

# Parameter optimization of a computer-aided diagnosis scheme for the segmentation of microcalcification clusters in mammograms

Marios A. Gavrielides<sup>a)</sup>

*Department of Biomedical Engineering, Duke University, Durham, North Carolina 27708*

Joseph Y. Lo and Carey E. Floyd, Jr.

*Digital Imaging Research Division, Department of Radiology, Duke University Medical Center, Durham, North Carolina 27710 and Department of Biomedical Engineering, Duke University, Durham, North Carolina 27708*

(Received 9 May 2001; accepted for publication 11 January 2002; published 12 March 2002)

Our purpose in this study is to develop a parameter optimization technique for the segmentation of suspicious microcalcification clusters in digitized mammograms. In previous work, a computer-aided diagnosis (CAD) scheme was developed that used local histogram analysis of overlapping subimages and a fuzzy rule-based classifier to segment individual microcalcifications, and clustering analysis for reducing the number of false positive clusters. The performance of this previous CAD scheme depended on a large number of parameters such as the intervals used to calculate fuzzy membership values and on the combination of membership values used by each decision rule. These parameters were optimized empirically based on the performance of the algorithm on the training set. In order to overcome the limitations of manual training and rule generation, the segmentation algorithm was modified in order to incorporate automatic parameter optimization. For the segmentation of individual microcalcifications, the new algorithm used a neural network with fuzzy-scaled inputs. The fuzzy-scaled inputs were created by processing the histogram features with a family of membership functions, the parameters of which were automatically extracted from the distribution of the feature values. The neural network was trained to classify feature vectors as either positive or negative. Individual microcalcifications were segmented from positive subimages. After clustering, another neural network was trained to eliminate false positive clusters. A database of 98 images provided training and testing sets to optimize the parameters and evaluate the CAD scheme, respectively. The performance of the algorithm was evaluated with a FROC analysis. At a sensitivity rate of 93.2%, there was an average of 0.8 false positive clusters per image. The results are very comparable with those taken using our previously published rule-based method. However, the new algorithm is more suited to generalize its performance on a larger population, depends on two monotonic outputs making its evaluation much easier and can be trained in an automatic way making practical its application on a large database. © 2002 American Association of Physicists in Medicine. [DOI: 10.1118/1.1460874]

## I. INTRODUCTION

According to recently published statistics,<sup>1</sup> breast cancer is still the leading cause of cancer death among women aged 40 to 59 and the second cause of cancer death overall behind lung cancer. Breast cancer is expected to account for 192 200 new cancer cases (or 31% of all new cancer cases) in 2001. However, breast cancer mortality has been decreasing an average of 2.2% per year between 1990 and 1997,<sup>1</sup> and screening mammography has contributed to this outcome by enabling the detection of breast cancer at its early stage. At present, screening mammography is the only reliable and practical method used for the early diagnosis of breast cancer. Despite its success, mammography has limitations: approximately 10%–30% of breast cancers retrospectively visible on the mammograms were missed or misinterpreted due to human or technical factors.<sup>2</sup> A suggested way for improving diagnostic accuracy is the double reading of mammograms. Previous studies have shown that two radiologists working together perform significantly more accurately than one alone.<sup>3</sup> While double reading is not economically fea-

sible in a clinical environment, an accurate computer aid could provide some of the benefit of a second reader at an acceptable cost and thus improve the accuracy of the diagnosis.

One of the most important and sometimes the only sign of breast cancer on mammograms is the presence of microcalcification clusters.<sup>4</sup> In a clinical study by Sickles,<sup>5</sup> more than half of nonpalpable cancers had mammographically visible calcifications, and in 36% of nonpalpable cancers, calcifications were the only sign of abnormality. In another study of cancers missed in screening mammography,<sup>6</sup> the presence of microcalcifications was the predominant feature in 18% of missed cancers. The task of detecting microcalcifications on mammograms can be difficult due mainly to their small size, low contrast, and the similarity of their radiographic appearance to dense tissue.

The clinical significance of microcalcification clusters and the risk of misdiagnosis have prompted significant research efforts during the last decade. These methods typically include a preprocessing step for noise suppression and contrast

enhancement, a feature extraction step for locating suspicious signals, and a feature analysis step to reduce the number of false positive signals. Preprocessing is usually done using conventional image processing and filtering methods. For the feature extraction step several techniques have been developed, including a difference-image technique,<sup>7</sup> local area thresholding,<sup>8</sup> wavelet transform-based methods,<sup>9-11</sup> image fuzzification and morphological operators,<sup>12</sup> statistical texture analysis,<sup>13</sup> and Laplacian scale-space signatures.<sup>14</sup> Methods to reduce the number of false positive clusters include rule-based techniques,<sup>8</sup> artificial neural networks,<sup>2,15</sup> and a combination of neural networks and rule-based systems.<sup>16</sup>

We have recently developed a multistage algorithm for the segmentation of suspicious microcalcification clusters in digitized mammograms.<sup>17</sup> The algorithm consisted of three main steps: (a) image preprocessing, (b) segmentation of individual microcalcifications, and (c) clustering and the removal of false positive clusters. In the first step, the breast region was segmented and unsharp masking was applied to enhance potential microcalcification signals within it. The second step operated sequentially on subimages. For each subimage, the gray level histogram was analyzed using a set of histogram features and a fuzzy rule-based classifier, in order to identify subimages containing microcalcifications and set appropriate local thresholds to segment any microcalcifications within them. The decision rules of the classifier involved the fuzzy membership values of the histogram features within specific intervals. Certain combinations of fuzzy membership values were selected to provide decision rules. In the last step of the algorithm, individual microcalcifications were clustered and a set of features was extracted for each cluster. Another fuzzy rule-based classifier used the cluster features to remove false positive or typically benign clusters. The output of the algorithm was in the form of a binary image containing the suspicious microcalcification clusters. The performance of the algorithm was evaluated using a database of 98 images, with 48 images containing one or more microcalcification clusters.

