A SIMPLE COMPARISON OF DYNAMIC CRITERIA FOR BREAST MRI CLASSIFICATION

F.A. Cardillo, A. Starita, D. Caramella, A. Cilotti, F. Odoguardi

University of Pisa, Italy

Abstract: Contrast-Enhanced magnetic resonance of the breast (CE-MRI) is a useful adjunct to standard breast examinations. A typical CE-MRI examination of the breast produces hundreds of images, that need to be analyzed one by one. The complete analysis of a single dataset requires a long time to complete and is an error-prone process due to the errors caused by the fatigue and the habituation of the radiologist. Computer Assisted Diagnosis (CAD) could help the clinicians in the analysis of such big datasets. In this paper we present an experimentation of simple classifiers for the classification of signal intensity curves extracted from CE-MRI images. We used 60 examinations with histological confirmed results collected by the Diagnostic and Interventional Radiology Department of the University of Pisa. From this data, we extracted 1800 patterns and used them to train and test three different classifiers using a 10-fold crossvalidation: two threshold based ones and a multilayer feedforward neural network. The result show that such algorithms are able to reach accuracies greater than 90%. Quite surprisingly, we found that a very simple threshold-based classifier reaches the best accuracy, even greater than those of the neural network and of a simple ensemble built using a voting scheme.

INTRODUCTION

Magnetic Resonance Imaging (MRI) of the breast is a useful complementary technique to standard breast imaging techniques, such as, for example, x-ray mammography, when specific clinical indications exist. As discussed in Heywang-Köbrunner et al [1], breast MRI is often performed in order to eliminate the ambiguity from uncertain mammographic findings. However, its role is not restricted to that function: MRI is performed to identify the extent and the multifocality of detected lesions, to evaluate the post-operative follow-up, to study the dense breasts of young women. Even if it provides images with a resolution lower than x-ray, the three-dimensional dataset can give precious hints in establishing exactly the location of the lesions. CE-MRI of the breast has been a very controversial theme, but, recently, its importance has been recognized: as stated in Takeda [2], the diagnosis of breast cancer has progressed owing much to the improvement in the breast MRI examination.

However, plain MRI is not useful for detecting breast cancers. In order to be effective, a contrast agent, typically a Gadolinium compound, must be injected in the body of the patient. The basis of contrast-enhanced MRI (CE-MRI) is the fact that tumors enhance and enhance more than normal tissues. Since tumors need many blood vessels to grow, the concentration of the contrast agent at their location will be higher than in surrounding tissues and they will consequently appear as brighter areas in the images. Even if there is not a standard diagnostic protocol, consult, for example, Kuhl and Schild [3], dynamic and morphological patterns related, respectively, to the diffusion of the contrast agent and to the shape, edges, or the internal pattern of the enhancing region, are usually able to discriminate among different types of lesion. However, there is a considerable overlap between dynamic and morphological patterns of benign and malignant lesions. A frequenty used scheme has been proposed by Fischer et al. [4]. In this paper, however, we will study dynamic features only.

A CE-MRI requires the acquisition of one series of images before the injection of the contrast agent, called pre-contrast series, and of several series of images, after the injection, called post-contrast series. A CE-MRI examinations produces hundreds of images, which need to be analyzed one by one. The role of a Computer-Assisted Diagnosis (CAD) tool would be important since it could provide a second-opinion to the radiologist, helping him to reduce the number of errors caused by the fatigue and habituation and providing him a second opinion to take into account in difficult cases. Furthermore, computer algorithms can be used to exploit and combine features, which are not directly human readable, reaching higher level of accuracy.

In this paper we present the results of a simple experimentation of classifiers applied to the problem of learning and classification of dynamic patterns extracted from CE-MRI examinations.

DIAGNOSTIC PROTOCOL AND DYNAMIC CRITERIA

In order to establish the presence of lesions, either malignant or benign, each image in the study must be analyzed using two different sets of criteria: morphological and dynamic. The first step of the interpretation is the search for enhancement in subtracted images, i.e. images obtained by post-contrast images by subtracting their corresponding pre-contrast images. Due to the overlap between benign and malignant lesions in their dynamic behaviour and their morphological appearance, an enhancing region must be classified according to dynamic and morphological criteria. Anyway, in most cases it is possible to make a correct diagnosis using only dynamic criteria. Morphological criteria, described by Nunes [5], that are basically the same as in other examinations, concern the shape of the lesion and its borders. In CE-MRI the set of signs that can be studied is enriched by the diffusion of the contrast agent in the tissues.

