Computerized analysis of mammographic microcalcifications in morphological and texture feature spaces

Heang-Ping Chan[a] and Berkman Sahiner
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109

Kwok Leung Lam
Department of Radiation Oncology, University of Michigan, Ann Arbor, Michigan 48109

Nicholas Petrick, Mark A. Helvie, Mitchell M. Goodsitt, and Dorit D. Adler
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109

(Received 24 September 1997; accepted for publication 20 July 1998)

We are developing computerized feature extraction and classification methods to analyze malignant and benign microcalcifications on digitized mammograms. Morphological features that described the size, contrast, and shape of microcalcifications and their variations within a cluster were designed to characterize microcalcifications segmented from the mammographic background. Texture features were derived from the spatial gray-level dependence (SGLD) matrices constructed at multiple distances and directions from tissue regions containing microcalcifications. A genetic algorithm (GA) based feature selection technique was used to select the best feature subset from the multi-dimensional feature spaces. The GA-based method was compared to the commonly used feature selection method based on the stepwise linear discriminant analysis (LDA) procedure. Linear discriminant classifiers using the selected features as input predictor variables were formulated for the classification task. The discriminant scores output from the classifiers were analyzed by receiver operating characteristic (ROC) methodology and the classification accuracy was quantified by the area, A_z, under the ROC curve. We analyzed a data set of 145 mammographic microcalcification clusters in this study. It was found that the feature subsets selected by the GA-based method are comparable to or slightly better than those selected by the stepwise LDA method. The texture features ($A_z = 0.84$) were more effective than morphological features ($A_z = 0.79$) in distinguishing malignant and benign microcalcifications. The highest classification accuracy ($A_z = 0.89$) was obtained in the combined texture and morphological feature space. The improvement was statistically significant in comparison to classification in either the morphological ($p = 0.002$) or the texture ($p = 0.04$) feature space alone. The classifier using the best feature subset from the combined feature space and an appropriate decision threshold could correctly identify 35% of the benign clusters without missing a malignant cluster. When the average discriminant score from all views of the same cluster was used for classification, the A_z value increased to 0.93 and the classifier could identify 50% of the benign clusters at 100% sensitivity for malignancy. Alternatively, if the minimum discriminant score from all views of the same cluster was used, the A_z value would be 0.90 and a specificity of 32% would be obtained at 100% sensitivity. The results of this study indicate the potential of using combined morphological and texture features for computer-aided classification of microcalcifications. [S0094-2405(98)00910-9]

Key words: computer-aided diagnosis, mammography, microcalcifications, genetic algorithm, linear discriminant analysis, ROC analysis

I. INTRODUCTION

Mammography is the most sensitive method for early detection of breast cancers. However, its specificity for differentiating malignant and benign lesions is relatively low. In the United States, the positive predictive value of mammography ranges from about 15% to 30%.[1,2] Various methods are being developed to improve the sensitivity and specificity of breast cancer detection.[3] Computer-aided diagnosis (CAD) is considered to be one of the promising approaches that may improve the efficacy of mammography.[4] Properly designed CAD algorithms can automatically detect suspicious lesions on a mammogram and alert the radiologist to these regions. They can also extract image features from regions of interest (ROIs) and estimate the likelihood of malignancy for a given lesion, thereby providing the radiologist with additional information for making diagnostic decisions.

There are two major approaches to the development of CAD schemes for classification of mammographic abnormalities. One approach uses computer vision techniques to extract image features from the digitized mammograms and classify the lesions based on the computer-extracted features. The computer-extracted features can include morphological features that are commonly used by radiologists for diagno-

sis, as well as texture features that may not be readily perceived by human eyes. The computerized analysis may therefore increase the utilization of mammographic image information and improve the accuracy of differentiating malignant and benign lesions. The other approach uses radiologists' ratings of mammographic features or encodes the radiologists' readings with numerical values. The lesions are then classified based on these radiologist-extracted features. This approach assists radiologists by systematically extracting image features and by optimally merging the features with a statistical classifier to reach a diagnostic decision. Additional risk factors based on patient demographic information and medical or family histories may also be included as input in either approach.