Even though the results of the above-mentioned technique were quite encouraging, further development was needed in order to meet the overall objective of this project, which was to develop an automated computer-aided diagnosis (CAD) scheme for the detection of suspicious microcalcification clusters in mammograms. Each stage of the previous computer scheme used a number of parameters, which were optimized empirically based on the performance of the algorithm on the training set. Manual training was time consuming and would make it impractical to retrain the algorithm when one of its stages was modified or when the algorithm was applied on another database with different image characteristics. Moreover, the combinations of features in each rule were chosen empirically, in a way that does not necessarily result in the best combinations for maximum performance over the training set, nor necessarily permit generalization to an independent testing set. Finally, the rule-based algorithm results in a particular combination of multiple decision variables, making it difficult for its performance to be

evaluated using ROC/FROC analysis, which is most easily implemented when a single threshold can be varied.

In order to overcome the above limitations, a new algorithm was developed in a manner explained in detail in this manuscript. The new algorithm was improved over the first one in the following ways. First, the rule-based classifier used in the previous study for segmentation of individual microcalcifications, was replaced with a neural network with fuzzy-scaled inputs. The fuzzy-scaled inputs were created by processing the histogram features with a set of membership functions, the parameters of which were extracted from the distribution of the feature values in a training set. The neural network was trained to combine these fuzzy inputs (or feature memberships) in order to classify feature vectors as either positive (vectors taken from subimages containing one or more microcalcifications) or negative (vectors taken from subimages containing no microcalcifications). Second, the rule-based classifier used in the previous study for the classification of microcalcification clusters as true positive or false positive, was replaced by another neural network that was trained to classify clusters using a set of cluster features.

With these substitutions, the significant parameters of both the individual microcalcification segmentation stage as well as the cluster classification stage were embedded into neural network weights, which could be optimized through the gradient decent technique. The continuous-valued output of the second neural network allowed the performance of the modified CAD scheme to be evaluated using FROC analysis on a set of digitized mammograms.

## II. MATERIALS

A database of 98 images, provided by the Digital Medical Imaging Program, University of South Florida and Moffitt Cancer Center and Research Institute was used for the development and evaluation of the algorithm. It contained 48 images with a total of 55 biopsy-proven malignant microcalcification clusters, 42 control images with no sign of abnormality after at least two consecutive annual screening tests, and 8 images with benign clusters. The benign cases were not biopsied, indicating that they were not considered suspicious. The images were digitized at 105  $\mu\text{m}$  and 12 bits per pixel with an NDT Scan II scanner (DBA, Melbourne, FL). The true locations of the malignant clusters, defined as three or more microcalcifications within 1  $\text{cm}^2$ , were marked on copies of the images by an experienced mammographer. This database has been described previously by Kallergi *et al.*<sup>18</sup> and was used in our previous study.<sup>17</sup>

In the multistage implementation described in the next section, a classifier in the middle stage had to be trained to distinguish between positive and negative subimages representing individual microcalcifications. The supervised training and evaluation algorithms used to optimize this stage, needed ground truth information describing the location of the individual microcalcifications. While establishing ground truth for suspicious clusters is typically performed by an experienced mammographer, the task of identifying individual microcalcifications is usually not provided since they do not

carry clinical significance on their own, but only as part of a cluster. In addition, this task represents an unreasonable burden of effort on a human investigator given the large number of individual calcifications that typically are present. As a practical solution, we used the corresponding intermediate output of our previous rule-based CAD scheme. Specifically, those subimages that were both classified as positive by the CAD scheme and belonged to a biopsy-proven malignant cluster, were labeled as positive subimages. This approach identifies many true positive individual microcalcifications, while also creating a few false positives and false negatives. It was assumed that the majority of the positives located within a biopsy-proven malignant cluster would be true positive, and moreover that all positives would be suspicious enough that they would be appropriate for training the classifier. Examples of negative subimages were generated by drawing squares in areas of normal images and extracting the histogram vectors from subimages within those squares. Care was taken to select these normal subimages from regions in the mammograms that included both dense and fat parenchyma. The histogram features were extracted from the positive and negative subimages and were used as inputs to the classifier. Subimages from the small number of images containing benign clusters were included as negative examples for testing purposes even though they were not used as training examples.

In order to enable a comparison with our previous study, we used the same training and testing sets as before. The training set consisted of 24 images (10 malignant, 10 normal, and 4 benign) with a total of 10 malignant microcalcification clusters and 1664 positive subimages. A total of 2030 negative subimages were drawn from normal images. The testing set consisted of 74 images (38 malignant, 32 normal, and 4 benign) containing 44 malignant clusters.

### III. METHODS

#### A. Preprocessing

The preprocessing stage of the algorithm remained unchanged as the one used in Ref. 17 and is briefly described here. All images were preprocessed using the following two steps: First, the breast region was segmented using a method by Bick *et al.*<sup>19</sup> The objective for this step was to reduce the overall processing time by further analyzing only the breast region and to eliminate possible sources of false positives such as view markers. Then, the high spatial frequency content of the segmented breast region was enhanced using unsharp masking, implemented by smoothing the image with a median filter of size  $11 \times 11$  pixels and then subtracting the smooth image from the original.

#### B. Segmentation of individual microcalcifications

The same features used in the rule-based classifier for the segmentation of individual microcalcifications were also used in the new algorithm. However, the new algorithm employed an automated method for feature scaling and replaced

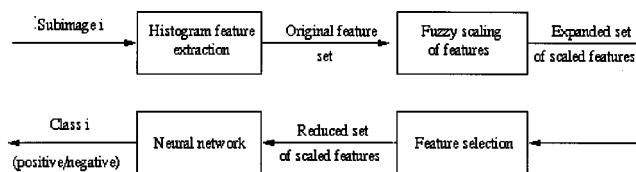


FIG. 1. Block diagram of the algorithm used for the segmentation of individual microcalcifications.

the rule-based classifier with an artificial neural network with fuzzy inputs as will be described in this section.