In our work we considered only dynamic criteria, re-

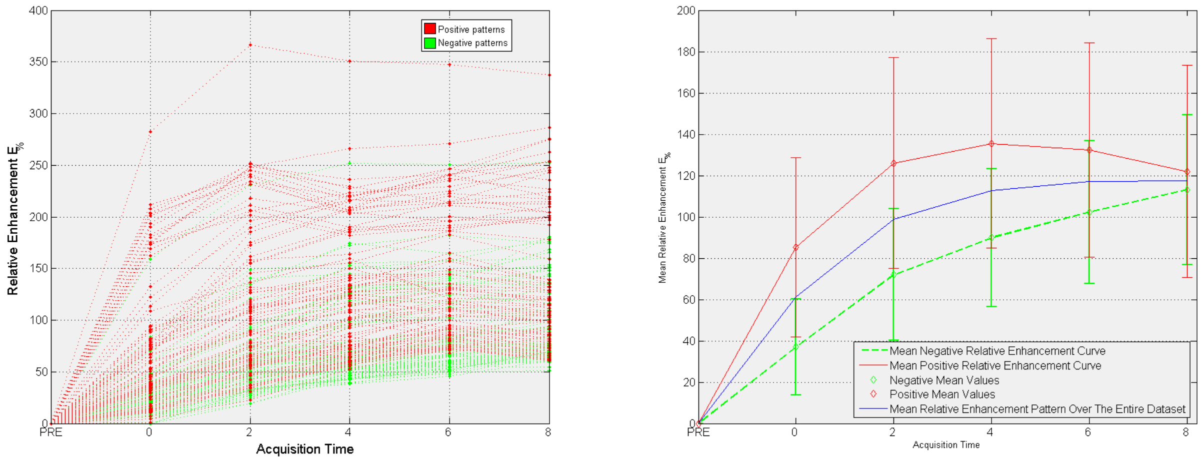

Figure 2: Left: Curves extracted from our datasets. The picture contains 90 negative relative enhancement curves (green) and 90 positive curves (red). Right: Mean relative enhancement curves computed on our dataset (1800 patterns, half positive and half negative) with standard deviation. The red curve is the mean computed over malignant patterns, while the green one is computed on negative patterns. The blue curve is the mean relative enhancement pattern computed over the entire dataset.

lated to the diffusion of the contrast agent in the tissues. There is not a common and accepted standard of diagnostic critetia, but there is agreement on the fact that cancers more often show early strong enhancement with rapid washout, while benign lesions show a slowly rising and persistent signal intensity curve. Szabó et al. [6].

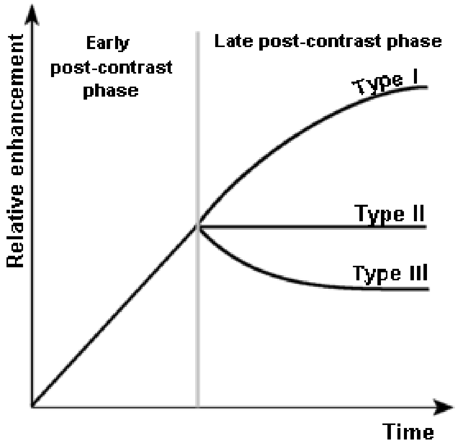

Figure 1: Left: Abstract classification scheme. Type I curves are usually found in benign lesions. Type III found are considered a strong sign of malignancy. Type II curves can be found in both types of lesions and do not have a clear label.

The most important dynamic criterion is the relative-enhancement, introduced in Kaiser & Zeitler [7]. The relative-enhancement represents the increase in signal intensity relative to the pre-contrast phase:

$$E_k^{\%} = \frac{\mathcal{I}_k(\mathcal{R}) - \mathcal{I}_0(\mathcal{R})}{\mathcal{I}_0(\mathcal{R})} \cdot 100 \; k \in [1, n_{series}) \qquad (1)$$

where $\mathcal{I}_k$ is the examined image, k is the series index, $\mathcal{I}_0$ is the corresponding pre-contrast image, $\mathcal{R}$ is a user-selected region of interest, usually nine voxels wide. In our experimentation, $\mathcal{R}$ correspond to a 3×3 square window centered on the voxel selected by the user, the value $\mathcal{I}_i(\mathcal{R})$ is simply the arithmetic mean of the intensities of the voxels in $\mathcal{R}$. The values computed according to equation (1) are then interpolated to construct relative-enhancement curves. The curves are classified in three classes according to their behaviour:

Type I a kinetic behaviour with a persistent uptake is considered a sign of benignity. Curves extracted from benign lesions are often in this class;

Type II a kinetic behaviour with a clear plateau phase can be found both in benign and in malignant features. The classification of Type II curves is thus unknown;

Type III a kinetic behaviour with a strong uptake followed by a rapid washout is considered a sign of malignancy. Type III are, in fact, often found in malignant lesions.

