RENDICONTI DEL CIRCOLO MATEMATICO DI PALERMO
Serie II, Suppl. 83 (2011), pp. 229-236

EXPLORING BREAST CANCER DIAGNOSIS WITH FRACTAL ANALYSIS AND CLASSIFICATION METHODS

GRAZIA RAGUSO1* - ANTONIETTA ANCONA[2] - LOREDANA CHIEPPA[1]
SAMUELA L'ABBATE[3] - MARIA LUISA PEPE[4]
FRANCESCO D. D'OVIDIO[3].
[1]Department of Mathematics, University of Bari, Bari, Italy.
[2]Radiology Unit, San Paolo Hospital of Bari, Bari, Italy.
[3]Department of Statistics, University of Bari, Bari, Italy.
[4]Department of Radiology, University Hospital - Policlinico of Bari, Bari, Italy.
[*]To whom all correspondence should be addressed
e-mail: raguso@dm.uniba.it
Phone: +39 – 080 – 544 – 2682
Fax: +39 – 080 – 544 – 3610

Abstract. Screening and diagnostic mammography are the most effective tools available for detection and diagnosis of breast cancer. In the last decade many techniques based upon measures of the shape of the contours of breast masses are been developed to investigate the nature of lesions between malignant tumours and benign masses. This paper presents methods for statistical analysis on a data set of 192 contours of breast masses. Results of these analysis lead to levels of accurate prediction in 90% of the cases, overcoming 98% for the diagnosis of malignant lesions. In this study we applied multivariate statistical techniques for examining relationships among more variables at the same time. We used in addition to the shape factors of contour masses also the age of the patients at the time of mammography, using both ROC analysis and segmentation analysis through Classification and Regression Tree.

1 Introduction.

Recent studies have shown that early detection through mammographic screening of asymptomatic women reduce breast cancer mortality. The true-positive and false-positive rates of

mammography vary in different age groups; the sensitivity of mammography is higher in women older than 50 years [1]. Mammography is the best method available for early detection of breast cancer. In order to assess a contour mass on mammograms, shape parameter are taken into consideration. On the basis of the notable shape differences we can distinguish between benign masses and malignant tumours. These observations have led to the idea of applying the concept of fractal dimension (FD) to analyze the contours of breast lesions [2]. Fractal analysis can characterize the degree of complexity of a contour or shape, and can provide parameters to discriminate between benign masses and malignant tumours [3].

In [3] we studied a data set of 192 mammograms were obtained from 192 patients at the Senology Unit, San Paolo Hospital, Bari, Italy, ASL Ba/4. The patients were diagnosed to have breast disease via screen-film mammography and confirmed from histological data; 163 of the cases were malignant and 29 were benign. The most useful mammographic projections were selected to analyze the contours of lesions. During an initial phase, contours of the present mammary lesions on the film images were traced by a team of radiologists specialized in mammography and successively, by a graphic tablet, we obtained a digital representation of the contour using Matlab software. Furthermore, we reported on a morphological study of 192 contours, with the aim of discriminating between benign masses and malignant tumours. From the contour of each mass, we computed the fractal dimension (FD) and a few shape factors, including compactness, 3 fractional concavity, and spiculation index. We calculated FD by using four different methods: the ruler and box-counting methods applied to each 2-dimensional (2D) contour and its 1-dimensional signature. Analysis using receiver operating characteristics (ROC) was performed with each shape feature to determine the diagnostic accuracy achievable in order to discriminate between benign masses and malignant tumours. ROC analysis indicated the area under the curve, A_z, of up to 0.92, having the individual shape features. The combination of compactness, FD with the 2D ruler method, and the spiculation index had as result in the highest A_z value of 0.93.

The data set, the shape features calculated for all data and the results obtained in the previous work [3], are used in this paper to

implement a different algorithm called CRT to have binary statistical classification of the variables (shape features). In addition to the shape factors of mass contours we introduced also the age of the patients at the time of mammography.