Individual microcalcifications were segmented using local thresholds assigned to subimages that were classified as positive. Subimages were classified as either positive or negative using the algorithm shown in Fig. 1. As a first step in developing this classifier, a set of eight histogram features was extracted from each training subimage. Then, the histogram features were processed using a family of fuzzy membership functions, resulting in an expanded set of scaled features. Stepwise discriminant analysis was then used to eliminate redundant features. Finally, a neural network was trained to distinguish between positive and negative subimages using the reduced set of features as inputs. After the classifier was trained, it was applied to overlapping subimages from each full image in the testing set. Individual microcalcifications were segmented from subimages classified as positive.

#### 1. Histogram feature extraction

A set of histogram features was extracted for overlapping subimages, the size of which was chosen as  $16 \times 16$  pixels based on clinical information about the size of microcalcifications and the resolution of the images. Specifically, breast cancers rarely produce calcifications larger than 1 mm, and most are under 0.5 mm in diameter.<sup>20</sup> The test images were digitized at  $105 \mu\text{m}$  per pixel, giving a 1-mm object an image size of about 10 pixels. A subimage with a size of  $16 \times 16$  pixels was found to be both large enough to include calcifications of clinical interest as well as small enough to enable individual microcalcifications to affect the local histogram. All subimages were allowed to overlap in order to minimize the risk of missing any microcalcifications. The degree of overlap was chosen experimentally as 75%.<sup>17</sup> It has to be noted that the size of the each subimage and the degree of their overlap, were not included in the parameter optimization scheme since they can be considered constant assuming the calcifications of clinical interest have a certain size as described above. For images of higher resolution or pixel value, the size of subimages would have to be increased. Similarly, if smaller microcalcifications were to be targeted, the degree of overlap would have to be increased.

Subimages with very small variance of pixel intensities were assumed to have no microcalcifications present. For all other subimages, features were extracted from their gray level histograms as follows. First, the histogram was smoothed with a moving average filter. Then, a *main lobe* was identified as the histogram area in the range  $[h_1, h_2]$ , where  $h_1$  and  $h_2$  were the lower and upper nonzero histo-

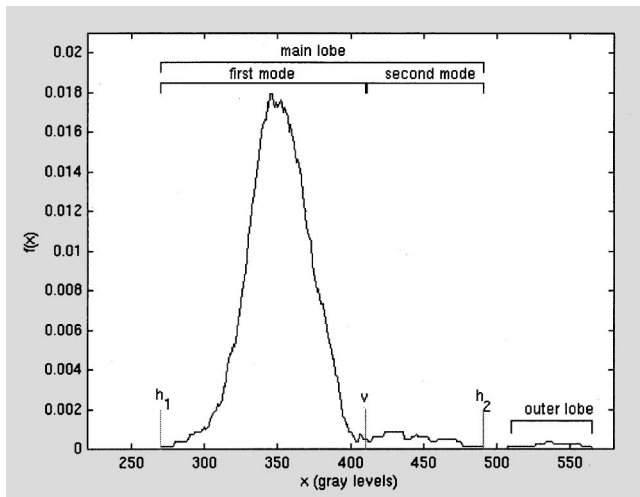


FIG. 2. Notation used for local histogram analysis. This example histogram was taken from a  $16 \times 16$  subimage containing microcalcifications.

gram limits, respectively, starting from the peak of the smoothed histogram. Any nonzero values beyond  $h_2$  defined an *outer lobe*. If a minimum within the histogram range  $[h_1, h_2]$  was found, the *main lobe* was considered to have two modes, defined as *first* and *second mode*, separated by this minimum. An estimate for the optimal gray level  $\nu$  separating the two modes was found using a thresholding technique,<sup>21</sup> where the histogram in the range  $[h_1, h_2]$  was fitted by a mixture density. The parameters of this fit were estimated with a least squares method for nonlinear models.<sup>22</sup> In the case where no minimum was found between  $[h_1, h_2]$ , the parameter  $\nu$  was set equal to  $h_2$ . The above histogram parameters are illustrated in Fig. 2 for an example histogram.

The presence of microcalcifications in a subimage and their size and contrast relative to the surrounding tissue are reflected in the histogram of that subimage. A more detailed discussion on the properties of histograms taken from positive and negative subimages, along with a set of examples is included in Ref. 17. Briefly, histograms from subimages containing microcalcifications appear to have more separation between the *first* and *second* mode compared to histograms from negative images, a larger spread of the *second modes*, and a nonzero *outer lobe*. In order to describe these differences in a quantitative way, a set of eight histogram features was extracted for each subimage. The histogram features are defined in Table I. The ratios,  $ptvr$ ,  $\sigma_{2,1}$ , and  $\sigma_{3,1}$  were inverted versions of the corresponding  $vtpr$ ,  $\sigma_{1,2}$ ,  $\sigma_{1,3}$  features defined in Ref. 17, so that a larger value for each feature extracted from a subimage corresponds to a greater likelihood of that subimage being positive. Histograms with  $a_1 > 0.99$  were assumed to be unimodal and their corresponding subimages were classified as negative. The remaining subimages were classified as positive or negative using the procedure described in the following sections.

## 2. Automated feature scaling using fuzzy membership functions

The rule-based classifier of our previously published method for the detection of individual microcalcifications employed fuzzy-scaled histogram features. The scaling was done by calculating the fuzzy membership of each feature within certain intervals, which were selected based on empirical observations. Even though the performance of the rule-based classifier supported the use of the fuzzy-scaled features, the scaling procedure was time consuming and not necessarily optimal. In this section, a method for automatic scaling of the histogram features, will be described. Moreover, an experiment will be presented, testing the hypothesis that scaling the histogram features within certain intervals increases the discriminating ability of the histogram features.

