Classification of Malignant and Benign Masses Based on Hybrid ART2LDA Approach

Lubomir Hadjiiski,* *Member, IEEE*, Berkman Sahiner, *Member, IEEE*, Heang-Ping Chan, Nicholas Petrick, *Member, IEEE*, and Mark Helvie

***Abstract*—A new type of classifier combining an unsupervised and a supervised model was designed and applied to classification of malignant and benign masses on mammograms. The unsupervised model was based on an adaptive resonance theory (ART2) network which clustered the masses into a number of separate classes. The classes were divided into two types: one containing only malignant masses and the other containing a mix of malignant and benign masses. The masses from the malignant classes were classified by ART2. The masses from the mixed classes were input to a supervised linear discriminant classifier (LDA). In this way, some malignant masses were separated and classified by ART2 and the less distinguishable benign and malignant masses were classified by LDA. For the evaluation of classifier performance, 348 regions of interest (ROI's) containing biopsy proven masses (169 benign and 179 malignant) were used. Ten different partitions of training and test groups were randomly generated using an average of 73% of ROI's for training and 27% for testing. Classifier design, including feature selection and weight optimization, was performed with the training group. The test group was kept independent of the training group. The performance of the hybrid classifier was compared to that of an LDA classifier alone and a backpropagation neural network (BPN). Receiver operating characteristics (ROC) analysis was used to evaluate the accuracy of the classifiers. The average area under the ROC curve (A_z) for the hybrid classifier was 0.81 as compared to 0.78 for the LDA and 0.80 for the BPN. The partial areas above a true positive fraction of 0.9 were 0.34, 0.27 and 0.31 for the hybrid, the LDA and the BPN classifier, respectively. These results indicate that the hybrid classifier is a promising approach for improving the accuracy of classification in CAD applications.**

***Index Terms*— Computer-aided diagnosis, hybrid classifier, mammography, neural networks.**

I. Introduction

MAMMOGRAPHY is the most effective method for detection of early breast cancer [1]. However, the specificity for classification of malignant and benign lesions from mammographic images is relatively low. Clinical studies have shown that the positive predictive value (i.e., ratio of the number of breast cancers found to the total number of biopsies) is only 15% to 30% [2]–[4]. It is important to increase the positive predictive value without reducing the sensitivity of breast cancer detection. Computer-aided diagnosis (CAD) has the potential to increase the diagnostic accuracy by reducing the false-negative rate while increasing the positive predictive values of mammographic abnormalities.

Classifier design is an important step in the development of a CAD system. A classifier has to be able to merge the available input feature information and make a correct evaluation. Commonly used classifiers for CAD include linear discriminants (LDA) [5], [6] and backpropagation neural networks (BPN) [7]–[9] which have been shown to perform well in lesion classification problems [10]–[22]. These classifiers are generally designed by supervised training. However, these types of classifiers have limitations dealing with the nonlinearities in the data (in case of LDA) and in generalizability when a limited number of training samples are available (especially BPN). Another classification approach is based on unsupervised classifiers, which cluster the data into different classes based on the similarities in the properties of the input feature vectors. Therefore, unsupervised classifiers can be used to analyze the similarities within the data. However, it is difficult to use them as a discriminatory classifier [29], [30]. They also have limited generalizability when the training sample set is small.

We propose here a hybrid unsupervised/supervised structure to improve classification performance. The design of this structure was inspired by neural information processing principles such as self organization, decentralization and generalization. It combines the adaptive resonance theory network (ART2) [26], [27] and the LDA classifier as a cascade system (ART2LDA). The self-organizing unsupervised ART2 network automatically decomposes the input samples into classes with different properties. The ART2 network has been found to perform better compared to conventional clustering techniques in terms of learning speed and discriminatory resolution for the detection of rare events in many classification tasks [28]–[30]. The supervised LDA then classifies the samples belonging to a subset of classes that have greater similarities. By improving the homogeneity of the samples, the classifier designed for the subset of classes may be more robust.

The ART2LDA design implements both structural and data decomposition. Decomposition is a powerful approach that can reduce the complexity of a problem. Both structural decom-

Manuscript received January 27, 1999; revised October 26, 1999. This work was supported by in part by the USPHS under Grant No. CA 48129 and in part by the U.S. Army Medical Research and Materiel Command (USAMRMC) under Grant DAMD 17-96-1-6254. The work of L. Hadjiiski was supported in part by the USAMRMC under Career Development Award DAMD 17-98-1-8211. The work of B. Sahiner was supported in part by the USAMRMC under Career Development Award DAMD 17-96-1-6012. The work of Nicholas Petrick was supported in part by a grant from The Whitaker Foundation. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was N. Karssemeijer. *Asterisk indicates corresponding author*.

*L. Hadjiiski, B. Sahiner, H.-P. Chan, N. Petrick, and M. Helvie are with the Department of Radiology, The University of Michigan, Ann Arbor, MI 48109-0904 USA.

Publisher Item Identifier S 0278-0062(99)10410-5.

0278–0062/99$10.00 © 1999 IEEE

position and data decomposition can improve classification accuracy [23] as well as model accuracy [24]. However, decomposition can also reduce the prediction accuracy due to overfitting the training data. We will demonstrate in this paper that the proposed hybrid structure can reduce the overfitting problem and improve the prediction capabilities of the system. The performance of the hybrid ART2LDA classifier will be compared with those of an LDA alone or a BPN classifier.

The rest of the paper is organized as follows. In Section II the ART2 unsupervised network is described. A hybrid ART2LDA classifier is introduced in Section III. Section IV describes the data set used in this study. The results are presented in Section V. Section VI contains discussion of these results. Finally, Section VII concludes this investigation.

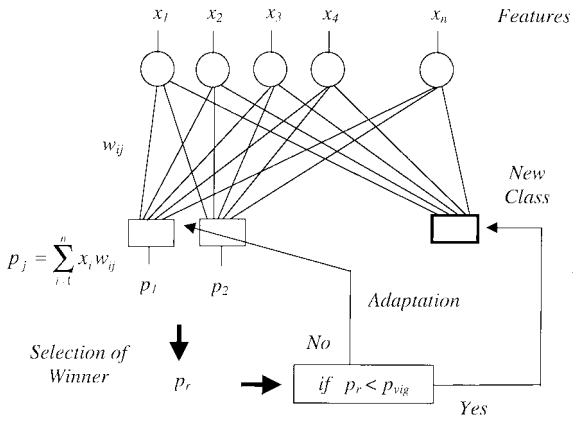

Fig. 1. Structure of the ART2 network.