1.1 Fractal dimension and shape factors.

Fractals are irregular figures, and can be generated by the iteration of linear or nonlinear functions [4, 5]. Sometimes they are self-similar, and have a fine structure which reveals new details at every level of magnification [4]. In order to measure the degree of complexity or irregularity of a fractal, the concept of FD was introduced; this concept is derived from the more general notion of the Hausdorff dimension [6]. Cancerous tumours exhibit a certain degree of randomness associated with their growth, and are typically irregular and complex in shape. The degree of irregularity of the contour of a mass is the first parameter assessed: benign masses are often smooth, rounded, well-circumscribed, whereas a malig- 4 nant tumour is often characterized by an irregular contour with the spicules, that could be considered as a fractal pattern. Therefore, fractal analysis can provide a better measure of complex patterns. The Hausdorff dimension generalizes the concept of the self-similarity dimension in the sense that it is applicable to any set of the plane, and therefore, to a fractal set that is not strictly self-similar. The difficulties involved in defining the Hausdorff dimension have led many authors to find alternative methods for estimating FD. The common numerical methods are the box-counting and the ruler methods, which have been extensively described in the literature [6, 2].

On the basis of the differences in shape between benign masses and malignant tumours, various measures can be associated with a contour or curve: these are the so-called shape factors, which have proven to be effective in describing shapes in many research fields, in particular in the medical field [2, 7]. The shape factors used are compactness *cf*, fractional concavity f_{cc}, spiculation index SI; these measures have been proven to be effective in the classification of breast masses [8, 2, 9, 7]. See Rangayyan and Nguyen [2] and Rangayyan [7] for details on the shape factors.

Compactness is defined as [7]

$$cf = 1 - \frac{4\pi}{P^2}, \tag{1}$$

where P and A are the perimeter and the area of the contour, respectively. A high compactness value indicates a long perimeter enclosing a small area. Fractional concavity is defined as [8, 7]

$$f \propto = 1 - \frac{CC}{L}, \tag{2}$$

where CC represents sum of the lengths of the concave segments of the contour and L is the total length of the contour. Spiculation index is defined as

$$SI = 1 - \frac{\sum_{n=1}^{N} (1 + \cos \theta_n) S_n}{\sum_{n=1}^{N} S_n}, \tag{3}$$

where θ_n and S_n are the narrowness angle and the spicule length, respectively.

1.2 Statistics method: CRT algorithm.

The segmentation analysis allows researchers to determine (starting from a learning sample [10] of n independent units whose determinations are known in both dependent and explanatory variables) a classification rule able to divide the population in groups as homogeneous as possible inside them. Such rule will also be able to estimate the probability to detect a specific response, for other cases with unknown values of dependent variable but predictors with known determinations [11, 12]: in our case, the distribution of patients with unknown type of lesion (benign / malignant), based on some combinations of predictors. The segmentation analysis, in itself, is a recursive computing method which has some conceptual similarity with the cluster analysis: in both the methods will define some groups of observations which are homogeneous within the group and different from those of

other groups. Their basic principles, however, are different: the cluster analysis joins together the individual units sampled in groups according to all the considered variables, with the constraint of minimum variability "within" and maximum variability "between", without any constraints of hierar- 6 chy or dependence [13, 14]. The segmentation analysis, instead, divides a sample in aggregates which are more and more internally homogeneous with respect to a dependent variable, based on the values assumed by other variables, taken as explanatory, and on the relations between such variables and the dependent ones [13]. The best segmentation among all possible ones, based on the combination of different predictors, is that one that best meets the criteria of internal homogeneity of the groups generated (also known as "purity" [10]). Ideally, all cases of an final node should have the same value as the dependent variable (maximum purity). There are several methods of segmentation, but currently the most used are, among the algorithms of binary division, the CRT method [10, 15], and the CHAID type [16] among ternary or multiple algorithms.

2 Results and discussion.

The combined use of the shape parameters in [3] led to slight improvements in terms of accuracy: the combination of FD calculated using the 2D ruler method with *cf* and *SI* gave the highest A_z of 0.927. However, the combinations do not have a significant difference between one another. Nevertheless, in this work, we show that the conditioned combination of such factors can give us further information. Applying a segmentation analysis to data set (through CRT algorithm), we obtain classification trees that analyze the phenomena in the best way. The best result involves, in various combinations, (see Figure 1), both FD-2D calculated with ruler method and *SI*, as well as FD-1D calculated with the same method and age at diagnosis: i.e., women with $SI>0.232$ and age > 45, 5 years in the 98, 6% of analyzed cases have malignant lesions, but the disease probability is clearly lower (<52%) in women which $SI < 0.232$ (none of those aged $7 < 50.5$ years presents malignant lesions). Stopping the algorithm to first levels, the correct classification is near 91%, but further levels of

the analysis drive us to predict true positives and true negatives in 96% of the cases, with most capability in the analysis of the malignant disease (98%) than of benign masses (83%) (see Table 1, Table 2). Robustness of the algorithm showed in figure (classification tree) is sufficient: the jackknife re-sampling method (100 different sub-samples with half number of cases) give a maximal error risk < 0.12.