Feature scaling was done using fuzzy membership functions in a manner analogous to window and leveling, using three window and three level settings for a total of nine combinations of membership functions. The membership functions used for scaling the histogram features were different versions of the  $S$  function given below:

$$\mu_A(y) = \mathfrak{S}(\alpha, \beta) = \begin{cases} 0, & \beta \leq -\alpha, \\ 2 \left( \frac{\beta + \alpha}{2\alpha} \right)^2, & -\alpha < \beta \leq 0, \\ 1 - 2 \left( \frac{\beta - \alpha}{2\alpha} \right)^2, & \alpha > \beta > 0, \\ 1, & \beta \geq \alpha, \end{cases}$$

where  $\mu_A(y)$  is the membership value of feature value  $y$  within interval  $\{A\}$ ,  $\beta = y - y_C$ ,  $\alpha = \min\{y_{\max} - y_C, y_C - y_{\min}\}$ ,  $y_{\max}$  and  $y_{\min}$  are the largest and smallest feature values, respectively, defining interval  $\{A\}$ , and  $y_C$  is the crossover point for which the membership value is 0.5.

By setting  $y_{\min}$  to zero, the function above depends on two parameters,  $y_{\max}$  and  $y_C$ . For this study, we used values for  $y_{\max}$  taken from the distribution of the feature to be scaled. Specifically,  $y_{\max}$  was allowed to take three values, (a)  $mean + 0.5 * std$ , (b)  $mean + 1.0 * std$ , and (c)  $mean + 1.5 * std$ , where  $mean$  and  $std$  were taken from the distribution of the feature values in the training set. Parameter  $y_C$  was allowed to take three distinct values, (a)  $0.25 * y_{\max}$  (b)  $0.50 * y_{\max}$ , and (c)  $0.75 * y_{\max}$ . Using all the combinations of  $y_{\max}$  and  $y_C$ , a total of nine membership functions were created to scale each of the histogram features. Figure 3 shows the distribution of the nine membership functions for an example feature array and Table II summarizes the parameters used to create the nine membership functions. The result of the scaling procedure was an expanded set of 80 features, including the 72 fuzzy-scaled features (8 features each scaled by the 9 different membership functions) and the 8 raw histogram features.

In order to validate the hypothesis that the expanded set of fuzzy-scaled features had a better discrimination ability compared to the set of the eight raw histogram features, we classified the training set of vectors using linear discriminant analysis (LDA) and cross-validated the results for perfor-

TABLE I. Histogram features extracted from each 16×16 pixel subimage.

Feature description	Mathematical description
Area outside the first mode, toward the high end of the histogram, $a_1$	$a_1 = \sum_{x=\nu}^{h_{\max}} f(x),$
Area of second mode of the main lobe, $a_2$	where $h_{\max}$ denotes the maximum gray level $a_2 = \sum_{x=\nu}^{h_2} f(x)$
Area of the outer lobe, $a_3$	$a_3 = \sum_{x=h_2+1}^{h_{\max}} f(x)$
Peak to valley ratio, $ptvr$	$ptvr = \frac{\max \text{ of } f(x) \text{ in } [h_1, h_2]}{f(\nu)}, \quad f(\nu) > 0$
Mean contrast of the second mode of main lobe relative to the first, $con_{1,2}$	$con_{1,2} = \frac{m_2 - m_1}{m_2 + m_1} \text{ where,}$ $m_1 = \frac{1}{a_1} \sum_{x=h_1}^{\nu} x \cdot f(x) \text{ and}$ $m_2 = \frac{1}{a_2} \sum_{x=\nu}^{h_2} x \cdot f(x)$
Ratio of the standard deviations of the two modes of the main lobe, $\sigma_{2,1}$	$\sigma_{2,1} = \frac{\sigma_2}{\sigma_1} \text{ where,}$ $\sigma_1 = \sqrt{\frac{1}{a_1} \sum_{x=h_1}^{\nu} f(x) \cdot (x - m_1)^2} \text{ and}$ $\sigma_2 = \sqrt{\frac{1}{a_2} \sum_{x=\nu}^{h_2} f(x) \cdot (x - m_2)^2}$
Mean contrast of the outer lobe relative to the first mode of the main lobe, $con_{1,3}$	$con_{1,3} = \frac{m_3 - m_1}{m_3 + m_1}, \text{ where,}$ $m_3 = \frac{1}{a_3} \sum_{x=h_2+1}^{h_{\max}} x \cdot f(x)$
Ratio of the standard deviations of the outer mode and the first mode of the main lobe, $\sigma_{3,1}$	$\sigma_{3,1} = \frac{\sigma_3}{\sigma_1}, \text{ where}$ $\sigma_3 = \sqrt{\frac{1}{a_3} \sum_{x=h_2+1}^{h_{\max}} f(x) \cdot (x - m_3)^2}$

mance evaluation. The LDA procedure was implemented using SAS software (SAS Institute, Cary, NC). The performance of the LDA classifier was measured and compared using two indices of performance: (a) the ROC curve area index  $A_z$  and (b) the ROC partial area index,  $pA_z$ . The value of the  $A_z$  index can be interpreted as the average value of sensitivity over all possible values of specificity or, alternatively, as the average value of specificity over all possible values of sensitivity.<sup>23</sup> Even though the  $A_z$  index is often used to summarize the diagnostic performance as described with the ROC curve, it may not be a meaningful measure for some situations where high sensitivity is required. The index  $pA_z$ <sup>24</sup> was used in addition to the  $A_z$  index because in this intermediate step of detecting individual microcalcifications, very high sensitivity was required. The  $pA_z$  index summarizes the high sensitivity region of the ROC curve, by providing the average specificity in the sensitivity region between 0.9 and 1, where 1 denotes perfect sensitivity. CLABROC software (provided by Charles Metz, University

of Chicago, IL) was used for the ROC analysis.