II. ART2 Unsupervised Neural Network

The ART2 is a self-organizing system that can simulate human pattern recognition. ART2 was first described by Grossberg [25] and a series of further improvements were carried out by Carpenter, Grossberg, and coworkers [26]–[28]. The ART2 network clusters the data into different classes based on the properties of the input feature vectors. The members within a class have similar properties. The process of ART2 network learning is a balance between the plasticity and stability dilemma. Plasticity is the ability of the system to discover and remember important new feature patterns. Stability is the ability of the system to remain unchanged when already known feature patterns with noise are input to the system. The balance between plasticity and stability for the ART2 training algorithm allows fast learning [28], i.e., rare events can be memorized with a small number of training iterations without forgetting previous events. The more conventional training algorithms, such as back propagation [7]–[9], perform slow learning, i.e., they tend to average over occurrences of similar events and require many training iterations.

The structure of the ART2 system is shown in Fig. 1. It consists of two parts: the ART2 network and the learning stage. Suppose that there are n input features x_i $(i = 1, \cdots, n)$ and k classes in the ART2 network. When a new vector is presented to the input of the ART2 network, an activation value p_j for class j is calculated as

$$p_j = \sum_{i=1}^{n} x_i w_{ij}, \qquad j = 1, \cdots, k \tag{1}$$

where w_{ij} is the connection weight between input i and class j. The activation value is a measure of the membership of the particular input feature vector to class j. The higher the value p_j is, the better the input vector matches class j. The maximum value p_r is selected from all p_j $(j = 1, \cdots, k)$ to find the best class match. Furthermore, in order to balance the contribution to the activation value from all feature components, the input feature values applied to the ART2 system are scaled between zero and one [30]. This normalization will allow detection of similar feature patterns even when the magnitudes of the input feature components are very different.

The learning stage of the ART2 system can influence the weights of the selected class or the complete ART2 network structure by adding a new class. An additional parameter, the vigilance, is used to determine the type of learning [26]. The vigilance parameter p_{vig} is a threshold value that is compared to the maximum activation value p_r. If p_r is larger than p_{vig} then the input vector is considered to belong to class r. The adaptation of the weights connected with class r is performed as follows:

$$w_{ir}^{new} = w_{ir}^{\text{old}} + \eta(x_i - w_{ir}^{\text{old}}), \qquad \text{for } i = 1, \cdots, n \tag{2}$$

where η is a learning rate. The adaptation of the class r weights (2), aims at maximization of the p_r value for the particular input vector. In an iterative manner the weights are adjusted so that the activation values produced for similar input vectors will be maximum only for the class to which they belong and these maximum activation values will be higher than p_{vig}.

If the maximum activation value p_r is smaller than p_{vig}, it is an indication that a novelty has appeared and a new class will be added to the ART2 structure. The new weights connecting the input with the new class $(k+1)$ are initialized with the scaled input feature values of this novelty. In such a way, the activation value p_{k+1} will be maximum $(p_r = p_{k+1})$ higher than p_{vig} when computed for this novelty in further training iterations. The value of the vigilance parameter p_{vig} determines the resolution of ART2. It can be chosen in the range between zero and one. In the case that p_{vig} is relatively small, only very different input feature vectors will be distinguished and separated in different classes. If p_{vig} is relatively large, the input feature vectors that are more similar will be separated into different classes. The value of p_{vig} is selected differently depending on the particular application.

III. ART2LDA Classifier

Despite the good performance of ART2 for efficient clustering and detection of novelties, the fast learning approach can cause problems associated with the generalization capability of the system and the correct classification of unknown cases. Supervised classifiers such as linear discriminants or backpropagation neural network classifiers can have better generalization capability than ART2, because they are trained by averaging over similar event occurrences. However, the learning process in these traditional learning algorithms tends

to erase the memory of previous expert knowledge when a new type of expertise is being learned. Therefore, these classifiers do not have as good an ability to correctly classify rare events as ART2 [28], [29].

In order to improve the accuracy and generalization of a classifier, we propose to design a hybrid classifier that combines the unsupervised ART2 network and a supervised LDA classifier. This hybrid classifier (ART2LDA) utilizes the good resolution capability of ART2 and the good generalization capability of LDA. The ART2 first analyzes the similarity of the sample population and identifies a subpopulation that may be separated from the main population. This will improve the performance of the second-stage LDA if the subpopulation causes the sample population to deviate from multivariate normal distributions for which LDA is an optimal classifier. Therefore, the ART2 serves as a screening tool to improve the homogeneity of the sample distributions by classifying outlying samples into separate classes.

The ART2LDA hybrid classifier can be described as

$$y_{AL} = g(f_2(x))f_1(x) + 1 - g(f_2(x)) \tag{3}$$

where x is the input vector, $f_1(\cdot)$ is the LDA classifier, $f_2(\cdot)$ is the ART2 classifier, and $g(\cdot)$ is a binary membership function, which labels the classes identified by ART2 to be one of the two types: malignant class or mixed class. A particular class is defined as malignant if it contains only malignant members. It is defined as mixed if it contains both malignant and benign members. The membership function is defined as follows:

$$g(c) = \begin{cases} 0, & \text{if } c \text{ is a malignant class} \\ 1, & \text{if } c \text{ is a mixed class.} \end{cases} \tag{4}$$

The type of a given class is determined based on ART2 classification of the training data set.

The structure of the ART2LDA classifier is shown in Fig. 2. The ART2 classifies the input sample x into either a malignant or a mixed class. Depending on the class type the function $g(\cdot)$ determines whether the LDA classifier will be used. If x is classified into a mixed class, the final classification will be obtained based on the LDA classifier. However, if x is classified by ART2 into a malignant class, then the mass will be considered malignant, without using the LDA classifier. Therefore, in the ART2LDA structure, the ART2 is used both as a classifier and a supervisor. This can be seen in (3). The first term in (3), $g(f_2(x))f_1(x)$, is the LDA classifier multiplied by the ART2 control part $g(f_2(x))$. The second term in (3), $(1 - g(f_2(x)))$, gives the classification result of the ART2 stage. If $f_2(x)$ is a malignant class, then $g(f_2(x)) = 0$, the LDA stage is eliminated, and the classifier output y_{AL} is equal to 1. On the other hand, if $f_2(x)$ is a mixed class, then $g(f_2(x)) = 1$, the ART2 term is eliminated, and the final classification is determined by the LDA classifier $(y_{AL} = f_1(x))$.

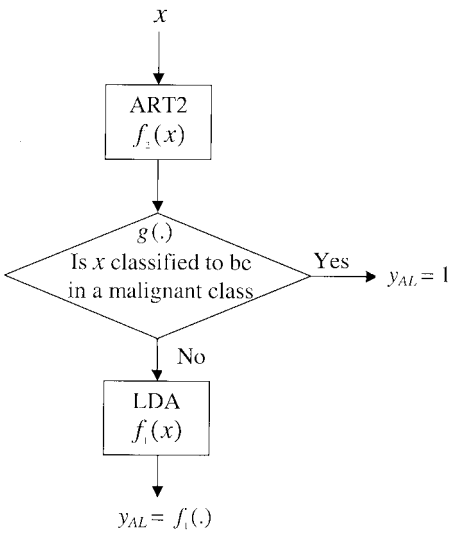

Fig. 2. Structure of the ART2LDA classifier.

