Mammogram segmentation by contour searching and massive lesions classification with Neural Network

F. Fauci [gh], S. Bagnasco [b], R. Bellotti [a], D. Cascio [g], S.C. Cheran [bc], G. De Nunzio [k], M. E. Fantacci [f], G. Forni [ij], A. Lauria [dj], E. Lopez Torres [l], R. Magro [gh], G. L. Masala [de], P. Oliva [de], M. Quarta [k], G. Raso [gh], A. Retico [f], S. Tangaro [a], E. Varrica [g]

a) Dipartimento di Fisica, Università di Bari and INFN Sezione di Bari, Italy.
b) INFN, Sezione di Torino, Italy.
c) Dipartimento di Informatica, Università di Torino, Italy.
d) Struttura Dipartimentale di Matematica e Fisica, Università di Sassari, Italy.
e) INFN, Sezione di Cagliari, Italy.
f) Dipartimento di Fisica, Università di Pisa, and INFN, Sezione di Pisa, Italy.
g) Dipartimento di Fisica e Tecnologie Relative, Università di Palermo, Italy.
h) INFN, Sezione di Catania, Italy.
i) Dipartimento di Scienze Fisiche, Università di Napoli, Italy.
j) INFN, Sezione di Napoli, Italy.
k) Dipartimento di Fisica, Università di Lecce and INFN, Sezione di Lecce, Italy.
l) CEADEN, Havana, Cuba.

Summary

Breast cancer is one of the most common kinds of cancer as well as the leading cause of mortality among women. It first appears as an asymptomatic lesion of the breast, then it may spread all over the organism. For this reason, an early diagnosis in asymptomatic women may greatly reduce the mortality. Mammography is the most effective procedure for an early diagnosis of the anomalies which could mark a tumour, even if, the detection of tumour in a mammographic image is a difficult task, due to the great number of non-pathological structures. In this paper, an algorithm for detecting massive lesions is presented. The database consists of 3762 digital images (1153 containing massive lesions), related to 1093 patients, acquired in several hospitals belonging to the MAGIC-5 collaboration; to the best of our knowledge, this is the greater mammographic database in Europe. The films are digitalized through a scanner with a resolution of 85 micron, and 12 bits per pixel which allows 4096 grey levels. All images are characterized by a full medical description, e.g. tissue, type of lesion, istotype and so on. The diagnosis is supported by an histological test, while non-pathological images correspond to patients with at least three years of follow-up.

Identification of massive lesions requires a dedicated strategy: shape and dimension of massive lesions are often irregular, the borders are ill-defined, thus making difficult the discrimination from parenchyma's structures. Due to the large size of the mammographic image, a reduction of the surface under investigation is desirable: this is achieved through segmentation of the whole image, without loss of meaningful information, by means of a ROI (Region Of Interest) Hunter algorithm, whose characteristics should be overall high efficiency (fraction of correctly identified lesion), minimum percentage area (fraction of selected area) and low false positive number.

The ROI Hunter algorithm here developed searches high intensity region contour: this goal is not trivial as the malignant mass often spreads into the surrounding tissue. Iso-intensity curves may be used to solve this problem once the threshold level is correctly set. The threshold depends on any parameters which may be tuned to obtain the best performance. In this way, the algorithm selects limited regions including peculiar characteristics as, for example, structures with a high density at the centre, sloping down with the distance (see Fig. 1).

Detection of these regions appreciably reduces the area under analysis in the following classification step.

