International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:1, No:12, 2007

Massive Lesions Classification using Features based on Morphological Lesion Differences

U. Bottigli, D.Cascio, F. Fauci, B. Golosio, R. Magro, G.L. Masala, P. Oliva, G. Raso, and S.Stumbo

***Abstract*—**Purpose of this work is the development of an automatic classification system which could be useful for radiologists in the investigation of breast cancer. The software has been designed in the framework of the MAGIC-5 collaboration.

In the automatic classification system the suspicious regions with high probability to include a lesion are extracted from the image as regions of interest (ROIs). Each ROI is characterized by some features based on morphological lesion differences.

Some classifiers as a Feed Forward Neural Network, a K-Nearest Neighbours and a Support Vector Machine are used to distinguish the pathological records from the healthy ones.

The results obtained in terms of sensitivity (percentage of pathological ROIs correctly classified) and specificity (percentage of non-pathological ROIs correctly classified) will be presented through the Receive Operating Characteristic curve (ROC). In particular the best performances are 88% ± 1 of area under ROC curve obtained with the Feed Forward Neural Network.

***Keywords*—**Neural Networks, K-Nearest Neighbours, Support Vector Machine, Computer Aided Diagnosis.

I. Introduction

BREAST cancer is reported as one of the first causes of women mortality [1] and an early diagnosis in asymptomatic women makes it possible to reduce the breast cancer mortality: in spite of a growing number of detected cancers, the death rate for this pathology decreased during the last 10 years[2], thanks to the screening programs and the relative early diagnosis[3].

With the project MAGIC-5 (Medical Application on Grid Infrastructure Connection), a collaboration among Italian physicists and radiologists, it was possible to built a large database of digitized mammographic images; all these images was used to develop a CAD (Computer Aided Diagnosis) for medical applications, such as breast cancer detection through mammographic images and lung cancer detection by Computed Tomography (CT) imaging modality. This collaboration group has developed an integrated station that is available either to digitize the analogical images or to archive or to perform statistical analysis. Furthermore this prototype of station can represent also a very good system for mammographic educational programs. With a GRID configuration it would be possible for the clinicians tele- and co-working in new and innovative groupings. Using the whole database, several analysis can be performed by the MAGIC-5 tools.

The mammographic images (18x24 cm^2, digitized by a CCD linear scanner with a 85 μm pitch and 4096 gray levels) are fully characterized: pathological ones have a consistent description which includes the radiological diagnosis and the histological data, while non pathological ones correspond to patients with a follow up of at least three years [4]. The focus is the automated analysis of massive lesions, i.e. the search for 'large objects' in the image, usually characterized by peculiar shapes. The detection is made by extracting features based on morphological lesion differences.

In this work we report the results obtained with some classifiers as a Feed Forward Neural Network, a K-Nearest Neighbours and a Support Vector Machine[6]-[14], used to select from the region of interest (ROI) the pathological massive lesion.

II. Methods

The CAD system here presented is an expert system based on three steps : a ROI-hunter, a features extractor module and a classifier.

The ROI-hunter was already described in ref. [5]. The aim of this stage is to reduce the data amount to process by searching for Regions Of Interest (ROIs) that include a lesion with high probability. Only selected regions are stored for the next processing steps, rather than the whole mammogram as shown in Fig. 1.

G.L. Masala is with the Struttura Dipartimentale di Matematica e Fisica dell'Università di Sassari and Sezione INFN di Cagliari, Italy, Via Vienna 2, Sassari, 07100, Italy (corresponding author; phone:+39079229486; fax: +39079229482; e-mail: giovanni.masala@ca.infn.it).

U. Bottigli, was with the Struttura Dipartimentale di Matematica e Fisica dell'Università di Sassari and Sezione INFN di Cagliari. He is now with the Dipartimento di Fisica dell'Università di Siena and Sezione INFN di Cagliari, Italy (e-mail: bottigli@uniss.it).

B. Golosio, P. Oliva, S. Stumbo are with the Struttura Dipartimentale di Matematica e Fisica dell'Università di Sassari and Sezione INFN di Cagliari, Italy (e-mail: golosio@uniss.it; oliva@uniss.it; stumbo@uniss.it).