IV. Methods

A. Data Set

The mammograms used in this study were randomly selected from the files of patients who had undergone biopsies at the University of Michigan. The criterion for inclusion of a mammogram in the data set was that the mammogram contained a biopsy-proven mass. The data set contained 348 mammograms with a mixture of benign $(n = 169)$ and malignant $(n = 179)$ masses. On each mammogram, a region of interest (ROI) containing the mass was identified by a radiologist experienced in breast imaging. The visibility of the masses was rated by the radiologist on a scale of 1 to 10, where the rating of 1 corresponds to the most visible category. The distributions of the visibility rating for both the malignant and benign masses are shown in Fig. 3. The visibility ranged from subtle to obvious for both types of masses. It can be observed that the benign masses tend to be more obvious than the malignant ones. Additionally the likelihood of malignancy for each mass was estimated based on its mammographic appearance. The radiologist rated the likelihood of malignancy on a scale of 1 to 10, where 1 indicated a mass with the most benign appearance. The distribution of the malignancy rating of the masses is shown in Fig. 4.

The data set can be considered as representative of the patient population that is sent for biopsy under current clinical criteria. Some characteristics of many malignant and benign masses can be visually distinguished by radiologists. However, there is also a nonnegligible fraction of malignant masses that are very similar to benign masses (the low malignancy rating region in Fig. 4). The estimated likelihood of malignancy of malignant and benign masses that are sent for biopsy basically overlaps over the entire range. This is consistent with the fact that in order not to miss malignant masses radiologists must recommend biopsy for even very low suspicion lesions.

Three hundred and five of the mammograms were digitized with a LUMISYS DIS-1000 laser scanner at a pixel resolution of 100 μm $\times$ 100 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly and inversely proportional to the optical density (OD) within the range of 0.1 to 2.8 OD units, with a slope of -0.001 OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually. The OD range of the digitizer was 0

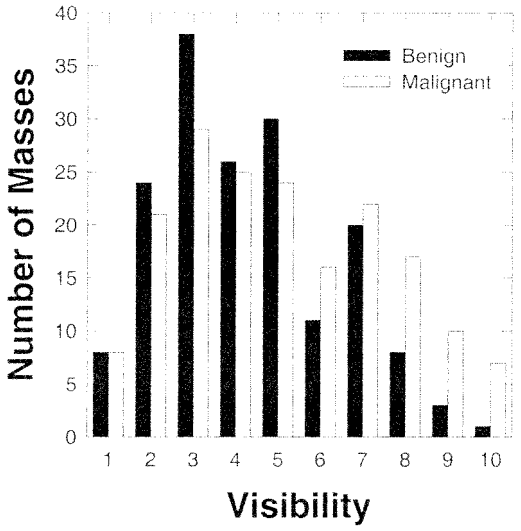

Fig. 3. The distribution of the visibility ranking of the masses in the dataset. The ranking was performed by an experienced breast radiologist (1: very obvious, 10: very subtle).

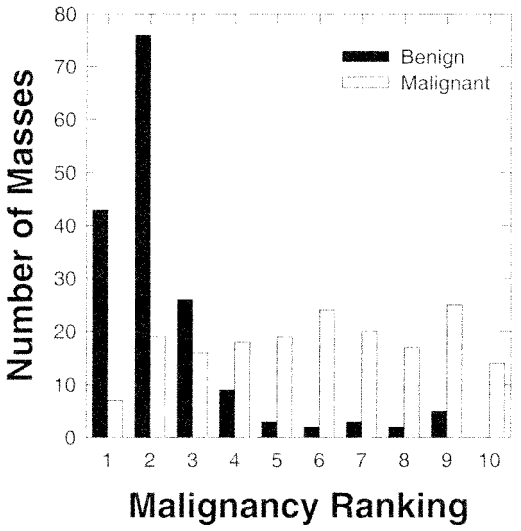

Fig. 4. The distribution of the malignancy ranking of the masses in the dataset. The ranking was performed by an experienced breast radiologist (1: very likely benign, 10: very likely malignant).

to 3.5. The remaining 43 mammograms were digitized with a LUMISCAN 85 laser scanner at a pixel resolution of 50 μm $\times$ 50 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly and inversely proportional to the OD within the range of 0 to 4 OD units, with a slope of -0.001 OD/pixel value. In order to process the mammograms digitized with these two different digitizers, the images digitized with LUMISCAN 85 digitizer were averaged with a 2×2 box filter and subsampled by a factor of two, resulting in 100 μm images.

In order to validate the prediction abilities of the classifier, the data set was partitioned randomly into training and test subsets on a 3:1 ratio, under the constraints that both the malignant and the benign samples were split with the 3:1 ratio and that the images from the same patient were grouped into the same (training or test) subset. These constraints caused the subsets to deviate from an exact 3:1 ratio. The data set was repartitioned randomly ten times. On average, 73% of the samples were grouped into the training set and 27% into the test set. The training and test results from the ten partitions were averaged to reduce their variability.

B. Feature Extraction

A rectangular ROI was defined to include the radiologist-identified mass with an additional surrounding breast tissue region of at least 40 pixels wide from any point of the mass border. A fully automated method was then used for segmentation of the mass from the breast tissue background within the ROI. The rubber band straightening transform (RBST) was previously developed [12] to map a band of pixels surrounding the mass onto the Cartesian plane (a rectangular region). In the transformed image, the border of mass appears approximately as a horizontal edge and spiculations appear approximately as vertical lines. The transformation of the radially oriented textures surrounding the mass margin to a more uniform orientation facilitates the extraction of texture features.

The texture features used in this study were calculated from spatial gray-level dependence (SGLD) matrices [10]–[12], [31], and run-length statistics (RLS) matrices [32] computed from the RBST images. The (i, j)th element of the SGLD matrix is the joint probability that gray levels i and j occur in a direction at a distance of θ pixels apart in an image. Based on our previous studies [10], a bit depth of eight was used in the SGLD matrix construction, i.e., the four least significant bits of the 12-bit pixel values were discarded. Thirteen texture measures, including correlation, energy, difference entropy, inverse difference moment, entropy, sum average, sum entropy, inertia, sum variance, difference average, difference variance, and two types of information measure of correlation were used. These measures were extracted from each SGLD matrix at ten different pixel pair distances $(d = 1, 2, 3, 4, 6, 8, 10, 12, 16$ and $20)$ and in four directions (0°, 45°, 90°, and 135°). Therefore, a total of 520 SGLD features were calculated for each image. The definitions of the texture measures are given in the literature [10]–[12], [31]. These features contain information about image characteristics such as homogeneity, contrast, and the complexity of the image.