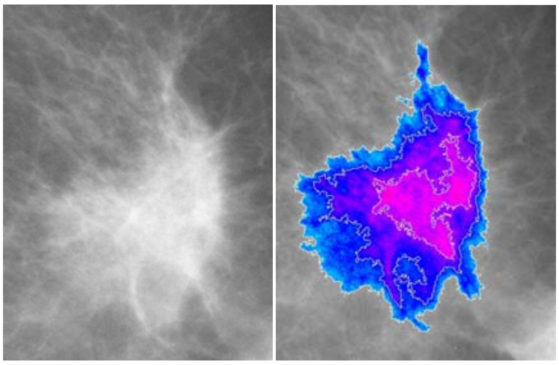

Fig.1

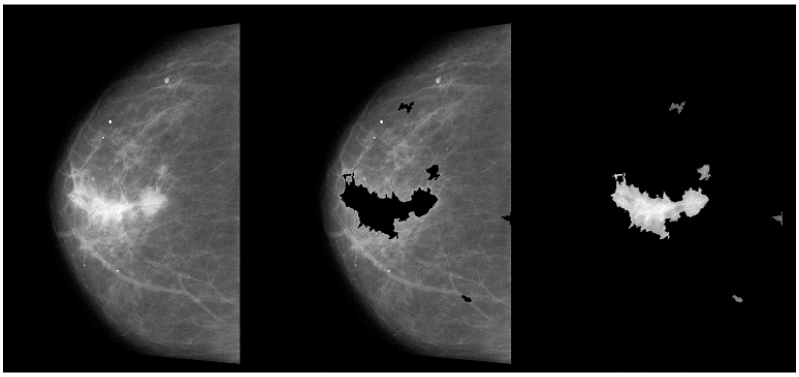

Fig. 2

As an example, the Fig. 2 displays the original image (left), the selected ROI (right), and the image without the ROI (middle). The performance of the overall ROI Hunter is:

- efficiency = 91 %
- covered area = 18 %
- FP per image = 7

Features extraction plays a fundamental role in many pattern recognition tasks. Some features give geometrical information as eccentricity, area and mean radius; others provide shape parameters as fractal dimension, entropy and inertial momentum; moreover, features can be extracted from the grey level co-occurrence matrix (GLCM) relative to the selected ROI. As the name suggests, the GLCM is constructed from the image by estimating the pair-wise statistics of pixel intensity, thus relying on the assumption that the texture content information of an image is contained in overall or average spatial relationship between pairs of pixel intensities. Textural features can be derived from the GLCM and used for classification purpose in place of the single GLCM elements. These features refer to textural property of the ROI such as homogeneity, contrast, presence of organized structure, complexity and nature of grey tone transitions.

Once the features are extracted from each image pattern (ROI), they are used as input to a feed forward back propagation supervised neural network with momentum. The number of hidden neurons was varied in order to obtain the best performance. The output neuron provides the probability that the ROI is pathological. The whole sample (1100 pathological patterns and 9500 healthy patterns) was divided in three subset: learning, validation and test sample.

Fig. 3 shows the ROC curve on the test sample after 500 learning epochs. As known, the area under ROC curve (A_Z) may be used to evaluate the overall performance of a diagnostic system. In our case we obtain $A_Z = (85.6 \pm 0.8)$ %.

This software is included in the CAD station working in the hospitals belonging to the MAGIC-5 collaboration and it is actually under the test of the radiologists.

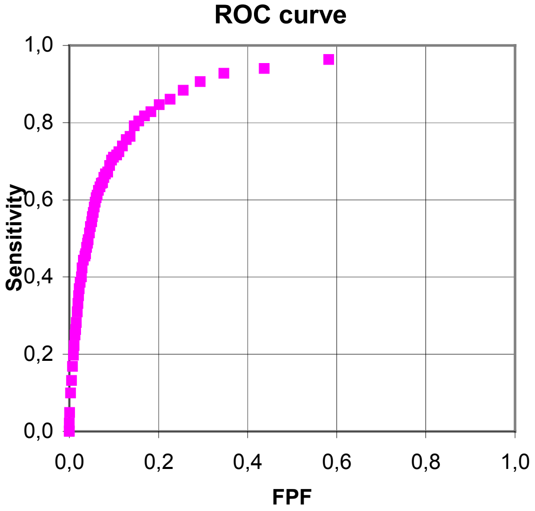

Fig. 3