The three prototypes that represent the three classes of curves are shown in Fig. 1. The distinction among the three types of curve is not always clear: there is a significant overlap among curves extracted from benign and malignant lesions. Furthermore, from the same lesion, it is possible to extract type I or type III curves. Some real curves, extracted from the examinations we have collected, are shown in the left picture of Fig. 2. However, the mean relative enhancement curves, computed over positive and negative patterns, plotted in the right picture of Fig. 2, confirm the ideal classification scheme of Fig. 1.

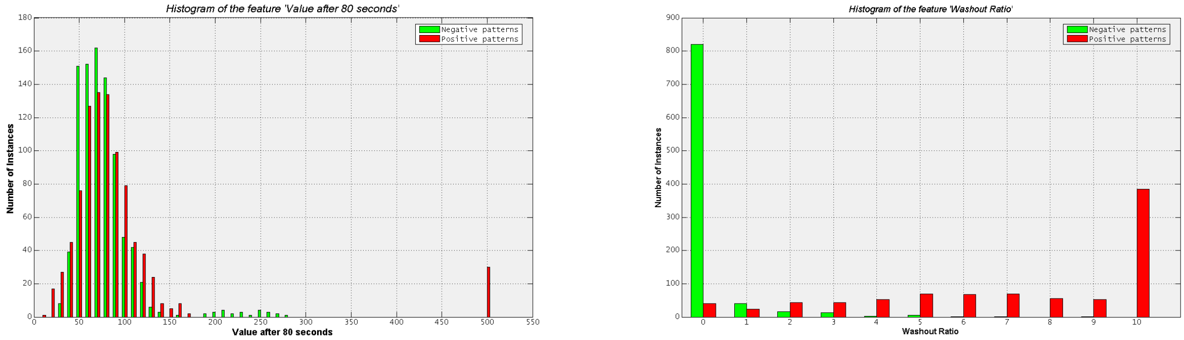

Figure 3: Left: Histogram of the values of the feature ’Values after 80 seconds’. Right: Histogram of the values of the feature ’Washout Ratio’

The mean curve of positive patterns has a strong washout phase, while the mean curve of negative is a steady increasing curve.

The second dynamic criterion we used is the absolute value of the signal intensity at t second after the injection. In our case $t = 80$s.

The third criterion is the initial slope, Szabó [6], defined as:

$$Slope_i = \frac{E_{peak}}{T_{peak}} \tag{2}$$

where E_{peak} is the maximum relative enhancement and T_{peak} is the time elapsed till E_{peak} is reached.

The fourth criterion is the washout-ratio, described in Ikeda et al. [8]:

$$W_{peak-k} = \frac{\mathcal{I}_{peak}(\mathcal{R}) - \mathcal{I}_k(\mathcal{R})}{\mathcal{I}_{peak}(\mathcal{R})} \cdot 100 \tag{3}$$

where $\mathcal{R}$ is again a user-selected region of interest. In our experimentation, where $k \in [0, 6)$, we computed and used W_{peak-5}, that is, we computed the washout-ratio between the maximum signal intensity and the signal intensity in the last post-contrast image, acquired at eight minutes after the contrast injection.

MATERIAL AND METHODS

The dMRI examinations used in our study were collected by the Diagnostic and Interventional Radiology Department of the University of Pisa. In such department, the dMRI is performed on women with uncertain mammographic findings, on young women and on women who present an high risk for cancer. The datasets were acquired on a General Electric 1.5T Signa Contour scanner using 3D fast spoiled gradient echo sequences (FSPGR) with 12.7 ms repetition time, 2.5 echo time and 30° flip angle. Each examination is composed by size series of images: one pre-contrast series, i.e. a series depicting the breast before injecting the agent, and five post-contrast series, acquired at 0, 2, 4, 6, and 8 minutes after the injection of the contrast agent. The images in each series have 256×256 voxels, each voxel being 1.5mm $\times 1.5$mm $\times 3$mm in dimension. There is no gap between successive and adjacent images. The number of images in the datasets depends on the dimension of the volume that needs to be acquired. The gray level of each voxel is stored using 9 bits, providing a colour depth equal to $2^9 = 512$.