The results of the LDA for the two feature sets are given in Table III. It can be seen from the table that using the expanded set of features improved the performance of both indices in a statistically significant way. This result supports our hypothesis that the discriminating ability of the histogram features as used here for segmenting individual microcalcifications, is increased when they are scaled using certain membership functions.

### 3. Feature selection

The automatic scaling of the histogram features produced an expanded set of 80 fuzzy-scaled features. Even though the expanded set was shown to significantly outperform the set of raw histogram features, it was possible that some of the features were redundant. In order to minimize redundancies on the feature set, we used stepwise discrimination analysis (SDA). The SDA procedure began with no selected features

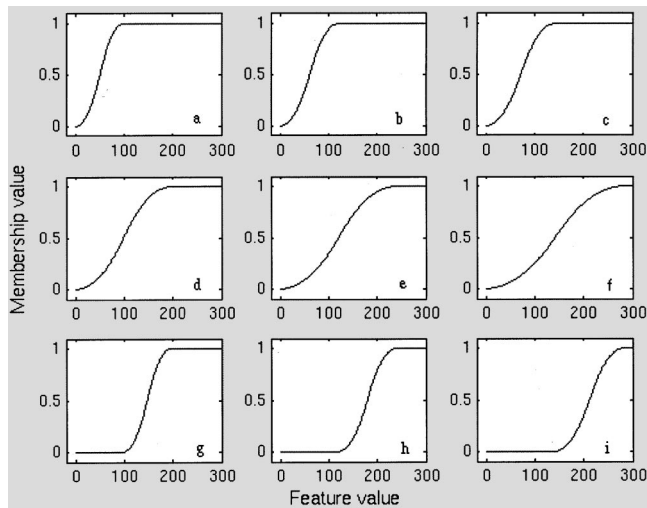


FIG. 3. Feature scaling using a family of fuzzy membership functions. Plots (a)–(i) show the distribution of fuzzy membership values of an example array, calculated using the  $S$  function given in Sec. III B 2 with the corresponding parameters of Table III.

and then added or removed a feature at each step. A feature already in the selected set was removed if it did not significantly improve the discriminating power as measured by the Wilks'  $\Lambda$  criterion<sup>25</sup> statistical criterion. If no feature was removed at a given step, then the feature that improved most the discriminating power was added to the set. The procedure stopped when no feature was either added or removed. The SDA procedure was implemented using SAS software (SAS Institute, Cary, NC).

By identifying this reduced subset of features, the resulting models would have less complexity in terms of parameters to be optimized, which in turn should reduce bias due to overtraining on finite training data and improve the ability of the model to generalize to new data.

#### 4. Neural network for subimage classification

A neural network (ANN1) merged the selected fuzzy-scaled features to predict whether a subimage was positive or negative. The neural network in this stage replaced the role of the rule-based classifier in our previous study. Instead of a set of distinct rules resulting in a multiple number of deci-

TABLE II. Parameters used to create the fuzzy membership functions  $a$  to  $i$  shown in Fig. 3. The parameters  $mean$  and  $std$  were taken directly from the distribution of the feature to be scaled.

Membership function	$y_{max}$	$y_t$
$a$	mean + 0.5*std	0.25* $y_{max}$
$b$	mean + 1.0*std	0.25* $y_{max}$
$c$	mean + 1.5*std	0.25* $y_{max}$
$d$	mean + 0.5*std	0.50* $y_{max}$
$e$	mean + 1.0*std	0.50* $y_{max}$
$f$	mean + 1.5*std	0.50* $y_{max}$
$g$	mean + 0.5*std	0.75* $y_{max}$
$h$	mean + 1.0*std	0.75* $y_{max}$
$i$	mean + 1.5*std	0.75* $y_{max}$

TABLE III. A comparison of classifier performance using linear discriminant analysis for (a) an original set of 8 raw histogram features and (b) an expanded set of 80 features (72 fuzzy-scaled+8 raw histogram features). The indexes of performance  $A_z$ , defined as the ROC area index, and  $pA_z$ , defined as the ROC partial area index describing the high sensitivity region (0.9–1.0) of the ROC curve, were used for the comparison.

Feature set	$A_z$	$pA_z$
Original set	0.936±0.004	0.733±0.016
Expanded set	0.962±0.003	0.860±0.009
$p$ value	<0.001	<0.001

sion variables, the network resulted in a function of weighted inputs and a monotonic output value that could be used as a decision variable. The weights were determined by training the network using supervised backpropagation with the delta rule, where the network was presented with a set of examples with known outputs. As stated before, there were 3694 feature vectors, each described by the fuzzy-scaled input features selected from the SDA procedure. During training, the error between the predicted output and the known output was used to modify all the weights of the network via the backpropagation algorithm. This process was repeated over many iterations (defined as one complete presentation of all training cases) until the cross-validation error was minimized. Target values ranged between 0 for negative subimages and 1 for positive subimages.

The ANN1 was a three-layer feedforward network, consisting of an input layer of the fuzzy-scaled inputs, a hidden layer of up to ten nodes and the output layer with one node. The performance of the classifier was evaluated with a 20-fold cross-validation of the training set of subimages in order to determine network parameters such as the number of hidden nodes, momentum, learning rate, and number of iterations. For the 3694 cross-validation testing outputs, the sensitivity and specificity over a range of decision thresholds were expressed as a ROC curve, from which  $A_z$  and  $pA_z$  were calculated as indices of performance.