RLS texture features were extracted from the vertical and horizontal gradient magnitude images, which were obtained by filtering the RBST image with horizontally or vertically oriented Sobel filters and computing the absolute gradient value of the filtered image. A gray level run is a set of consecutive, collinear pixels in a given direction which have the same gray level value. The run length is the number of pixels in a run [32]. The RLS matrix describes the run length statistics for each gray level in the image. The (i, j)th element of the RLS matrix is the number of times that the gray level i in the image possesses a run length of j in a given direction. In our previous study, it was found experimentally that a bit depth of five in the RLS matrix computation could provide good texture characteristics [12].

Five texture measures, namely, short run emphasis, long run emphasis, gray level nonuniformity, run length nonuniformity,

and run percentage were extracted from the vertical and horizontal gradient images in two directions, $\theta = 0°$ and $\theta = 90°$. Therefore, a total of 20 RLS features were calculated for each ROI. The formal definition of the RLS feature measures can be found in [32].

A total of 540 features (520 SGLD and 20 RLS) were therefore extracted from each ROI.

C. Feature Selection

In order to reduce the number of the features and to obtain the best feature set to design a good classifier, feature selection with stepwise linear discriminant analysis [33] was applied. At each step of the stepwise selection procedure one feature is entered or removed from the feature pool by analyzing its effect on the selection criterion. In this study, the Wilks' lambda (the ratio of within-group sum of squares to the total sum of squares [34]) was used as a selection criterion. The optimization procedure used a threshold F_{in} for feature entry and a threshold F_{out} for feature removal. On a feature entry step, the features not yet selected are entered into the selected feature pool one at a time, the significance of the change in the Wilks' lambda caused by this feature is estimated based on F statistics. The feature with the highest significance is entered into the feature pool if its significance is higher than F_{in}. On a feature removal step, the features which have already been selected are analyzed one at a time from the selected feature pool and the significance of the change in the Wilks' lambda is estimated. The feature with the least significance is removed from the selected feature pool if the significance is less than F_{out}. Since the appropriate values of F_{in} and F_{out} are not known *a priori*, we examined a range of F_{in} and F_{out} values and chose the appropriate thresholds in such a way that a minimum number of features were selected to achieve a high accuracy of classification by LDA for the training sets. More details about the stepwise linear discriminant analysis and its application to CAD can be found in [10]–[12].

D. Performance Analysis

To evaluate the classifier performance, the training and test discriminant scores were analyzed using receiver operating characteristic (ROC) methodology [35]. The discriminant scores of the malignant and benign masses were used as decision variables in the LABROC1 program [36], which fit a binormal ROC curve based on maximum likelihood estimation. The classification accuracy was evaluated as the area under the ROC curve, A_z. For the ART2LDA classifier, the discriminant scores of all case samples classified in the two stages are combined. All masses classified into the malignant group by the ART2 stage were assigned a constant positive discriminant score higher than or equal to the most malignant discriminant score obtained from the LDA stage .

The performance of ART2LDA was also assessed by estimation of the partial area index $(A_z^{(0.9)})$ and compared with the corresponding performance index of the LDA and BPN classifiers. The partial area index $(A_z^{(0.9)})$ is defined as the area that lies under the ROC curve but above a sensitivity threshold of 0.9 $(\text{TPF}_0 = 0.9)$ normalized to the total area above TPF_0, (1-TPF_0). The partial $A_z^{(0.9)}$ indicates the performance of the classifier in the high-sensitivity (low false negative) region which is most important for clinical cancer detection task. In addition, the performance of the LDA stage of the ART2LDA classifier was evaluated by the estimation of the area under the ROC curve, denoted as A_z (LDA), for the case samples passed onto the LDA classifier.

TABLE I
NUMBER OF SELECTED FEATURES FOR THE TEN DATA GROUPS WITH THE CORRESPONDING F_{IN} AND F_{OUT} PARAMETERS

Data Group No.	Number of selected features	F_{in}	F_{out}
1	12	1.8	1.6
2	15	2.4	2.2
3	13	2.4	2.2
4	18	2.4	2.2
5	14	2.4	2.2
6	14	2.1	1.8
7	13	2.4	2.2
8	18	1.8	1.6
9	14	2.4	2.2
10	14	2.4	2.2

V. RESULTS

In this section the ART2LDA classification results for malignant and benign masses will be presented and compared with those of the LDA or BPN classifiers. The important point in this study is the fact that the test subset is truly independent of the training subset. Only the training subset is used for feature selection and classifier training, and only the test subset is used for classifier validation. In order to validate the prediction abilities of the classifier, ten different partitions of the training and test sets were used. A different ART2LDA classifier was trained using each training set and the corresponding set of selected features. The classification result was estimated as the average performance for the ten partitions.

For a given partition of training and test sets, feature selection was performed based on the training set alone. The feature selection results for the ten different training groups are shown in Table I. The average number of selected features was 14. An average of two RLS features and twelve SGLD features were selected for each of the training sets which represented 10% of all RLS features and 2.3% of all SGLD features, respectively. Both types of features (RLS and SGLD) are necessary in order to obtain good classification. The most often selected RLS features for the ten training sets were: horizontal short run emphasis (four times), horizontal long run emphasis (six times), vertical run length nonuniformity (three times), horizontal run length nonuniformity (three times). The most often selected SGLD texture measures for the ten training sets were: inverse difference moment (eight times), information measure of correlations one and two (19 times), difference average (nine times), and correlation (ten times). For a given texture measure, features at different angles or distances may be selected, but these features are usually highly correlated so

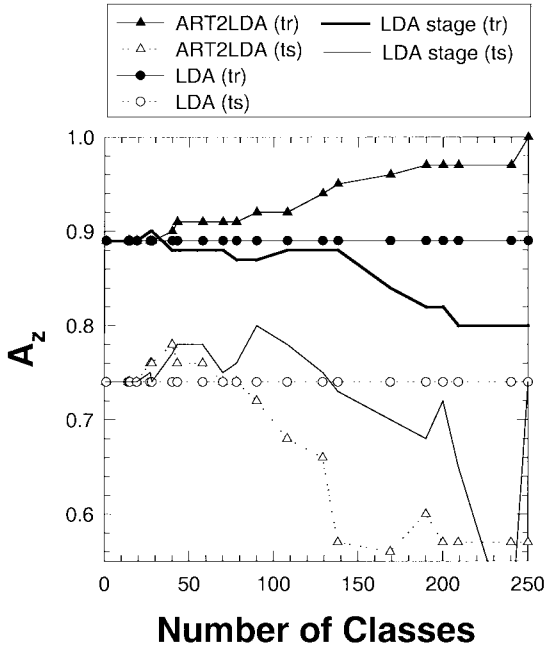

Fig. 5. ART2LDA and LDA classification results for training and test sets from data group three as a function of the generated number of classes. Additionally the results for the LDA stage from the ART2LDA classifier are plotted.