Among the datasets that we have collected, we used 60 examinations with MR diagnosis, i.e. the diagnosis made by the radiologist during reading, confirmed by a histological examination. 30 examinations contain malignant tumors, the remaining 30 contain benign tumors. For each examination, we extracted data from 30 points located inside the lesions. The dataset used in the experimentations described in the next section is thus composed by $30 \cdot 30 = 900$ benign patterns and $30 \cdot 30 = 900$ malignant patterns, for a total number of patterns equal to 1800. Each pattern is labeled according to the label assigned by the histological examination to the lesion that it is extracted from. In the following, we will refer to malignant patterns as positive, and to benign ones as negative (even if they have been extracted from a benign pathological lesion). We did not include patterns from normal tissues and from blood vessels. Normal tissues can be easily filtered out by applying a threshold, blood vessels have a

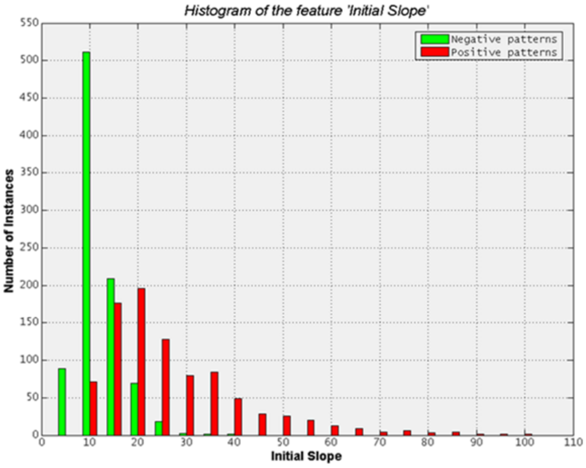

Figure 4: Histogram of the values of the feature ’Initial Slope’

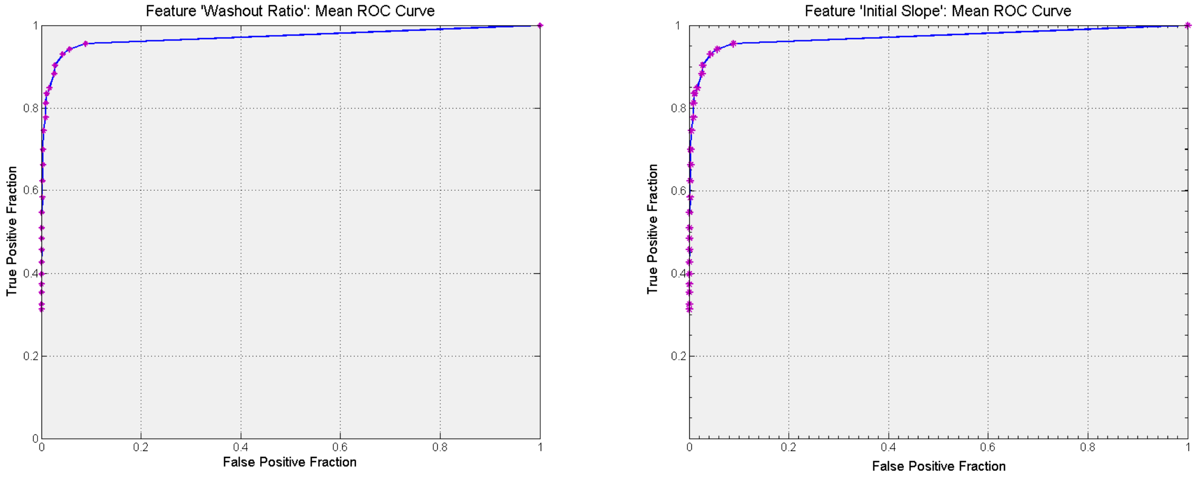

Figure 5: Mean ROC curves of the threshold-based classifier. Left: Classifier applied to the ‘Washout Ratio’ feature. Right: Classifier applied to the ’Initial Slope’ feature.

chaotic dynamic behavior, with very irregular curves, and can be easily detected.