D. Cascio, F.Fauci, R.Magro and G. Raso are with Dip. di Fisica e Tecnologie Relative, Università di Palermo, Palermo, Italy and INFN, Sezione di Catania, Catania, Italy (e-mail: donato@difter.unipa.it, ffauci@mail.unipa.it, rmagro@unipa.it, graso@unipa.it).

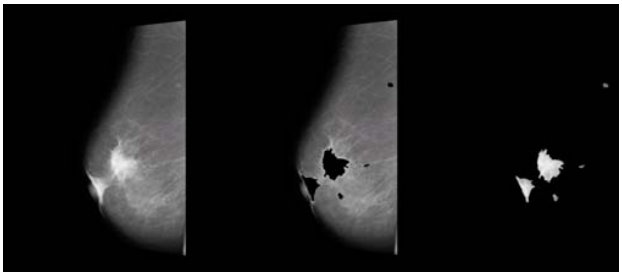

Fig. 1 The original mammogram (left), the remaining image (middle), the selected patterns containing the ROIs (right)

The Features Extractor Module

In this paper twelve features are extracted from the segmented masses. The criteria for the features selection are based on morphological lesion differences [15]-[20]. For example the excessive lengthening is often symptom of pathology absence. In Table I the complete extracted features list is reported.

TABLE I
MORPHOLOGICAL FEATURES EXTRACTED FROM THE REGION OF INTEREST

Fractal index	Area
Eccentricity	Contour Gradient Entropy
Average Intensity	Standard Deviation of Intensity
Average Radial Length (ARL)	Standard Deviation of ARL
Entropy of intensity distribution	Anisotropy
Inertial Momentum	Circularity

The Features extraction [5]-[6] plays a fundamental role in many pattern recognition tasks. Some features give geometrical information as eccentricity, area and average radial length; others provide shape parameters as fractal index and inertial momentum. In order to verify the feature discrimination capability between the two classes (pathologic or healthy patients), the feature value histograms are drawn. As an example in Fig. 2, 3, and 4 the histograms of average radial length, entropy of intensity distribution and circularity are shown.

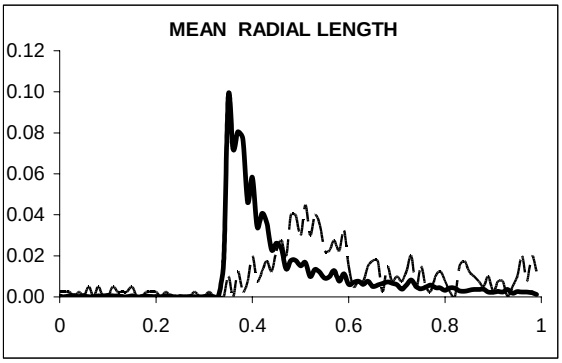

Fig. 2 Feature "Mean radial lenght" distribution for pathological (dashed) and healthy (continuous) ROIs

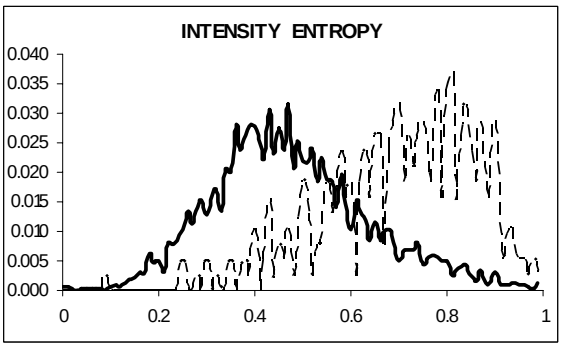

Fig. 3 Feature "Entropy of intensity" distribution for pathological (dashed) and healthy (continuous) ROIs

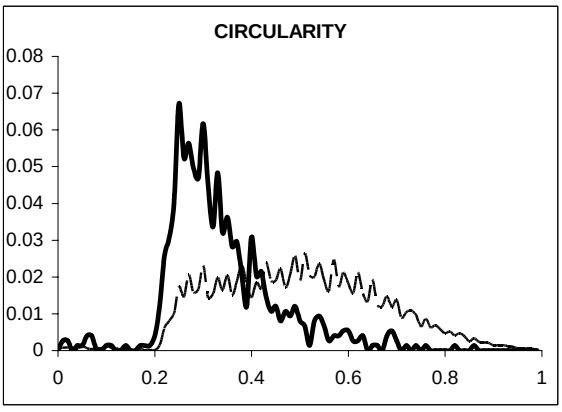

Fig. 4 Feature "Circularity" distribution for pathological (dashed) and healthy (continuous) ROIs

More dataset extracted from the CALMA database [4] are shown in the Table II. It is illustrated the composition (positive samples vs total samples) of the training set, validation set and testing set.