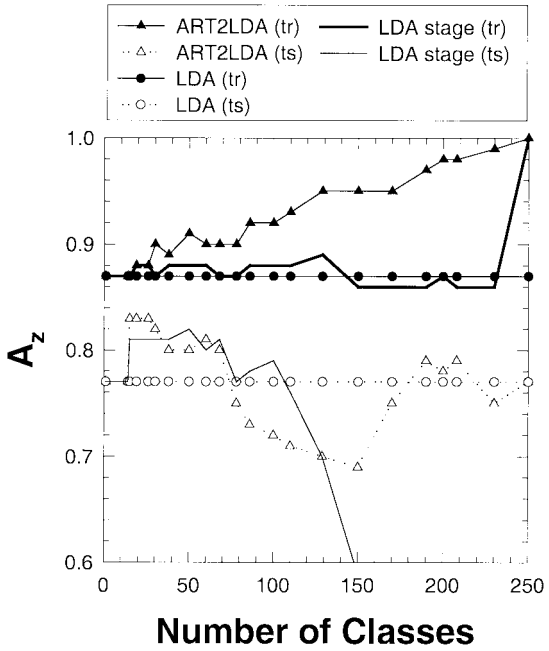

Fig. 6. ART2LDA and LDA classification results for training and test sets from data group one as a function of the generated number of classes. Additionally the results for the LDA stage from the ART2LDA classifier are plotted.

that they can be considered to be similar and counted together as described above.

A. ART2LDA Classification Results

For the ART2LDA classifier, the number of selected features determines the dimensionality of the input vector of the ART2 classifier and the dimensionality of the LDA classifier. By applying different values for the vigilance parameter, ART2 classifiers with different number of classes were obtained. In this study, the vigilance parameter p_{vig} was varied from 0.9 to 0.99, resulting in a range of 10 to 240 classes. The overall performance of the ART2LDA classifier was evaluated for different numbers of ART2 classes because different subset of the samples were separated and classified by ART2 when p_{vig} was varied. In Fig. 5, the classification results for the ART2LDA are compared to the results from LDA alone for the training and test set partition three. The classification accuracy, A_z, was plotted as a function of the number of ART2 classes. For this training and test set partition, when the number of classes was between 20 and 60, the ART2LDA classifier improved the classification accuracy for the test set in comparison to LDA. As the number of classes increased to greater than 60, the A_z value increased for the training data set, but decreased for the test data set and was lower than that of the LDA alone. The two solid lines in Fig. 5 show the A_z values for the LDA stage in the ART2LDA classifier for both the training and test sets. It can be observed that the test A_z for the LDA stage is higher than the A_z for the LDA classifier alone, but not as high as A_z obtained by ART2LDA when the number of classes is small.

In Fig. 6 the classification results of LDA and ART2LDA for the partition one training and test sets are shown. In this case it appeared that in the test set there were two large malignant outliers which degraded the LDA performance. Only 15 classes at the ART2 stage in the ART2LDA was enough to cluster the outliers into a separate malignant class and to improve the performance of the LDA stage and the overall result. The rest of the outliers required more ART2 classes before they were clustered into separate classes and correctly classified as malignant. This is the reason for the similar behavior of the classifiers for partitions three and one in the range of 40 to 70 classes as seen in Figs. 5 and 6. When the number of classes was less than 70, the test A_z for the LDA stage (A_z(LDA)) was higher than the LDA alone, but not as high as the A_z for ART2LDA with less than 30 classes (Fig. 6). The best A_z values for the test data sets of the ten training and test partitions are presented in Table II and Fig. 7. The ART2LDA classifier achieved higher A_z values than the LDA alone in nine of the ten partitions. The average A_z is 0.81 for ART2LDA and 0.78 for LDA alone. The standard deviations of the A_z values for the ten groups range from 0.03 to 0.05 for the ART2LDA classifier and from 0.04 to 0.05 for the LDA classifier.

The performance of ART2LDA was also assessed by estimation of the partial area under the ROC curve $A_z^{(0.9)}$ at a TPF higher than 0.9. The results are presented in Table III and Fig. 7. In the lower part of Fig. 7, the $A_z^{(0.9)}$ values of the test set for the corresponding ten partitions of training and test sets are presented. The average test $A_z^{(0.9)}$ value is 0.34 for the ART2LDA and 0.27 for LDA. For nine of the ten partitions, the $A_z^{(0.9)}$ value was improved at the high-sensitivity operating region (TPF $>$ 0.9) of the ROC curve.

The classifier performance was also evaluated when the ART2LDA classifiers were designed using a fixed number

TABLE II
CLASSIFIERS PERFORMANCE FOR THE TEN TEST SETS. THE A_z VALUES REPRESENT THE TOTAL AREA UNDER ROC CURVE

Data Group No.	LDA	ART2LDA	BPN	ART2LDA(1)
1	0.77	0.83	0.85	0.80
2	0.78	0.80	0.82	0.77
3	0.74	0.78	0.77	0.78
4	0.77	0.77	0.75	0.77
5	0.77	0.78	0.76	0.77
6	0.80	0.83	0.82	0.81
7	0.80	0.81	0.82	0.77
8	0.77	0.80	0.74	0.75
9	0.77	0.80	0.81	0.80
10	0.86	0.89	0.84	0.89
Mean	0.78	0.81	0.80	0.79

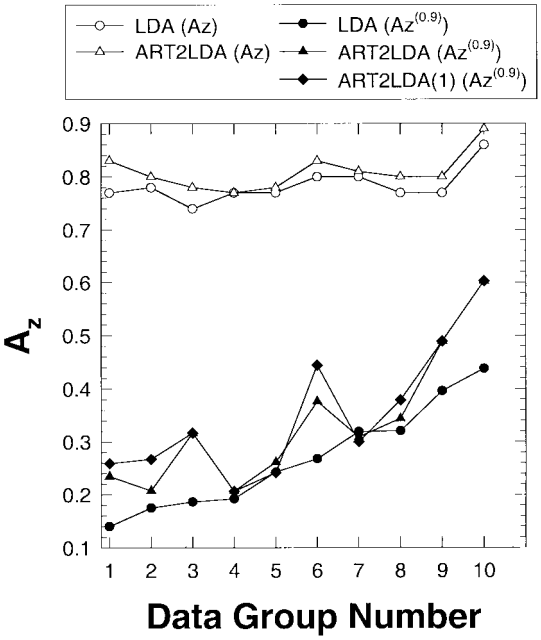

Fig. 7. Average A_z classification results for the 10 test sets. The top graphs represent the ART2LDA and LDA A_z values for the total area under the ROC curve. The bottom graphs represent the ART2LDA, ART2LDA(1) and LDA A_z values for the partial area of the ROC curve above the true positive fraction of 0.9.