#### 5. Testing on full images

Once its architecture was determined, the ANN1 classifier was trained on all the training vectors and tested on all subimage vectors from the full images of the testing set. Subimages with  $a_1 > 0.99$  were not tested since they were already classified as negative. The remaining testing vectors were prescaled with the membership functions, as determined in the previous section. Note that although there were 3694 training subimages, they only covered a relatively small fraction of the breast area in a full image. In comparison, a full image contained an average of 99 700 subimages (std. deviation of 38 486). The result of this step was that for all testing images, a corresponding number of files were created containing the network outputs of their subimages. By applying a threshold on the network outputs, defined as  $ann\_seg$ , subimages were classified as either positive or negative.

From each positive subimage, individual microcalcifications were segmented by assigning the valley  $\nu$  of its histogram (identified earlier in the feature extraction section) as a local threshold. Due to the 75% overlap of the subimages, each pixel in the image was selected up to 16 times. Pixels that were selected less than a specified number of times, defined as *overlap threshold*, were set to zero. The variability in the pixel intensities was used as a feature in the false positive cluster reduction stage, as will be explained in the next section.

In a similar fashion, the ANN1 classifier was used to segment individual microcalcifications in the full area of the training images. These segmented microcalcifications were grouped in order to produce training microcalcification clusters, which were used in the next stage of the algorithm.

### C. Clustering and false positive cluster reduction

In the final stage of the algorithm, segmented individual microcalcifications were clustered and false positive clusters were eliminated using the following procedure. First, microcalcifications were clustered using nearest neighbor clustering with a distance of 0.5 cm. Then, for each cluster a set of three features was extracted and served as input to a neural network which classified clusters as true malignant or false positive. After false positive clusters were eliminated, microcalcifications in the remaining clusters were clustered again using nearest neighbor clustering with a distance of 1.0 cm to allow merging of close clusters.

#### 1. Cluster feature extraction

Individual segmented microcalcifications were grouped into clusters using nearest neighbor clustering. Specifically, a microcalcification was assigned to the closest cluster provided that the Euclidean distance between it and the nearest member of that cluster was less than 0.5 cm. If a microcalcification was not assigned to a cluster, a new cluster was formed and the procedure continued until all individual microcalcifications were grouped. Clusters with less than three microcalcifications were removed.

For each cluster, three features were extracted: (a) number of microcalcifications in the cluster, *numc*, (b) average distance between microcalcifications within the cluster, *mcd*, and (c) the average number of times the segmented pixels within the cluster were selected during the segmentation of individual microcalcifications, *mism*. In a previous clinical study,<sup>26</sup> the first two features were found to be the most important for discriminating between benign and malignant clusters. The *numc* feature was categorized into four groups: (1) 3 microcalcifications, (2) 4–5, (3) 6–9, and (4)  $\geq 10$ , similar to the approach used in Ref. 26.

#### 2. Neural network for cluster classification

A second artificial neural network (ANN2) used the three cluster features to eliminate false positive clusters. The ANN2 classifier was a three-layer feedforward neural network trained using back-propagation, consisting of the input layer of the three cluster features, a hidden layer of up to five

hidden nodes and the output layer with one node. Output values ranged between 0 and 1 with true malignant clusters generating larger outputs.

The ANN2 was trained using true positive and false positive cluster vectors extracted from the training images. Training cluster vectors were labeled as positive if they satisfied the following criterion: The ratio of the intersection of area X, defined as the area of the smallest rectangle enclosing the detected cluster, and area Y, defined as the area of the smallest rectangle enclosing the true cluster, over the union of areas X and Y, had to exceed a predetermined threshold. That threshold was set to 0.3 based on empirical observation that using the above criterion with that threshold value, a detected cluster included the majority of the microcalcifications of the true cluster and at the same time the area of the true cluster was not expanded. Training cluster vectors not satisfying the cluster criterion were labeled as negative. It has to be noted that microcalcification clusters used for training were produced directly from grouping segmented microcalcifications. The ground truth information was used just for determining which of these training clusters were positive and which negative.

In order to determine the network parameters of the ANN2 classifier, its performance was evaluated using the leave-one-out method on the set of training cluster vectors. This process resulted in a number of testing outputs equal to the sum of all positive and negative cluster vectors. By sweeping a threshold over the range of output values, a ROC curve was created from which  $A_z$  and  $pA_z$  were calculated as indices of performance.

#### 3. Final algorithm output

The final output of the CAD scheme was created in two steps. First, the ANN2 classifier was applied on cluster vectors from the testing images. Clusters with output values below a decision threshold, *ann\_clu*, were removed from the final output. Second, microcalcifications in the remaining clusters were clustered again using nearest neighbor clustering with the distance raised to 1.0 cm to allow merging of close clusters. The final output was in the form of a binary image containing the segmented microcalcification clusters.

### D. Performance evaluation

The performance of the overall CAD scheme was evaluated on the independent testing set that was not used in any of the optimization steps. A family of free response receiver operating characteristic (FROC) curves was formed from the classifier outputs of the two stages (segmentation of individual microcalcifications and clustering). FROC analysis<sup>27</sup> is a well-known approach for describing the performance of detection schemes and is a plot of *sensitivity* (% of true clusters detected by the scheme) and a corresponding number of false positive clusters per image (FPi). Each FROC curve was acquired by sweeping the *ann\_clu* threshold of the ANN2 classifier after a fixed threshold was used for the ANN1 classifier to segment individual microcalcifications. For each value of *ann\_clu*, a pair of *sensitivity* and FPi num-

bers were calculated and defined a point in the FROC curve. A family of FROC curves was generated for different values of the *ann\_seg* threshold of the ANN1 classifier.