TABLE II
COMPOSITION OF THE MAMMOGRAPICH DATASET

Dataset	# of samples (ROIs)	# of positive samples (ROIs)
Training set	4230	318
Validation set	4230	315
Testing set	4230	320

The Classifier

We make a comparative study of the following classifiers:

→A K-Nearest Neighbors (K-NN) classifier. For this type of deterministic classifier, it is necessary to have a training set which is not too small, and a good discriminating distance. KNN performs well in multi-class simultaneous problem solving. There exists an optimal choice for the value of the parameter K, which brings to the best performance of the classifier. This value of K is often approximately close to $N^{1/2}$.

→A Multi Layer Perceptron (MLP). The selected MLP is a feed-forward back-propagation supervised neural network trained with gradient descent learning rule with "momentum", so as to quickly move along the direction of decreasing gradient, thus avoiding oscillations around secondary minima.

→ the SVM algorithm creates a hyperplane that separates the data into two classes with the maximum-margin. Given training examples labeled either "yes" or "no", a maximum-margin hyperplane is identified which splits the "yes" from the "no" training examples, such that the distance between the hyperplane and the closest examples (the margin) is maximized. There is way to create non-linear classifiers by applying the kernel trick to maximum-margin hyperplanes. The resulting algorithm is formally similar, except that every dot product is replaced by a non-linear kernel function. This allows the algorithm to fit the maximum-margin hyperplane in the transformed feature space. The transformation may be non-linear and the transformed space high dimensional; thus though the classifier is a hyperplane in the high-dimensional feature space it may be non-linear in the original input space.

III. RESULTS

Using sensitivity (percentage of pathologic ROIs correctly classified) and specificity (percentage of non pathologic ROIs correctly classified), the results obtained with this analysis are described in terms of the ROC (Receiver Operating Characteristic) curve[21]-[22], which shows the true positive fraction (sensitivity), as a function of the false positive fraction (1-specificity) obtained varying the threshold level of the ROI selection procedure. In this way, the ROC curve produced allows the radiologist to detect massive lesions with predictable performance, so that he can set the CAD sensitivity value.

The results of the K-Nearest Neighbours, the Support Vector Machine and the Multi Layer Perceptron (MLP) are supplied in the diagram through the ROC curve calculated on the testing set after the optimization on the training set-validation set of the classifiers. The KNN is optimized for K = 21; as shown in Fig. 5.

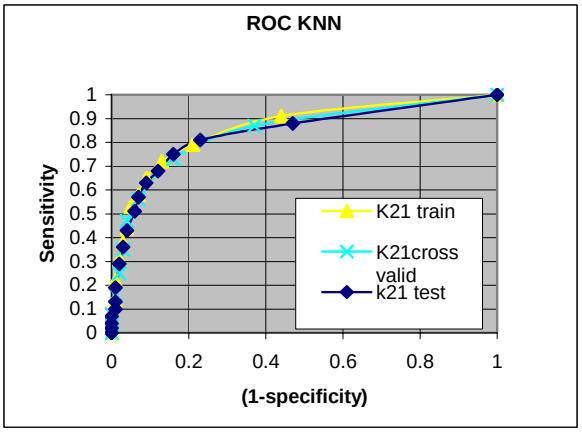

Fig. 5 ROC curve about KNN in normal training and cross validation on validation set and in normal learning on testing set

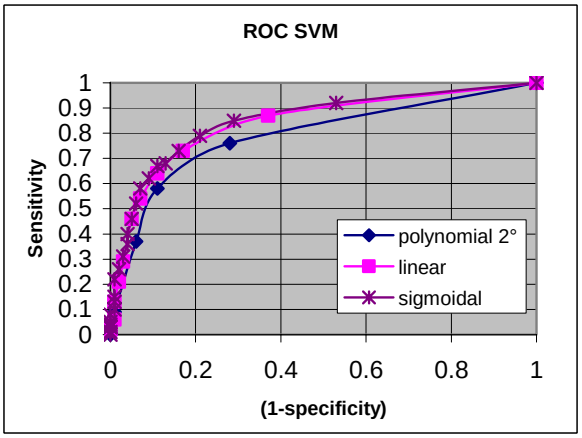

Fig. 6 ROC curve about SVM with polynomial, linear and sigmoidal kernel on the validation set

