5$	$c_3, c_7, c_6, c_8, c_1, c_5, c_2, c_4$

STEP 3. Use the classifier based on artificial neural network with three layers: input, middle andoutput.

The definition of the architecture of neural networks for microcalcifications and *clusters* of microcalcifications was defined by the steps described in [25]. The procedure was usedfor the following architectures of neural networks:8-n-1, 7-n-1, 6-n-1 e 5-n-1. For all these architectures the best combination of accuracy and convergence time fornetworks that used features of microcalcifications in entry was obtained with two neuronsin the hidden layer. As for the network that used the input characteristics of *clusters* of microcalcifications, the best combination was obtained with three neurons in theintermediate layer.

The training of neural networks was performed using the optimization method of *Levenberg-Marquardt* [24]. The convergence criterion used was a smaller mean square error than 10-4. Table 3defines the architectures of neural networks used in this work. All architectures were tested with groups of variables defined in Tables 1 and 2.

Table 3: Architecture of Neural Networks and input variables used

Entrance Variables	Architecture of RNA microcalcifications	Architecture of RNA Clusters of Microcalcifications
Group 1, 2, 3 ,4	5-3-1	5-2-1
Group 1, 2, 3 ,4	6-3-1	6-2-1
Group 1, 2, 3 ,4	7-3-1	7-2-1
Group 1, 2, 3 ,4	8-3-1	8-2-1

3. Results

For obtaining any results the cross validation method was used. The samples were divided into four subsets of twenty images each. Each neural network architecture was trained with three subsets and tested with the remaining subset. As there are fourpossibilities of combining the four subsets of images in threes, this procedure was repeated four times. For the four tests then took out the average accuracy, sensitivity and specificity.
Tables 4 and 5 show the best performances of the various architectures of neural networkusing as input variables the characteristics of clusters of microcalcifications, when usingdistance measurements FDR and area under the ROC curve, respectively. For each architecture the average values of accuracy, sensitivity and specificity of the group thatpresented the best result for accuracy is shown. It can be observed that the best results in terms of accuracy using the characteristics of clusters of microcalcifications were obtained with the architecture 6-2-1 using the group with six features and the distance measure FDR.

Table 4 - Best performances of the four architectures of neural networks using as input variables the characteristics of *clusters* of microcalcifications and as far away as the FDR.

Architecture	Ac (%)	S(%)	E(%)	Group
5-2-1	82.74	80.25	88.69	1
6-2-1	86.19	80.56	95.83	2
7-2-1	82.86	81.50	85.42	3
8-2-1	82.74	85.57	80.36	4

Table 5 - Best performances of the four architectures of neural networks using as input variables the characteristics of *clusters* of microcalcifications and how far away the area under ROC curve.

Architecture	Ac (%)	S(%)	E(%)	Group
5-2-1	80.83	80.90	82.44	3
6-2-1	84.52	87.15	82.89	2
7-2-1	86.01	86.61	85.71	2
8-2-1	81.07	85.42	79.18	4

Tables 6 and 7 show the best performances of the various architectures of neural networkusing as input variables the characteristics of microcalcifications, when using distance measurements FDR and area under the ROC curve, respectively. For each architecturethe values of accuracy, sensitivity and specificity of the médium group showed the best result for accuracy is shown. It can be observed that the best results in terms of accuracyusing the characteristics of microcalcifications were obtained with the architecture 6-3-1, using as input variables the group of seven features and as far away as FDR.

Table 6 - Best performances of the four architectures of neural networks using as input variables the characteristics of microcalcifications and as far away as FDR.

Architecture	Ac (%)	S(%)	E(%)	Group
5-3-1	66.11	69.22	64.87	1
6-3-1	72.81	67.15	59.69	2
7-3-1	68.47	70.49	66.34	3
8-3-1	72.57	79.83	67.71	4

Table 7 - Best performances of the four architectures of neural networks using as input variables the characteristics of microcalcifications and how far away the area under ROCcurve.

Architecture	Ac (%)	S(%)	E(%)	Group
5-3-1	68.61	73.75	63.75	1
6-3-1	72.36	75.48	70.18	2
7-3-1	68.47	70.49	66.34	3
8-3-1	71.57	78.83	68.71	4

4. Discussion and Conclusion.

From Tables 1 and 2 it appears that the distance measures FDR and area under the ROCcurve result in diferente sets of characteristics when using the selection technique SFS.

From Tables 4,5,6 and 7 it is seen that the best performance of neural classifiers used are not obtained with the maximum number of features available. This demonstrates the validity of using a feature selection technique, worrying so with the power of thediscriminant analysis of the characteristics of mammographic findings.
Classifiers using features of clusters of microcalcifications showed superior results.
Although not shown, the performance of neural classifiers simultaneously using features ofboth groups was lower than the results presented here and for groups of features used separately.
Soon, you realize that to identify and / or classify mammographic findings is not necessary to use all the features that describes.

References

[1] INCa, 2012, 2013 Forecasting: "Cancer incidence in Brazil", Rio de Janeiro.

[2] D. Kramer, F. Aghdasi, "Texture analysis techniques for the classification of microcalcifications in digitised mammograms," Proceedings of the 1999 Fifth IEEE AFRICON Conference Electrotechnical Service for Africa, September 28-October 1, 1999, pp. 395-400.