## IV. RESULTS AND DISCUSSION

### A. Feature selection

Stepwise discrimination analysis was used on the expanded set of 80 features resulting in a reduced set of features that was subsequently used for the classification of subimages. From the 24 training images, the histogram feature extraction step resulted in 3694 subimages, consisting of 1694 positive cases and 2030 negative cases. These feature vectors were the input cases for the above fuzzy scaling and SDA procedures. The SDA procedure, implemented with SAS software, reduced the number of features describing each vector from 80 down to 33. A *p* value of 0.05 was used as a statistical significance level for feature removal and entry. It should be noted that none of the 33 selected features belonged in the original set of the eight raw histogram features.

### B. Performance evaluation

The algorithm was evaluated on the testing set after being trained exclusively on the training set, in order to get an assessment of the performance of the algorithm when applied to an unknown set of cases. The overall performance of the algorithm was described using the FROC curves of Fig. 4. The curve points were acquired by varying the output threshold, *ann\_clu* of the ANN2 cluster classifier, while keeping the value of the *ann\_seg* constant. Three FROC curves were produced using *ann\_seg* values of 0.3, 0.4, and 0.5. It can be seen from Fig. 4 that the best performance of the algorithm is achieved using an *ann\_seg* of 0.4. Using that threshold, the algorithm achieves the highest measure of sensitivity of 93.2% or 41/44 truly detected clusters at an average of about 0.8 false positive clusters per image.

The performance of the new algorithm is very comparable to the one achieved using our previous algorithm, with 93.2% sensitivity at 0.7 false positive clusters per image. However, by eliminating *ad hoc* rules, the new algorithm is less constrained compared to the rule-based system and is more suited to generalize its performance on a larger population. Another advantage of the new algorithm is its dependence on only two monotonic decision variables, the *ann\_seg* and *ann\_clu* thresholds of classifiers ANN1 and ANN2, respectively, making its evaluation much easier. This can become an important issue when we evaluate the performance of our algorithm on a public database and seek to compare its performance with other methods. Finally, the new algorithm can be trained in an automatic way making practical its application on a large database.

Future work includes the training and evaluation of the CAD scheme on a large dataset taken from the Digital Database for Screening Mammography (<http://marathon.csee.usf.edu/Mammography/Database.html>) (DDSM) database, which is now publicly available. Applying our al-

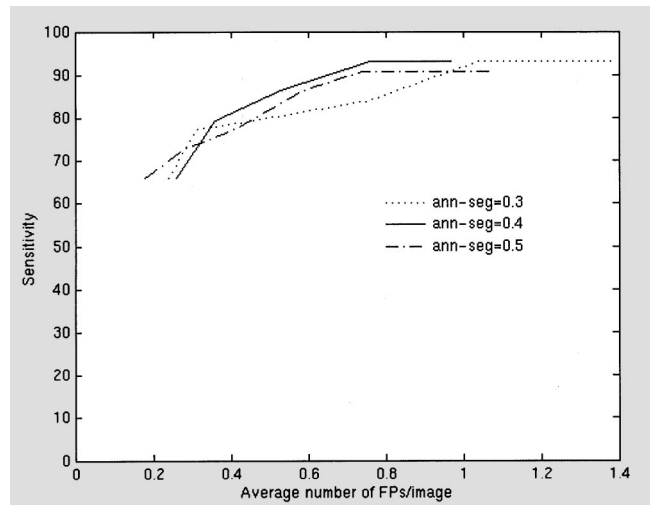


FIG. 4. A family of FROC curves describing the overall performance of the algorithm. Each curve was acquired by varying the threshold of the ANN2 classifier while keeping constant the threshold of the ANN1 classifier.

gorithm on the DDSM database will test the ability of the algorithm to generalize its performance on a large number of cases and will enable comparisons to be made with methods from other laboratories. Moreover, the DDSM includes images digitized with different scanners than the one used in this study, enabling observations to be made on the effect of digitizer type on the performance of the algorithm. Digitizer characteristics like pixel resolution may require changes in the size of subimages and the median filter used for preprocessing. The quality of the digitized images could also affect the performance of the algorithm, especially the presence of small, bright artifacts that could introduce false positive individual microcalcifications and change the values of cluster features such as the number and mean distance between microcalcifications. By retraining on examples from a new dataset acquired using another digitizer, the algorithm should be able to adapt to changes in the distribution of histograms due to the digitizer type. Also, some effort will be devoted into increasing the speed of the algorithm. The training of the new algorithm took much less time compared to the previous rule-based algorithm, but the execution of the new algorithm took about the same time, since it was mostly spent in the histogram analysis part of the algorithm. Using a threshold of 0.3 for the first classifier, it took an average of about 15 min to process each image using a single processor Ultra-60 work station (Sun Microsystems, Mountain View, CA). No effort has been devoted yet for making the algorithm software more efficient and fast. Other than reprogramming the algorithm software, the execution speed could be increased by replacing the histogram fitting part with a more direct way for finding the valley of each histogram.

## V. CONCLUSION

The new algorithm described in this paper performs comparably with our previously published algorithm. However, the neural network-based algorithm employed automated parameter optimization making practical the retraining and ap-

plication of the algorithm on other datasets and enabled the assessment of the algorithm performance using FROC analysis. The results of the performance evaluation of the CAD system and the fact that it can be trained without user interaction are encouraging for its use as an automatic computer tool for the early detection of breast cancer.

## ACKNOWLEDGMENT

This project was supported by Dissertation Research Award No. DISS 2000 729 from the Susan G. Komen Breast Cancer Foundation.

<sup>a)</sup>Corresponding author. Duke University Medical Center, DUMC 3302, Bryan Research Bldg., Room 135, Durham, North Carolina 27710. Phone: (919) 684-7751; fax: (919) 684-7122; electronic mail: Marios.Gavrielides@duke.edu

<sup>1</sup>R. T. Greenlee, M. Hill-Harmon, T. Murray, and M. Thun, "Cancer Statistics, 2001," *Ca-Cancer J. Clin.* **51**, 15–36 (2001).