A number of investigators have developed feature extraction and classification methods for characterization of mammographic masses or microcalcifications. Ackerman *et al.*[5] developed 4 measures of malignancy and classified lesions recorded on 120 digitized xeroradiographs by 3 decision methods. Kilday *et al.*[6] used 7 shape descriptors and patient age to classify 39 masses and could correctly classify 69% of the masses. Huo *et al.*[7] analyzed the spiculation of masses using a radial edge-gradient analysis technique and achieved an area, A_z, under the receiver operating characteristic (ROC) curve of 0.88 in a data set of 95 masses. Sahiner *et al.*[8,9] developed a rubber-band straightening image transformation technique to analyze the texture in the region surrounding a mass and obtained an A_z of 0.94 in a data set of 168 masses. Pohlman *et al.*[10] extracted 6 morphological descriptors to classify 47 masses and obtained A_z values ranging from 0.76 to 0.93. Wee *et al.*[11] analyzed 51 microcalcification clusters on specimen radiographs using the average gray level, contrast, and horizontal length of the microcalcifications and obtained 84% correct classification. Fox *et al.*[12] included cluster features in their classifier and obtained 67% correct classification in a data set of 100 clusters from specimen radiographs. Chan *et al.*[13–18] developed morphological and texture features and evaluated various feature classifiers for differentiation of malignant and benign microcalcifications. Shen *et al.*[19] used 3 shape features, compactness, moments, and Fourier descriptors to classify 143 individual microcalcifications with a nearest neighbor classifier and obtained 100% classification accuracy. Wu *et al.*[20] classified 80 pathologic specimens radiographs with a convolution neural network and obtained an A_z of 0.90. Jiang *et al.*[21] trained a neural network classifier to analyze 8 features extracted from microcalcification clusters and obtained an A_z of 0.92 in a data set of 53 patients. Thiele *et al.*[22] extracted texture and fractal features from the tissue region surrounding a microcalcification cluster for classification and achieved a sensitivity of 89% at a specificity of 83% for 54 clusters. Dhawan *et al.*[23] used features derived from first-order and second-order gray-level histogram statistics and obtained an A_z of 0.81 with a neural network classifier for a data set of 191 clusters.

Computerized classification of mammographic lesions using radiologist-extracted features has also been reported by a number of investigators. Ackerman *et al.*[24] estimated the probability of malignancy of mammographic lesions by analyzing 36 radiologist-extracted characteristics with an automatic clustering algorithm and obtained a specificity of 45% at a sensitivity of 100% in a data set of 102 cases. Gale *et al.*[25] analyzed 12 radiologist-extracted features of mammographic lesions with a computer algorithm and obtained a specificity of 88% at a sensitivity of 79% in a data base of 500 patients. Getty *et al.*[26] developed a computer classifier to enhance the differentiation of malignant and benign lesions by a radiologist during interpretation of xeromammograms. Using a similar approach, D'Orsi *et al.*[27] evaluated a computer aid and obtained an improvement of about 0.05 in sensitivity or specificity in mammographic reading. Wu *et al.*[28] trained a neural network to merge 14 radiologist-extracted features for classification of mammographic lesions and obtained an A_z of 0.89. Baker *et al.*[29] trained a neural network based on the lexicon of the Breast Imaging Recording and Data System of the American College of Radiology and found that the neural network could improve the positive predictive value from 35% to 61% in 206 lesions. Lo *et al.*[30] used a similar approach to predict breast cancer invasion and obtained an A_z of 0.91 for 96 lesions. Although the results of these studies varied over a wide range and the performances of the computer algorithms are expected to depend strongly on data set, they indicate the potential of using CAD techniques to improve the diagnostic accuracy of differentiating malignant and benign lesions.

In our early studies, we found that texture features extracted from spatial gray-level dependence (SGLD) matrices at multiple distances were useful for differentiating malignant and benign masses on mammograms. This may be attributed to the texture changes in the breast tissue due to a developing malignancy. The usefulness of SGLD texture measures in differentiating malignant and benign breast tissues was further demonstrated by analysis of mammographic microcalcifications.[17,18,31] In a preliminary study, we developed morphological features to describe the size, shape, and contrast of the individual microcalcifications and their variation within a cluster. We used these features to classify the microcalcifications and obtained moderate results.[13,15] In the present study, we expanded the data set and explored the feasibility of combining texture and morphological features for classification of microcalcifications. The classification accuracy in the combined feature space was compared with those obtained in the texture feature space or in the morphological feature space alone. We also studied the use of a genetic algorithm[32–34] (GA) to select a feature subset from the large-dimension feature spaces, and compared the classification results to those obtained from features selected with stepwise linear discriminant analysis (LDA).[35] Linear discriminant classifiers[36] were designed for the classification tasks. The performance of the classifiers was analyzed with ROC methodology[37] and the classification accuracy was quantified with the area, A_z, under the ROC curve.

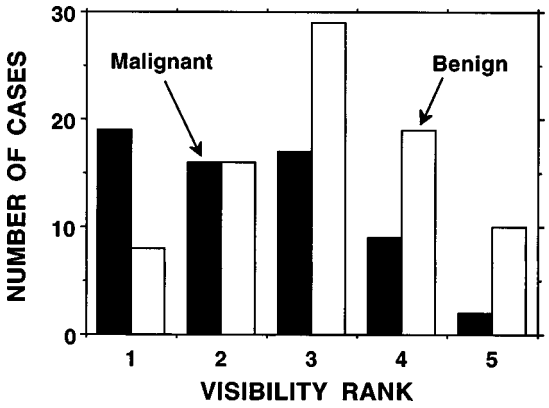

FIG. 1. Distribution of the visibility rankings of the 145 clusters of microcalcifications. Higher ranking corresponds to more subtle clusters.