TABLE III
CLASSIFIERS RESULTS FOR THE TEN TEST SETS. THE A_z VALUES REPRESENT THE PARTIAL AREA OF THE ROC CURVE ABOVE THE TRUE POSITIVE FRACTION OF 0.9 ($A_z^{(0.9)}$)

Data Group No.	LDA	ART2LDA	BPN	ART2LDA(1)
1	0.14	0.23	0.31	0.26
2	0.17	0.21	0.28	0.27
3	0.19	0.32	0.27	0.32
4	0.19	0.21	0.19	0.21
5	0.24	0.26	0.32	0.24
6	0.27	0.38	0.27	0.44
7	0.32	0.31	0.38	0.30
8	0.32	0.34	0.25	0.38
9	0.40	0.49	0.40	0.49
10	0.44	0.60	0.38	0.60
Mean	0.27	0.34	0.31	0.35

of ART2 classes. The A_z, and $A_z^{(0.9)}$ results, averaged over the ten test partitions, are presented in Table IV. The average A_z with the ART2LDA classifier, compared to that of LDA alone, was again improved between 15 and 40 classes. The maximum average A_z of 0.80 was achieved between 20 and 40 classes. The average $A_z^{(0.9)}$ results are improved for all

TABLE IV
AVERAGE A_z AND AVERAGE $A_z^{(0.9)}$ CLASSIFICATION RESULTS FOR THE TEN TEST SETS. CLASSIFIERS WERE DESIGNED USING A FIXED NUMBER OF ART2 CLASSES

	LDA	ART2LDA					
No. of classes		15	20	30	40	50	60
A_z	0.78	0.80	0.80	0.80	0.80	0.78	0.77
$A_z^{(0.9)}$	0.27	0.30	0.31	0.33	0.33	0.31	0.31

ART2LDA classifiers presented in Table IV. The maximum average $A_z^{(0.9)}$ value is 0.33 and it remains constant between 30 and 40 classes.

An alternative way to evaluate the performance of a classifier is its classification accuracy when a decision threshold for malignancy is selected based on the training set. For instance, a decision threshold may be selected such that all positive samples from the training set are classified correctly i.e., at a sensitivity of 100%. The ART2LDA with this decision threshold is referred to as ART2LDA(1). For a given training and test partitioning, ART2LDA classifiers with different number of classes in the ART2 stage were obtained (Figs. 5 and 6). For each of these models the decision threshold for a sensitivity of 100% was selected from the training set and the corresponding ART2LDA(1) classifier was obtained. Then the ART2LDA(1) classifier (with a specific number of classes in the ART2 stage) that correctly classified the maximum number of malignant masses in the test set is selected. By using all samples of the test set, the A_z value is calculated for the corresponding ART2LDA model. The A_z values for the ART2LDA(1) classifiers for the test sets of the ten data partitionings are shown in Tables II and III. For five of the partitions the overall A_z value for ART2LDA(1) is higher than that of LDA alone (Table II). The average A_z value was 0.79. The partial areas above the TP fraction of 0.9, $A_z^{(0.9)}$, for the ten test data sets obtained by the ART2LDA(1) classifier are also shown in Fig. 7. The ART2LDA(1) achieved the highest average $A_z^{(0.9)}$ value of 0.35 compared to ART2LDA and LDA (Table III).

B. BPN Classification Results

A multilayer perceptron back-propagation neural network with a single hidden layer and a single output node was used for comparison with the ART2LDA classifier. The number of selected features determined the number of input nodes to the BPN. The same ten training/test set partitions (as in the case of ART2LDA) were used for the training and validation of the BPN classifiers. BPN's with their number of hidden nodes ranging from two to ten were evaluated to obtain the best architecture. Back-propagation training was used. Each of the BPN's was trained for up to 18 000 training epochs. At every 1000 epochs the neural network weights were saved and the classification result for the corresponding test set was evaluated. This design procedure was repeated for each of the ten training/test groups. For each group, the best test result among all the BPN architectures (different number of hidden nodes) and all the training epochs examined was selected. The average test A_z over the ten groups for the BPN was 0.80, compared to 0.81 for ART2LDA (Table II). The standard deviations of the A_z values for the ten groups range from 0.04 to 0.05 for the BPN. The average partial $A_z^{(0.9)}$ for the BPN

was 0.31, compared to 0.34 for ART2LDA (Table III). The A_z and $A_z^{(0.9)}$ of the ART2LDA classifier were higher than those of the BPN in six of the ten training/test groups.

VI. Discussion

In the present study, a new classifier (ART2LDA) was designed and applied to the classification of malignant and benign masses. The results indicated that the ART2LDA classifier had better generalizability than an LDA classifier alone. The ART2 classifier grouped the case samples that were different from the main population into separate classes. The minimum number of classes needed to start the clustering of outliers into separate classes depended on how different the outliers were from the rest of the sample population. For the ten different partitions of training and test sets used in this study, the minimum number varied between 13 and 15 classes. When the number of ART2 classes was less than this minimum number of classes, the ART2 classifier generated only mixed malignant-benign classes and all samples were transferred to the LDA stage. In that case, the ART2LDA was equivalent to the LDA classifier alone. When a higher number of classes were generated, an increased number of cases that might be considered outliers of the general data population was removed (clustered in separate classes). For the ten training sets used in this study, the malignant outliers were gradually removed when the number of classes increased. The training accuracy increased when the number of classes increased and A_z could reach the value of 1.0. However, a large number of ART2 classes led to overfitting the training sample set and poor generalization in the test set. The classification accuracy of ART2 for the test set tended to decrease when the number of classes was greater than about 70. The large number of classes also led to a reduction in the generalizability of the second-stage LDA; the training of LDA with a small number of samples would again result in overfitting the training set, and poor generalizability in the test set. This effect was observed when more than 60 or 70 classes were generated by ART2 (see Figs. 5 and 6).

The classification accuracy of ART2LDA increased initially with an increased number of classes and then decreased after reaching a maximum. The correct classification of the outliers by the ART2 in combination with an improvement in the classification by the LDA resulted in the increased accuracy. When the number of ART2 classes was further increased, the effects of overfitting by the ART2 and the LDA became dominant and the prediction ability of the ART2LDA decreased. In some cases the second-stage LDA prediction was much worse than the ART2. In other cases the ART2 could not generalize well. The generation of a high number of classes is therefore impractical and unnecessary both from a computational and a methodological point of view.