The experimentation has been performed using a k-crossvalidation procedure: the patterns are split into k folders and each classifier is trained k times, for each iteration, $k - 1$ folders are used as training set and 1 folder as test set. In our experiments we used $k = 10$ folders, each folder containing 180 patterns from size examinations, $3 \cdot 30 = 90$ extracted from three negative examinations and $3 \cdot 30 = 90$ extracted from three positive examinations. Each examination is included in a single folder and is used either as part of the training or part of the test. The examinations included in each folder were chosen randomly.

From the left picture of Fig. 2, the overlap between malignant and benign patterns is manifest. The mean relative enhancement curves for positive and negative patterns, shown in the right picture of Fig. 2, respect the abstract classification scheme proposed in the left picture of the same figure. It is evident how negative relative enhancement patterns are basically steady, always increasing curves, while positive ones are characterized by a strong washout phase.

The feature ’Value after 80 seconds’, even if it is included in the Fischer scheme for lesion classification, does not provide a good classification of our dataset. The histogram plotted in the left picture or Fig. 3 shows that malignant and benign patterns are concentrated in the same region. The criterion could be useful in the classification only when the feature has a value greater than $\theta = 300$, since no negative patterns assume values greater than θ. However, we decided to exclude such feature by the current experimentation.

The feature ’Washout Ratio’ provides a good discrimination between malignant and benign patterns. In fact, negative patterns are characterized by a very low washout value, while positive patterns have a greater washout ratio. The mean value over positive patterns is $wr_+ = 10.26$ with a standard deviation $\sigma_{wr_+} = 8.29$, while the mean value over negative patterns is $wr_- = 0.18$ with a standard deviation $\sigma_{wr_-} = 0.73$. The washout ratio criterion, as it will be discussed later, is the criterion that best classifies our dataset.

The feature ’Initial Slope’, whose histogram is plotted in the right picture of Fig. 3, provides a good discrimination, even if negative and positive patterns are not clearly separated as in the washout-ratio. In this case, the mean value over positive patterns is $is_+ = 25.05$ with a standard deviation $\sigma_{is_+} = 14.73$, while the mean and the standard deviation over negative patterns are, respectively, $is_- = 11.85$ and $\sigma_{is_-} = 4.22$.

In order to classify the relative enhancement curves, trying to distinguish among the three different types, we tested a multilayer feedforward neural network, trained with the backpropagation algorithm. The network is a three-layer network, with five input units, seven hidden nodes and one output unit. The units have a log-sigmoid transfer function. The network architecture was chosen according to the results of previous experimentations and several tests on the current dataset. The neural network was trained for 10000 epochs. Before using the relative-enhancement curves as training patterns, they need to be normalized. We normalized them in the $[0, 1]$ range; the output of the network is in $[0, 1]$. It should be noted that the normalization implies a great loss of information: low enhancing curves and strong enhancing curves can be mapped onto the same normalized curve, loosing all the information about the absolute values of the relative enhancement. The backpropagation neural network is used by other research groups to classify similar data, as, for example, in Lucht et al. [9] and in Szabó et al. [10].

The features ’initial slope’ and ’washout ratio’ were classified training a simple threshold-based classifier. Basically, for each iteration of the crossvalidation, the accuracy of the classifier was computed for each iteration. The threshold used in the classification of the test set was the threshold allowing the classifier to reach the best ac-

Accuracy of \ on	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	Mean
BP on R.E.	*79.44*	92.78	92.78	90.56	96.11	**99.44**	87.22	90.56	97.22	97.22	92.33
Threshold on IS	71.11	84.44	87.22	*66.67*	83.89	**98.33**	75.56	88.89	81.11	86.11	82.33
Threshold on WR	*84.44*	97.22	96.67	92.78	**98.89**	97.22	90.00	92.78	97.78	94.44	94.22
Ensemble (Voting)	*80.0*	95.00	95.56	91.11	98.89	**99.44**	88.33	92.78	97.22	93.89	93.22

Sensitivity of \ on	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	Mean
BP on R.E.	*68.89*	97.78	90.00	96.67	**100**	**100**	78.89	94.44	95.56	98.89	92.11
Threshold on IS	*58.89*	95.56	95.56	67.78	82.22	**100**	67.78	94.44	63.33	85.56	81.11
Threshold on WR	*77.78*	97.78	96.67	94.44	**100**	97.78	84.44	95.56	97.78	90.00	93.22
Ensemble (Voting)	*68.89*	97.78	95.56	96.67	**100**	**100**	81.11	95.56	95.56	91.11	92.22