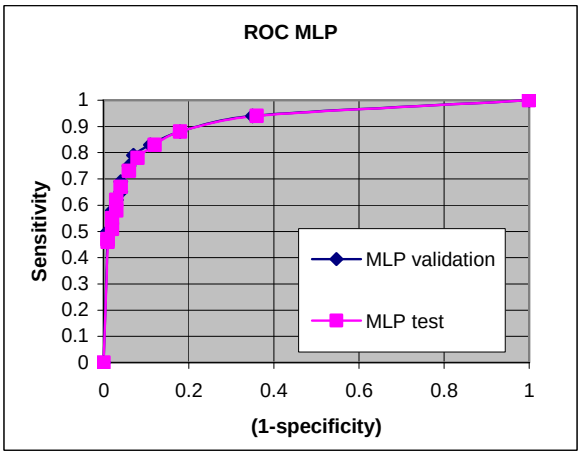

Fig. 7 ROC curve about MLP on validation set and testing set

The SVM supplies the best performances on the testing set for a sigmoidal kernel as shown in Fig. 7, while the best neural net MLP has 12 hidden neurons for the 12 input previously described.

Finally in the plot of Fig. 8 are shown the results of the classifiers on the testing set.

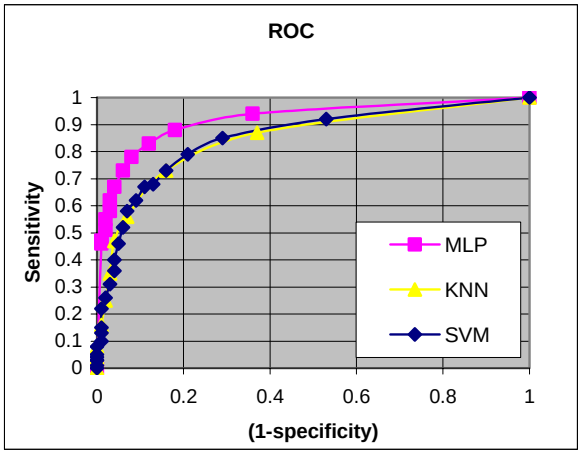

Fig. 8 Comparative ROC curves about MLP, KNN, SVM on testing set

Also the area under the curve [21]-[22], obtained in relation to the same ROC curves calculated on the test values, are reported below in Table III.

TABLE III
PERFORMANCE OF THE CLASSIFIERS IN TERMS OF AREA UNDER THE ROC CURVE S

Classifiers	Area under ROC curves	Error
KNN	81 %	± 1
SVM	81 %	± 1
MLP	88 %	± 1

The results of the Table III show that the neural net has better performances then the other classifiers for the dataset considered. A study carried out on the complementariness of the classifiers used on the dataset in examination show (in the case of the features previously used), the regions of decision of the three classifiers overlap. This fact and the better performance in comparison to the other two classifiers, indicates that it is not possible in this case to combine the output of the various classifiers with techniques of multi classification system (MCS) to improve the total performances.

IV. CONCLUSION

In this paper a comparison of some classification system for massive lesion classification has been presented. The features are extracted trough an algorithm based on morphological lesion differences. The features are used for the discrimination between the two classes (pathological or healthy ROIs). The discriminating performances of the algorithm were checked by means of a supervised neural network against other classifiers and the results have been presented in terms of ROC curve. The results are comparable or better than those obtained in other recent studies [5],[23]-[24] verifying that the new representation applied provides a better ability to distinguish pathological ROIs from the healthy ones.