For the optimal number of classes (usually less than 50 for the data sets used) the A_z value for the second-stage LDA in the ART2LDA was better than an LDA classifier alone, but it was not as good as the overall A_z from the ART2LDA. It is evident that the ART2 was a useful classifier for improvement of the second-stage classification.

When the partial area of the ROC curve above the true positive fraction (TPF) of 0.9 ($A_z^{(0.9)}$) was considered as a measure of classification accuracy, the advantage of ART2LDA over LDA alone became even more evident. By removing and correctly classifying the outliers, the accuracy of the classification was increased at the high sensitivity end of the curve.

The classifier performance was evaluated when the ART2LDA classifiers were designed using a fixed number of ART2 classes. The results showed improved performance of the ART2LDA in a range between 20 and 40 ART2 classes. Both the average A_z and the average $A_z^{(0.9)}$ reached a maximum within this region, and the maximum average A_z and the average $A_z^{(0.9)}$ values remained unchanged between 30 and 40 classes. These results indicated that the performance of a hybrid ART2LDA classifer was robust and stable and could be potentially useful in real clinical applications.

We have performed statistical tests with the CLABROC program to estimate the significance in the differences between the A_z values from the ART2LDA, the LDA alone, and the BPN, as well as in the differences in the partial $A_z^{(0.9)}$ from the three classifiers. The statistical tests were performed for each individual data set partition because the correlation among the data sets from the different partitions precludes the use of student's paired t test with the ten partitions. We found that the differences in both cases did not reach statistical significance because of the small number of test samples and thus the large standard deviation in the A_z values. However, the consistent improvements in A_z and $A_z^{(0.9)}$ by the ART2LDA (9 out of 10 data set partitions in both cases for LDA and six out of ten data set partitions in both cases for BPN) suggest that the improvement was not by chance alone, and that the accuracy of a classification task could be improved by the use of an ART2 network. In addition, one advantage of the ART2LDA is that the training process is more efficient than that of the BPN, especially when there is a subset of outlying samples. In such a case, the BPN will require a large number of training epochs to minimize the error function.

ART2LDA can be trained to classify the sample cases into more than two classes, such as a class of normal tissue regions in addition to malignant and benign masses. There will be an increase in the complexity of training and a larger training sample size will be desired, but these requirements will be comparable for the different classifiers. In a clinical situation, if the classification task is performed on all computer-detected lesions, the classifier has to distinguish the falsely detected normal tissue from malignant or benign lesions. However, it may be noted that a classifier that can distinguish only malignant and benign masses is applicable to the scenario that the radiologist identifies a suspicious lesion on the mammogram and would like to have a second opinion about its likelihood of malignancy before making a diagnostic decision. Therefore, the development of a classifier that can differentiate malignant and benign masses is the research of interest for many investigators.

Similarly, ART2 can be trained to discover and remove a pure benign mass class. The approach will be similar to the task of classifying and removing the pure malignant classes,

as described in this study. However, our approach of removing the malignant classes will reduce the chance of misclassification of malignant masses. In breast cancer detection, the cost of false-negative (missed cancer) is very high. Therefore, our goal in classifier design is to be conservative. By removing the malignant classes in the first stage, any misclassification to these classes will be regarded as malignant. The remaining classes will be classified again with the second-stage classifier so malignant masses will be less likely to be missed.

The problem of classification of malignant and benign masses has been studied by many investigators. Rangayyan *et al.* [15] used Mahalanobis distance classifer (a modification of an LDA classifier) and the leave-one-out method to evaluate the classification of 54 masses. Fogel *et al.* [16] compared LDA and BPN classifiers using the leave-one-out method and 139 masses (malignant and benign classification). Highnam *et al.* [17] used a morphological feature called a halo to classify 40 masses as malignant and benign. Huo *et al.* [22] employed BPN and a rule-based classifier to classify 95 masses using the leave-one-out evaluation method. Sahiner *et al.* [12] used an LDA classifier and the leave-one-out method to classify 168 masses. An important difference between the classifier designed in this study and the previous studies in the CAD field is the method of feature selection. In the above mentioned studies [12], [15]–[17], [22] and several other published studies [18]–[21] the features were selected from the entire data set first, and then the data set was partitioned into training and test sets. This meant that at the feature selection stage of the classifier design, the entire data set was used as a training set. Depending on the distribution of the features and the total number of samples used, the test results in these studies might be optimistically biased [37]. In our current study, the entire data set was initially partitioned into training and test sets and then feature selection was performed only on the training set. This method will result in a pessimistic estimate of the classifier performance when the training set is small [37]. However, it will provide a more conservative but realistic estimation of the classifier performance in the general patient population. We can expect that the performance would be improved if the classifier in this study were designed using a large data set. Since our main purpose in this study was to compare the ART2LDA classifier with the commonly used LDA and BPN, we did not attempt to quantify how pessimistic our results were in this study.

The most important contribution of this paper is to introduce a new approach that utilizes a two-stage unsupervised–supervised hybrid classifier. We believe that the hybrid approach will improve classification when the sample distribution contains subpopulations that may be difficult for a single classifier to classify. It will be useful for similar classification tasks although different classifiers may be used in each stage of the hybrid structure.

VII. Conclusion

A new classifier combining an unsupervised ART2 and a supervised LDA has been designed and applied to the classification of malignant and benign masses. A data set consisting of 348 films (179 malignant and 169 benign) was randomly partitioned into training and test subsets. Ten different random partitions were generated. For each training set, texture features were extracted and feature selection was performed. An average of features were selected for each group. A hybrid ART2LDA classifier, an LDA, and a BPN were trained by using each of the ten training sets. The A_z value under the ROC curve for the test sets, averaged over the ten partitions, was higher for ART2LDA $(A_z = 0.81)$ compared to those of the LDA alone $(A_z = 0.78)$ and of the BPN $(A_z = 0.80)$. A greater improvement was obtained when the partial ROC area above a true-positive fraction of 0.9 was considered. The average partial A_z for ART2LDA was 0.34, as compared to 0.27 for LDA and 0.31 for BPN. Additionally, for the ART2LDA classifiers that correctly classified the maximum number of malignant masses in the test sets with decision threshold defined with the training set, the average partial A_z was 0.35. These results indicate that the hybrid classifier is a promising approach for improving the accuracy of classifiers for CAD applications.

Acknowledgment

The authors would like to thank Prof. S. Grosberg and Dr. G. Carpenter for providing them with valuable information as well as for the useful discussions. Additionally the authors would like to thank C. E. Metz, Ph.D., for providing the LABROC1 and CLABROC programs.