Specificity of \ on	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10	Mean
BP on R.E.	90.00	87.78	95.56	*84.44*	92.22	**98.89**	95.56	86.67	**98.89**	95.56	92.56
Threshold on IS	83.33	73.33	78.89	*65.56*	85.56	96.67	83.33	83.33	**98.89**	86.67	83.56
Threshold on WR	91.11	96.67	96.67	91.11	97.78	96.67	95.56	*90.00*	97.78	**98.89**	95.22
Ensemble (Voting)	91.11	92.22	95.56	*85.56*	97.78	**98.89**	95.56	90.00	**98.89**	96.67	94.22

Table 1: Accuracy (%) of each classifier on the ten iterations of the 10-fold crossvalidation. First row: backpropagation trained to classify normalized relative enhancement curves. Second row: threshold-based classifier trained on 'Initial Slope' values. Third row: threshold-based classifier trained on 'Washout-ratios' values. For each row the last columns contains the mean accuracy value over the 10 iterations.
Second table: Sensitivity. Third table: Specificity.

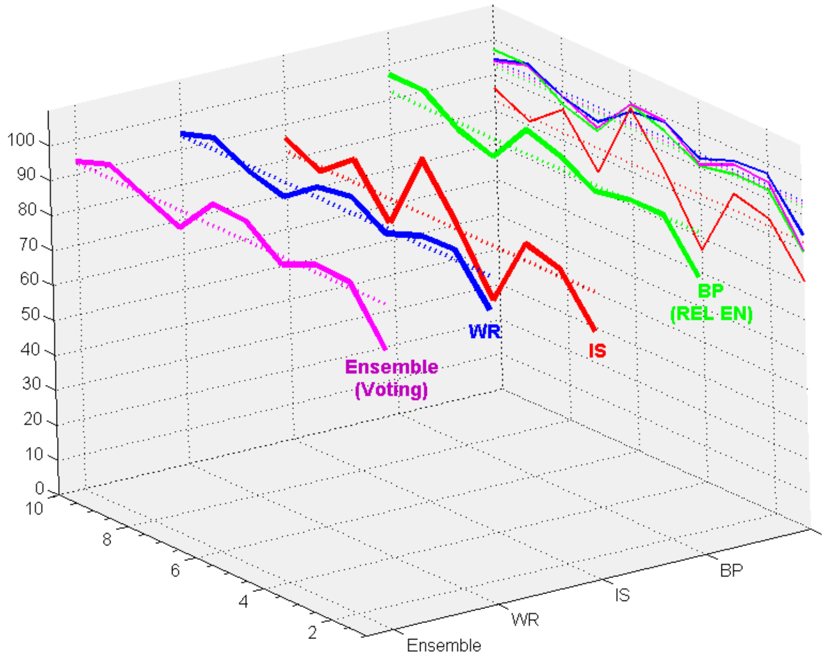

Figure 6: Accuracy curves of the three classifiers and of the committee. From right to left: backpropagation (green), threshold on 'Initial Slope' (red), threshold on 'Washout Ratio' (blue), committee (magenta). The dotted lines represent the mean value of the classifier accuracy.

curacy on the training set. Such threshold was chosen by computing ROC curves, as shown in Fig. 5. The thresholds chosen by the two classifier at each iteration were stable. In the 'Washout Ratios' case the chosen threshold was 1.5 nine times and 1.0 in only one iteration. In the 'Initial Slope' case the threshold was 15 in seven iterations and 16 in only three of them. The little variance of the two sets of thresholds can give precious hints in the future construction of more complex classifiers.

Before presenting and discussing the results, some specifications need to be made:

- in order to get reliable subtracted images, the post-contrast and the pre-contrast images should be registered in order to compensate for deformations caused by patient's movement and by her breathing and heartbeat. However, the curves used in the current experimentation are extracted from original images since clinicians prefer not to perform image registration in order to avoid modifying the original gray levels and loose information.
- The tested classifiers do not use architectural features and any prior knowledge, like, for example, lesion shape and location in previously acquired mammographies. If a pattern extracted from a malignant curve shows a clear and unambiguous benign behaviour, according to the classification scheme shown in Fig. 1, any classifier not using architectural features would make a wrong diagnosis.