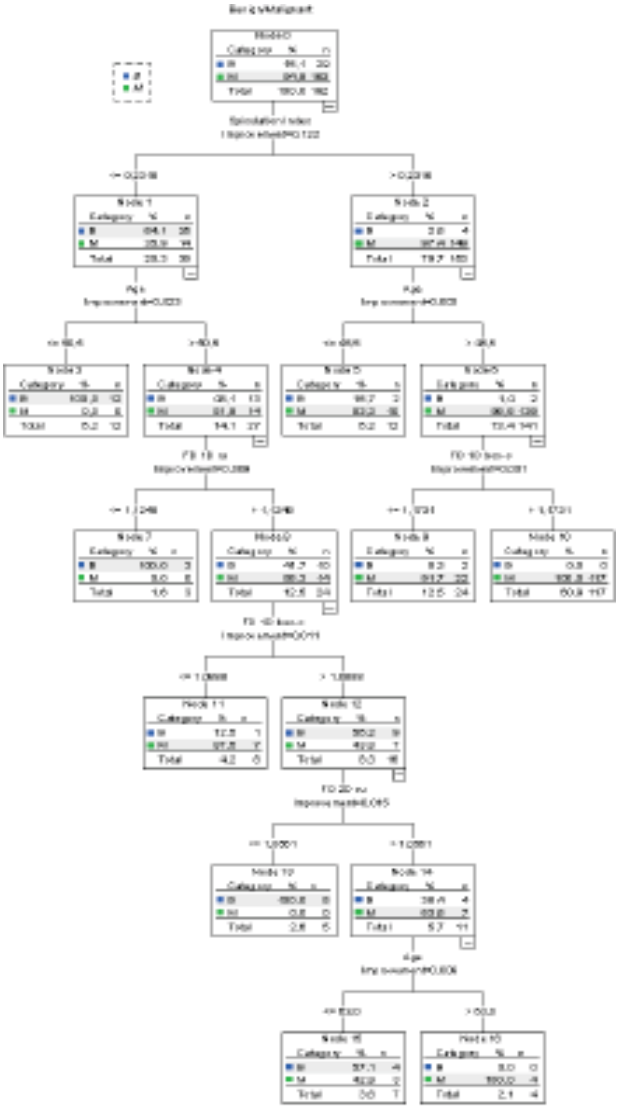

Figure 1: Segmentation Tree.

Table 1: Growing Method: CRT. Dependent Variable: Benign/Malignant.

Indipendent Variable	Importance	Normalized Importance
Spiculation Index	0.138	100.0%
FD-ruler 2D	0.137	99.3%
Compactness	0.125	90.6%
FD-ruler 1D	0.100	72.5%
FD-box 1D 0.	082	59.5%
FD-box 2D	0.081	58.7%
Fractional Concavity	0.055	39.8%
Age	0.047	34.1%

Table 2: Growing Method: CRT. Dependent Variable: Benign/Malignant.

Observed	Predicted		
	B	N	Percent Correct
B	24	5	82.8%
M	3	160	98.2%
Overall Percentage	14.1%	85.9%	95.8%

Acknowledgments

This work was supported by Fondazione Cassa di Risparmio di Puglia, Italy. We thank Rangayyan R.M., Thanh M. Nguyen and Naga R. Mudigonda for their contribution and discussion on the topic of this work.

REFERENCES

[1] R. W. Koudies A. I. Mushlin and D. E. Shapiro. Estimating the accuracy of screening maxnmography: a metaanalysis. *American Journal of Preventive Medicine,* 14:143–153, 1998.

[2] R. M. Rangayyan and T. M. Nguyen. Fractal analysis of contours of breast masses in mammograms. *Journal of Digital Imaging,* 20(3):223–237, 2007.