REFERENCES

[1] Smith R.A., "Epidemiology of breast cancer", in "A categorical course in physics. Imaging considerations and medical physics responsibilities", Madison, Winsconsin, Medical Physics Publishing, 1991.
[2] Peto R., Boreham J., Clarke m., Davies c., Deral V, correspondence "UK and USA Breast cancer deaths down 25% in year 2000 at ages 20-69 years", LANCET 2000, 355, (9217) pp. 1822-1823, 2000.
[3] Bird R., Wallace T., Yankaskas B., "Analysis of cancer missed at screening mammography", Radiology 1992: 184; pp.613-617, 1992.
[4] Bottigli U, Delogu P, Fantacci ME, Fauci F, Golosio B, Lauria A, Palmiero R, Raso G, Stumbo S, Tangaro S Search of Microcalcification clusters with the CALMA CAD station. The International Society for Optical Engineering (SPIE) 4684: 1301-1310, 2002
[5] F. Fauci, S. Bagnasco, R. Bellotti, D. Cascio, S. C. Cheran, F. De Carlo, G. De Nunzio, M. E. Fantacci, G. Forni, A. Lauria, E.Lopez Torres, R. Magro, G. L. Masala, P.Oliva, M. Quarta, G. Raso, A. Retico, S.Tangaro, Mammogram Segmentation by Contour Searching and Massive Lesion Classification with Neural Network, Proc. IEEE Medical Imaging Conference, October 16-22 2004, Rome, Italy; M2-373/1-5, 2004.
[6] O. Duda, P. E. Hart, D. G. Stark, "Pattern Classification", second edition, A Wiley-Interscience Publication John Wiley & Sons, 2001.
[7] S. J. Russel, P.Norvig, "Artificial Intelligence. A modern approach", UTET, 1998.
[8] S. Haykin "Neural Networks – A comprehensive foundation", second edition, Prentice Hall, 1999.
[9] V. N. Vapnik. "Statistical Learning Theory. Wiley", New York , 1998.
[10] M. Pontil, A. Verri "Properties of Support Vector Machines", Neural Computation, Vol. 10, pp 955-974, 1998.
[11] N. Cristianini, J. Shave-Taylor. "An Introduction to Support Vector Machine"(and other kernel-based learning methods). Cambridge University Press 2000.
[12] SVM_light software is available in the following location : ftp://ftp-ai.cs.unidortmund.de/pub/Users/thorsten/svm_light/current/ svm_light.tar.gz
[13] T. Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features, Proc. 10th European Conf. Machine Learning (ECML), Springer-Verlag, 1998.
[14] T. Mitchell "Machine Learning" , McGraw-Hill 1997.
[15] Timp S., Karssemeijer N., A new 2D segmentation method based on dynamic programming applied to computer aided detection in mammography, Medical Physics: 31; 958-971, 2004.
[16] Baydush A.H., Catarious D.M., Abbey C.K., Floyd C.E., Computer aided detection of masses in mammography using subregion Hotelling observers, Medical Physics: 30; 1781-1787, 2003.
[17] Tourassi G.D., Vargas-Voracek R., Catarious D.M. Jr, Floyd C.E. Jr, Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information, Medical Physics: 30 (8); 2123-2130, 2003.
[18] Antonie M.L., Zaiane O.R., Coman A., Application of data mining techniques for medical image classification, Proc. of II Int. Work. On Multimedia Data Mining, USA, 2001.
[19] Vyborny CJ., Giger ML., Computer vision and artificial intelligence in mammography, AJR: 162; 699-708, 1994.
[20] Lai S., Li X., Bischof W., On techniques for detecting circumscribed masses in mammograms", IEEE Transaction on Medical Imaging: 8(4); 377-386, 1989.
[21] Hanley JA, McNeil B, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology: 143; 29-36, 1982.

[22] Hanley JA, McNeil B, A method of comparing the areas under receiver operating characteristic curves derived from the same cases, Radiology: 148; 839-843, 1983.
[23] U. Bottigli, B. Golosio, G. L. Masala, P. Oliva, S. Stumbo, D. Cascio, F. Fauci, R. Magro, G. Raso, R. Bellotti, F. De Carlo, S.Tangaro, I. De Mitri, G. De Nunzio, M. Quarta, A. Preite Martinez, P. Cerello, S. C. Cheran, E.Lopez Torres "Dissimilarity Application for Medical Imaging Classification" on proceedings of The 9th World Multi-Conference on Systemics, Cybernetics and Informatics WMSCI 2005, Orlando 10-13 July 2005, vol III pag 258-262, 2005.
[24] G. Masala, B. Golosio, D. Cascio, F. Fauci, S. Tangaro, M. Quarta, S. C Cheran, E. L. Torres, "Classifiers trained on dissimilarity representation of medical pattern: a comparative study" on Nuovo Cimento C, Vol 028, Issue 06, pp 905-912 , 2005.