RESULTS

As previously said, the classifiers were trained and tested using a 10-fold crossvalidation scheme. The accuracies of the three classifiers in the 10 iterations are shown in Table 1. The simple threshold-based classifier, applied to the washout ratios, is better than the more complex backpropagation. It has not only a greater mean accuracy, but it outperforms the neural network in nine out of ten iterations. Even if the feature it is applied to is one of the most reliable signs of malignancy, the simplicity of the ratio computation should have forced it to make more mistakes. In fact, the washout ratio, as implemented in the current experimentation, takes into account only the difference between the maximum value of a relative enhancement curve and the last value of the same curve. Curves like the false negative ones in the eigthth iteration, plotted in Fig. 7, are incorrectly labeled as benign by the WR classifier since they have a washout ratio equal to zero: the difference between the maximum value and the last one is small, but the washout phase is clear from the intermediate values. A potential extension of the current experimentation would be the combination of predictions based on washout ratios computed at different time instants. The lower accuracy of the backpropagation neural network could be caused by the uncertainty, that the network cannot decrease, related to the curves presenting, after the normalization, a quasi-plateau phase. The initial slope value does not provide, except for particular datasets, a reliable classification. That is an expected result since it is strongly related to the 'Values after 80 seconds' feature, which has been found almost useless.

Once the three classifiers have been trained, they can be joined in a single committee classifier. The simplest combination scheme is the 'voting' one: a test pattern is classified by the committee according to the majority of the single classifiers the committee is composed of. However, the committee has accuracy, specificity, and sensitivity lower than the WR-classifier. This fact suggests that the backpropagation and the IS classifier agree quite often in giving a wrong classification. In some cases, like for example Test set 6, the combination of the prediction made by the backpropagation and the IS classifier outperforms the WR classifier. Even if the results obtained by the committee in this experimentation are not very good, we consider the combination of simple classifiers the correct way to improve the diagnosis. In fact, one of the requirements of clinicians is to be able to understand what an algorithm is doing and the reasons of its final classification. The use of very simple, human-readable features allow them to understand every single step and to accept or reject the result with confidence.

In Fig. 7, the patterns, that have been misclassified by the committee, are plotted for each iteration. In some cases the misclassifed patterns really indicate a different type of lesion. For example, in iteration eight, the false positive presents a washout phase.

Lastly, we need to specify that, in real settings, the results of classifiers based on dynamic features might be better. In fact, we performed a curve-based test: in the practice, it is fundamental that the classifier highlight a lesion. The experimentation should be conducted on a per lesion basis.

CONCLUSIONS AND FUTURE WORK

In this paper we presented a simple experimentation of three classifiers based on basic features, such as, for example, the washout ratio. The classifiers were experimented on 1800 curves extracted from 60 examinations using a 10-fold crossvalidation. Such experimentation strategy clearly show that the classifiers are quite good in classifying lesions with typical patterns, but fail, like in Test Set 1, when lesions present atypical kinetic patterns. Such observation confirms that, by themselves, the kinetic patterns are not able to distinguish correctly between benign and malignant lesions. Anyway, the results are quite good since almost all of the classifiers reached a mean accuracy greater than 90% using only kinetic features.

There are several ways we plan our future work:

- the short-term goal is to enrich the set of kinetic features used by the classifiers and try to improve the classification of contrast-enhancement patterns. Furthermore, other classifiers need to be tested: in past experimentations we used other neural architectures and linear discriminant analysis after a principal component analysis. When choosing a new classifier, one must take into account that, as previously said, clinicians rightly pretend to understand why a computer program made a decision.