[3] G. Raguso, A. Ancona, L. Chieppa, S. L'Abbate, M. L. Pepe, F. Mangieri, M. De Palo, and R. M. Rangayyan. Application of fractal analysis to mammography. In *Proceedings of the 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2010,* Buenos Aires, Argentina, August 31-September 4, 2010.

[4] B. B. Mandelbrot. *The Fractal Geometry of Nature.* W. H. Freeman, San Francisco, CA, 1983.

[5] H. O. Peitgen and P. H. Richter. *The Beauty of Fractals - Images of Complex Dynamical Systems.* Springer - Verlag, Berlin Heidelberg, Germany, 1986.

[6] H. O. Peitgen, H. Jurgens, and D. Saupe. *Chaos and Fractals: New Frontiers of Science.* Springer, New York, NY, 2004.

[7] R. M. Rangayyan. *Biomedical Image Analysis.* CRC Press, Boca Raton, FL, 2005. 10

[8] R. M. Rangayyan, N. R. Mudigonda, and J. E. L. Desautels. Boundary modelling and shape analysis methods for classification of mammographic masses. *Medical and Biological Engineering and Computing,* 38:487–496, 2000.

[9] R. M. Rangayyan, N. M. El-Faramawy, J. E. L. Desautels, and O. A. Alim. Measures of acutance and shape for classification of breast tumors. *IEEE Transactions on Medical Imaging,* 16(6):799–810, 1997.

[10] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Chapman and Hall, New York, NY, 1993.

[11] J. A. Sonquist and J. N. Morgan. *The Detection of Interaction Effects.* Institute for Social Research, The University of Michigan, Ann Arbor, MI, 1964.

[12] J. A. Sonquist. *Multivariate Model Building. The Validation of a Search Strategy.* Institute for Social Research, The University of Michigan, Ann Arbor, MI, 1970.

[13] L. Fabbris. *Statistica multivariata. Analisi esplorativa dei dati.* McGraw-Hill, Milano, IT, 1997.

[14] F. Delvecchio. *Statistica per l'analisi di dati multidimensionali.* Cleup, Padova, IT, 2010.

[15] D. Biggs, B. De Ville, and E. Suen. A method of choosing multiway partitions for classification and decision trees. *Journal of Applied Statistics,* 18:49–62, 1991. 11

[16] G. Kass. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics,* 29(2):119–127, 1980.

DICHIARAZIONE SOSTITUTIVA DI ATTO DI NOTORIETÀ

(Art. 47 D.P.R. 28 Dicembre 2000, n. 445)

Il sottoscritto

d'Ovidio Francesco Domenico, nato il **25/07/1955** a **Bari** (prov. **BA**), residente a **Bari** (prov. **BA**) in via **Sardegna S.Spirito** n. **23,** C.F. **DVDFNC55L25A662V**

consapevole che chiunque rilascia dichiarazioni mendaci è punito ai sensi del codice penale e delle leggi speciali in materia, ai sensi e per gli effetti di cui all'art. 47 D.P.R. n. 445/2000,

DICHIARA

che l'articolo "Exploring Breast Cancer Diagnosis with Fractal Analysis and Classifications Methods", pubblicato sulla rivista *Rendiconti del Circolo Matematico di Palermo*, serie II, suppl. 83 (pp. 229-236) come opera comune di tutti gli Autori, è a lui attribuibile solo per quanto riguarda il paragrafo 1.2 e il paragrafo 2 nella parte riguardante l'analisi di segmentazione.

DECLARATION

(Art. 47 D.P.R. December 28, 2000, n. 445)

The undersigned Author

d'Ovidio Francesco Domenico, born on **25/07/1955** in **Bari** (**BA**), resident in **Bari** (**BA**), via **Sardegna S.Spirito** n. **23**

aware that anyone who makes false statements is punishable under the Penal Code and special laws, pursuant to D.P.R. n. 445/2000 (Art. 47) and for its purposes,

DECLARES

that the article "Exploring Breast Cancer Diagnosis with Fractal Analysis and Classifications Methods", published in the *Rendiconti del Circolo Matematico di Palermo*, series II, suppl. 83 (pp. 229-236) as a common work of all the Authors, is entirely attributable to him only with regard to § 1.2 and to §2, in the section about the segmentation analysis.

Bari, November 15, 2012

Francesco Domenico d'Ovidio