<sup>2</sup>H. P. Chan, S. C. Lo, B. Sahiner, K. L. Lam, and M. A. Helvie, "Computer-aided detection of mammographic microcalcifications: pattern recognition with an artificial neural network," *Med. Phys.* **22**, 1555–1567 (1995).

<sup>3</sup>R. A. Schmidt, "The role of CAD in mammography and missed lesions," *Computer-aided Diagnosis in Medical Imaging*, edited by K. Doi *et al.* (Elsevier Science, Amsterdam, 1998), pp. 177–184.

<sup>4</sup>L. W. Bassett and S. Gambhir, "Breast imaging for the 1990s," *Semin. Oncol.* **18**, 80–86 (1991).

<sup>5</sup>E. A. Sickles, "Mammographic features of 300 consecutive nonpalpable breast cancers," *Am. J. Roentgenol.* **146**, 661–663 (1986).

<sup>6</sup>R. E. Bird, T. W. Wallace, and B. C. Yankaskas, "Analysis of cancers missed at screening mammography," *Radiology* **184**, 613–617 (1992).

<sup>7</sup>H. P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu, and H. MacMahon, "Improvement in radiologists' detection of clustered microcalcifications on mammograms: The potential of computer-aided diagnosis," *Invest. Radiol.* **25**, 1102–1110 (1990).

<sup>8</sup>D. H. Davies and D. R. Dance, "Automatic computer detection of clustered calcifications in digital mammograms," *Phys. Med. Biol.* **35**, 1111–1118 (1990).

<sup>9</sup>W. Qian, M. Kallergi, L. P. Clarke, H. D. Li, P. Venugopal, D. Song, and R. A. Clark, "Tree structured wavelet transform segmentation of microcalcifications in digital mammography," *Med. Phys.* **22**, 1247–1254 (1995).

<sup>10</sup>H. Yoshida, K. Doi, R. M. Nishikawa, M. L. Giger, and R. A. Schmidt, "An improved computer-assisted diagnostic scheme using wavelet transform for detecting clustered microcalcifications in digital mammograms," *Acad. Radiol.* **3**, 621–627 (1996).

<sup>11</sup>W. Zhang, H. Yoshida, R. M. Nishikawa, and K. Doi, "Optimally weighted wavelet transform based on supervised training for detection of

microcalcifications in digital mammograms," *Med. Phys.* **25**, 949–956 (1998).

<sup>12</sup>H.-D. Cheng, Y. M. Lui, and R. I. Freimanis, "A novel approach to microcalcification detection using fuzzy logic technique," *IEEE Trans. Med. Imaging* **17**, 442–450 (1998).

<sup>13</sup>J. K. Kim and H. W. Park, "Statistical texture features for detection of microcalcifications in digitized mammograms," *IEEE Trans. Med. Imaging* **18**, 231–238 (1999).

<sup>14</sup>T. Netsch and H.-O. Peitgen, "Scale-space signatures for the detection of clustered microcalcifications in digital mammograms," *IEEE Trans. Med. Imaging* **18**, 774–786 (1999).

<sup>15</sup>W. Zhang, K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt, "An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms," *Med. Phys.* **23**, 595–601 (1996).

<sup>16</sup>R. H. Nagel, R. M. Nishikawa, J. Papaioannou, and K. Doi, "Analysis of methods for reducing false positives in the automated detection of clustered microcalcifications in mammograms," *Med. Phys.* **25**, 1502–1506 (1998).

<sup>17</sup>M. A. Gavrielides, J. Y. Lo, R. Vargas-Voracek, and C. E. Floyd, "Segmentation of suspicious clustered microcalcifications in mammograms," *Med. Phys.* **27**, 13–22 (2000).

<sup>18</sup>M. Kallergi, L. P. Clarke, W. Qian, M. Gavrielides, P. Venugopal, C. G. Berman, S. D. Holman-Ferris, M. S. Miller, and R. A. Clark, "Interpretation of calcifications in screen/film, digitized, and wavelet-enhanced monitor-displayed mammograms: a receiver operating characteristic study," *Acad. Radiol.* **3**, 285–293 (1996).

<sup>19</sup>U. Bick, M. L. Giger, R. A. Schmidt, R. M. Nishikawa, D. E. Wolverton, and K. Doi, "Automated segmentation of digitized mammograms," *Acad. Radiol.* **2**, 1–9 (1995).

<sup>20</sup>D. B. Kopans, "Discriminating analysis uncovers breast lesions," *Diagn. Imaging* **13**, 94–101 (1991).

<sup>21</sup>A. K. Jain and M.-P. Dubuisson, "Segmentation of x-ray and c-scan images of fiber reinforced composite materials," *Pattern Recogn.* **25**, 257–270 (1992).

<sup>22</sup>W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge University Press, Cambridge, UK, 1988).

<sup>23</sup>C. E. Metz, "ROC methodology in radiologic imaging," *Invest. Radiol.* **21**, 720–733 (1986).

<sup>24</sup>Y. Jiang, C. Metz, and R. Nishikawa, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests," *Radiology* **201**, 745–750 (1996).

<sup>25</sup>S. Sharma, *Applied Multivariate Techniques* (J. Wiley, New York, 1996).

<sup>26</sup>I. M. Freundlich, T. B. Hunter, G. W. Seeley, C. J. D'Orsi, and N. L. Sadowsky, "Computer-assisted analysis of mammographic clustered calcifications," *Clin. Radiol.* **40**, 295–298 (1989).

<sup>27</sup>D. P. Chakraborty and L. H. L. Winter, "Free-response methodology: Alternate analysis and a new observer-performance experiment," *Radiology* **174**, 873–881 (1990).