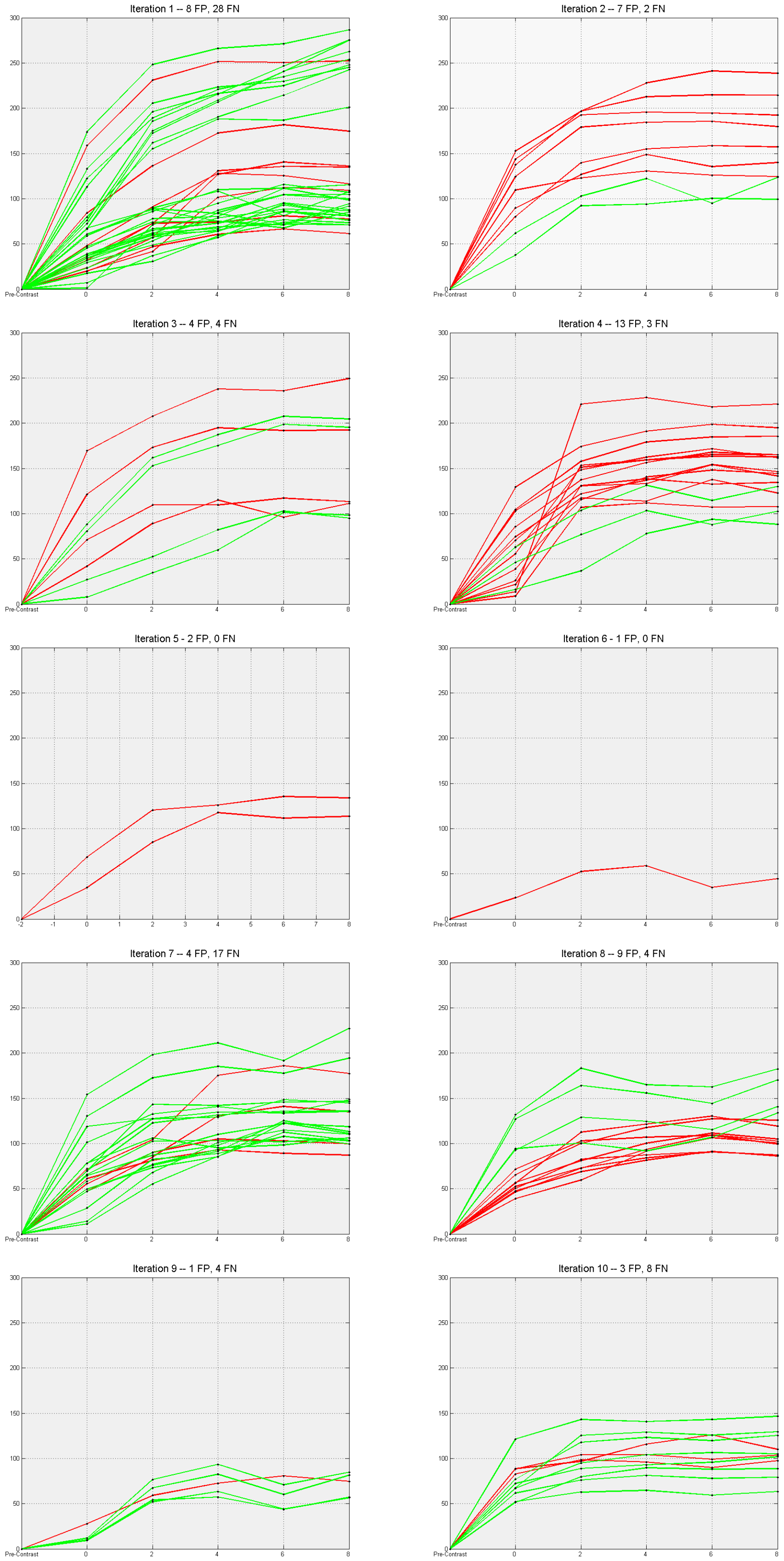

Figure 7: Relative-enhancement curves of false positive (red curves) and false negative (green curves) patterns for each iteration of the committee classifier.

- The long-term goal is to introduce morphological analysis. The resolution of the examinations (256×256 voxels for both breasts) in the current dataset does not allow us to perform such type of analysis. However, we are now collecting new data which offer a resolution high enough for applying morphological criteria.

REFERENCES

[1] Heywang-Köbrunner S.H., Viehweg P. Heinig A, and K´uchler Ch. 1997, Eur J Cancer, 24, 94–108.

[2] Takeda Y., Yoshikawa, K., 2005, Biomed Pharmacother, 59, 115–121.

[3] Kuhl C.K., and Schild H.H., 2000, J Magn Reson Imaging, 12, 965–974.

[4] Fischer H., Kopka L., Grabbe E., 1999, Radiology, 213, 881–889.

[5] Nunes L. W., Schnall M. D., and Orel S. G., 2001, Radiology, 219, 484–494.

[6] Szabó, B. K., Aspelin, P., Kristoffersen Wiberg, M., and Boné, B., 2003, Acta Radiol, 44, 379–386.

[7] Kaiser W.A. and Zeitler E., 1989, Radiology, 170, 681–686

[8] Ikeda D.M., Yamashita Y., Morishita S. et al., 1999, Acta Radiol, 40, 585–592

[9] Lucht R.E.A., Knopp M. V., and Brix G., 2001, Magn Reson Imaging, 19, 51–57

[10] Szabó B. K., Aspelin P., and Kristoffersen Wiberg M., 2004, Acad Radiol, 11, 1344–1354