References

[1] H. C. Zuckerman, "The role of mammography in the diagnosis of breast canser," in *Breast Canser, Diagnosis and Treatment*, I. M. Ariel and J. B. Cleary, Eds. New York: McGraw-Hill, 1987, pp. 152–172.
[2] D. B. Kopans, "The positive predictive value of mammography," *Amer. J. Roentgenol.*, vol. 158, pp. 521–526, 1992.
[3] D. D. Adler and M. A. Helvie, "Mammographic biopsy recommendations," *Curr. Opin. Radiol.*, vol. 4, pp. 123–129, 1992.
[4] M. Moskowitz, "Impact of a priory medical detection on screening for breast cancer," *Radiology*, vol. 184, pp. 619–622, 1989.
[5] P. A. Lachenbruch, *Discriminant Analysis*. New York: Hafner, 1975.
[6] R. O. Duda, and P.E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
[7] P. J. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral sciences," Ph.D. dissertation, Harvard Univ., Cambridge, MA, 1974.
[8] D. Rumelhart, G. E. Hinton, and R. J. Williams, in D. E. Rumelhart, Ed., *Parallel and Distributed Processing*. Cambridge, MA: MIT Press, 1986, vol. 1, p. 318.
[9] J. Herz, A. Krogh, and R. Palmer, *Introduction to the Theory of Neural Computation*. Reading, MA: Addison-Wesley, 1991.
[10] H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminat analysis in texture feature space," *Phys. Med. Biol.*, vol. 40, pp. 857–876, 1995.
[11] D. Wei, H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis," *Med. Phys.*, vol. 22, pp. 1501–1513, 1995.
[12] B. Sahiner, H. P. Chan, N. Petick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mamograms: The rubber band sraightening transform and texture analysis," *Med. Phys.*, vol. 25, no. 4, pp. 516–526, Apr. 1998.
[13] B. Sahiner, H. P. Chan, D. Wei, N. Petick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue," *Med. Phys.*, vol. 23, no. 10, pp. 1671–1683, Oct. 1996.
[14] H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Lam, D. D. Adler, and M. M. Goodsitt, "Computerized classification of malignant

and benign microcalsifications on mammograms: Texture analysis using an artificial neural network," *Phys. Med. Biol.*, vol. 42, pp. 549–567, 1997.

[15] R. M. Rangayyan, N. M. El-Farmawy, J. E. Desautels, and O. A. Alim, "Measures of acutance and shape for classification of breast tumors," *IEEE Trans. Med. Imag.*, vol. 16, pp. 799–810, Dec. 1997.

[16] D. B. Fogel, E. C. Wasson, E. M. Boughton, V. W. Porto, and P. J. "Angeline, linear and neural model for classifying breast masses," *IEEE Trans. Med. Imag.*, vol. 17, pp. 485–488, June 1998.

[17] R. P. Highnam, J. M. Brady, and B. J.Shepstone, "A quantitative feature to aid diagnosis in mammography," in *Proc. Digital Mammography'96*, pp. 201–206.

[18] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," *Radiology*, vol. 187, pp. 81–87, 1993.

[19] V. Goldberg, A. Manduca, D. L. Evert, J. J. Gisvold, and J. F. Greenleaf, "Improvements in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence," *Med. Phys.*, vol. 19, pp. 1475–1481, 1992.

[20] J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computerized image analysis," *IEEE Trans. Med. Imag.*, vol. 12, pp. 664–669, Dec. 1993.

[21] M. F. McNitt-Gray, H. K. Huang, and J. W. Sayre, "Feature selection in the pattern classification problem of digital chest radiograph segmentation," *IEEE Trans. Med. Imag.*, vol. 14, pp. 537–547, Sept. 1995.

[22] Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," *Acad. Radiol.*, vol. 5, pp. 155–168, 1998.

[23] M. Jordan, and R. A. Jacobs, "Hierarchical mixture of experts and EM algorithm," *Neural Comput.*, vol. 6, pp. 181–214, 1994.

[24] L. Hadjiiski and P. Hopke, "Design of large scale models based on multiple neural network approach," *Intelligent Engineering Systems Through Artificial Neural Networks*. ASME, 1997, vol. 7, pp. 61–66.

[25] S. Grossberg, "Adaptive pattern classification and universal recoding. I. Parallel development and coding of neural feature detectors," *Biolog. Cybern.*, vol. 23, no. 3, pp. 121–134, 1976.

[26] G. A. Carpenter and S. Grossberg, "ART 2: Self-organization of stable category recognition codes for analog input patterns," *Appl. Opt.*, vol. 26, no. 23, 1, pp. 4919–4930, Dec. 1987.

[27] G. A. Carpenter, S. Grossberg, and D. B. Rosen, "ART 2-A: An adaptive resonance algorithm for rapid category learning and recognition," *Neural Networks*, vol. 4, no. 4, pp. 493–504, 1991.

[28] G. A. Carpenter and S. Grossberg, "Integrating symbolic and neural processing in a self-organizing architeture for pattern recognition and prediction," in *Artificial Intelligence and Neural Networks: Steps toward Principled Integration*. New York: Academic, 1994.

[29] G. A. Carpenter and N. Markuzon, "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases," *Neural Networks*, vol. 11, no. 2, pp. 323–336, Mar. 1998.

[30] Y. Xie, P. K. Hopke, and D. Wienke, "Airborne particle classification with a combination of chemical composition and shape index utilizing an adaptive resonance artificial neural network," *Environ. Sci. Technol.*, vol. 28, no. 11, pp. 1921–1928, 1994.

[31] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. 3, pp. 610–621, Nov. 1973.

[32] M. M. Galloway, "Texture analysis using gray level run length," *Comput. Graph. Image Processing*, vol. 4, pp. 172–179, 1975.

[33] M. J. Norusis, *SPSS Professional Statistics 6.1*. Chicago, IL: SPSS, 1993.

[34] M. M. Tatsuoka, "Multivariate Analysis," *Techniques for Educational and Psychological Research*. New York: Macmillan, 1988.

[35] C. E. Metz, "ROC methodology in radiographic imaging," *Invest. Radiol.*, vol. 21, pp. 720–733, 1986.

[36] C. E. Metz, J. H. Shen, and B. A. Herman, "New methods for estimating a binomial ROC curve from continuously distributed test results," presented at the *1990 Annu. Meeting American Statistical Association*, Anahaim, CA, 1990.

[37] B. Sahiner, H. P. Chan, N. Petrick, R. Wagner, and L. Hadjiiski, "The effect of sample size on feature selection in computer-aided diagnosis," *Proc. SPIE*, vol. 3661, pp. 499–510, 1999.