[3] H.D. Cheng, X. Cai, X.Chen, L. Hu, X. Lou, " Computer-aided detection and classification of microcalcifications in mammograms: a survey," Journal of the Pattern Recognition Society, IEEE,v.36, pp. 2967-2991, March 2003.

[4] Y. Jiang, R.M. Nishikawa, D.E. Wolverton, C.E. Metz, R.A. Schmidt, K. Doi, "Malignant and Benign Clusterd Microcalcifications: Automated Feature Analysis and Classification," Radiology, v. 198, Number 3, pp.671-678, March, 1996.

[5] H. Nguyen, W.T. Hung, B.S. Thornton, E. Thornton, W. Lee, "Classification of Microcalcifications in Mammograms using Artificial Neural Networks,"Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 20, No 2, 1998.

[6] A. Sepehr, M.H. Jamarani, B. Gholamali Rezai-rad, C. Hamid Behnam," A Novel Method for Breast Câncer Prognosis Using Wavelet Packet Based Neural Network," Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, September 1-4, 2005.

[7] L. Wei, Y. Yang, R.M. Nishikawa, "Microcalcification Classification Assisted by Content-based Image Retrieval for Breast Câncer Diagnosis,"Pattern Recognition, 2008.

[8] D.Tsai, H.Fujita, K.Horita, T. Endo, "Classification of Breast Tumors in Mammograms using a Neural Network: Utilization of Selected Features,"Proceedings of 1993 International Joint Conference on Neural Networks.

[9] B. Verma, J. Zakos, "A Computer-Aided Diagnosis System for Digital Mammograms Based on Fuzzy-Neural and Feature Extraction Techniques, "IEEE Transactions on Information technology in Biomedicine, vol. 5, No. 1, march 2001.

[10] P. Zhang, B. Verma, K. Kumar," Neural vs. Statistical Classifier in Conjunction with Genetic Algorithm Feature Selection in Digital Mammography,"IEEE, pp. 1206-1213, 2003.

[11] K.Geetha, K. Thanushkodi, A.K.Kumar, "New Particle Swarm Otmizations for Feature Selection and Classification of Microcalcifications in Mammograms,"IEEE- International Conference on Signal Processing, Communications and Networking Madras Institute of Technology, Jan 4-6, pp. 458-463, 2008.

[12] M.A. Alofe, A.M.Youssef, Y.M.Kadh, A.S.Mohamed, "Computer - aided diagnostic system based on wavelet analysis for microcalcification detection in digital mammograms," IEEE Transactions on Medical Imaging, CIBEC, 2008.

[13] Y. Jiang, R.M. Nishikawa, D.E. Wolverton, C.E. Metz, "Computerized Classification of Malignat and Benign Clustered Microcalcifications in Mammograms," Proceedings – 18th International Conference – IEEE/EMBS Oct. 30-Nov. 2, pp. 521-523, 1997.

[14] L. Arbach, J.M. Reinhart, D.L. Bennett, G. Fallouh, "Mammographic Masses Classification: Comparision between BNN, KNN and Humam Readers," CECE-CCGEI, May, 2003.

[15] W. Chiracharit, R. Kongkachandra, "Clustered Microcalcification Classification Using CC-MLO- View Corresponding Shape and Distribuição Features," SICE Annual Conference, August 20-22,2008.

[16] S. Singh, V. Kumar, H.K. Verma, D. Singh, "SVM Based System for classification of Microcalcifications in Digital Mammograms," Proceedings of the 28th IEEE, EMBS Annual International Conference, Aug 30-Sept 3, 2006.

[17] F.V. Pimentel, "Extração de parâmetros das microcalcificações de imagens mamográficas usando morfologia matemática," COPPE –UFRJ, Tese de Mestrado, Março 2004.

[18] S. Sehad, S. Desarnaund, A. Strauss," Artificial Neural Classification of Clustered Microcalcifications on Digitized Mammograms," IEEE, pp. 4217-4222, 1997.

[19] T.H. Lin, H.Li, K.C. Tsai, "Implementing the Fisher's Discriminant Ratio in a K-means Clustering Algorithm for Feature Selection and Data Set Trimming," American Chemical Society, June, 2003.

[20] Y. Chitre, A. P. Dhawan, M. Moskowitz, "Artificial Neural Network Based Classification of Mammographic Microcalcifications Using Image Structure and Cluster Feature,"IEEE 1993.

[21] R. Panchal, B. Verma, "Classification of Breast Abnormalities in Digital Mammograms using Image and BI-RADS Features in Conjunction with Neural Network," Proceedings of International Joint Conference on neural networks, July 31-Aug 4,2005.

[22] S. Theodoridis e K. Koutroumbas, Pattern Pattern Recognition, 3th Edition, Elseveir Academic Press, 2003.

[23] C. Wang, W. Jiang, X. Dong, "Char.cterization of clustered microcalcifications in mammograms based on support vector machines with genetic algorithms," IEEE, 2006.

[24] A. Cichoki, R. Unbehauen, "*Neural Networks otmization and signal processing*,", John Wiley & Sons, New York, 1993.

[25] de Melo, C.L.S.; Costa Filho, C.F.F.; Costa, M.G.F.; Pereira, W.C.A.; "Matching Input Variables Sets and Feedforward Neural Network Architectures in Automatic Classification of Microcalcifications and Microcalcification Clusters", In: 3rd International Conference on Biomedical Engineering and Informatics (BMEI 2010), Yantai, p., 358-362, 18-21 Nov.