II. MATERIALS AND METHODS

A. Data set

The data set for this study consisted of 145 clusters of microcalcifications from mammograms of 78 patients. The cases were selected from the patient files in the Department of Radiology at the University of Michigan. The only selection criterion was that it included a biopsy-proven microcalcification cluster. We kept the number of malignant and benign cases reasonably balanced so that 82 benign and 63 malignant clusters were included. All mammograms were acquired with a contact technique using mammography systems accredited by the American College of Radiology (ACR). The dedicated mammographic systems had molybdenum anode and molybdenum filter, 0.3 mm nominal focal spot, reciprocating grid, and Kodak MinR/MinR E screen-film systems with extended processing. A radiologist experienced in mammography ranked the visibility of each microcalcification cluster on a scale of 1 (obvious) to 5 (subtle), relative to the visibility range of microcalcification clusters encountered in clinical practice. The histogram of the visibility ranking of the 145 clusters is shown in Fig. 1. The histogram indicated the mix of subtle and obvious clusters included in the data set.

The selected mammograms were digitized with a laser scanner (Lumisys DIS-1000) at a pixel size of 0.035 mm $\times$0.035 mm and 12-bit gray levels. The digitizer has an optical density (O.D.) range of about 0 to 3.5. The O.D. on the film was digitized linearly to pixel value at a calibration of 0.001 O.D. unit/pixel value in the O.D. range of about 0 to 2.8. The digitizer deviated from a linear response at O.D. higher than 2.8.

B. Morphological feature space

For the extraction of morphological features, the locations of the individual microcalcifications have to be known. We have developed an automated program for detection of individual microcalcifications.[38] However, the detection sensitivity is not 100% and the detected signals include false-positives. Furthermore, automated detection tends to have a higher likelihood of detecting obvious microcalcifications than subtle ones, which may bias the evaluation of the classification capability of the extracted features and the trained classifiers if microcalcifications detected by the automated program are used for classifier development. Since these variables are program dependent, we isolated the detection problem from the classification problem in this study by using manually identified true microcalcifications for the morphological feature analysis. The true microcalcifications were defined as those visible on the film mammograms with a magnifier. Magnification mammograms were used occasionally for verification when they were available, but in most cases only contact mammograms were used. At present, there is no other method that can more reliably identify individual microcalcifications on mammograms. Specimen radiographs can confirm the presence of the microcalcifications but the locations of the individual microcalcifications cannot be correlated with those on the mammograms because of the very different imaging geometry and techniques.

We have developed an automated signal extraction program to determine the size, contrast, signal-to-noise ratio (SNR), and shape of the microcalcifications from a mammogram based on the coordinate of each individual microcalcification. In a local region of 101×101 pixels centered at each signal site, the low frequency structured background is estimated by polynomial curve fitting in the horizontal and vertical directions and then averaging the fitted values obtained in the two directions at each pixel. This background estimation method is used because it can approximate the background more closely than two-dimensional surface fitting or the distance-weighted interpolation method (described below) used for texture feature extraction. The central $l\times l$ pixels that contain the signal are excluded from the curve fitting and noise estimation. The size l is chosen to be a constant of 15 pixels which is larger than the diameters of the microcalcifications of interest yet much smaller than the local region. The background pixel values in this $l\times l$ region are estimated from the fitted and smoothed background surface. The exclusion of the signal region is necessary so that the high contrast pixel values of the microcalcification will not affect the background estimation at the signal site. Other microcalcifications that may locate within the 101×101 pixel region are treated as background pixels because their effect on the estimated background levels at the signal site will be relatively small.

After subtraction of the structured background, the local root-mean-square (rms) noise is calculated. A gray-level threshold is determined as the product of the rms noise and an input SNR threshold. With a region growing technique, the signal region is then extracted as the connected pixels above the threshold around the manually identified signal location. A high threshold will result in extracting only the peak pixels of the microcalcification which may not represent its shape perceived on the mammogram. A low threshold will cause the microcalcification region to grow into the surrounding background pixels. Since there is no objective standard what the actual shape of a microcalcification is on a mammogram, the proper threshold to extract the signals was

(a)

(b)

FIG. 2. An example of a cluster of malignant microcalcifications in the data set: (a) the cluster with mammographic background, (b) the cluster after segmentation. Morphological features are extracted from the segmented microcalcifications.

determined by visually comparing the microcalcifications in the original image and the thresholded image of the microcalcifications superimposed on a background of constant pixel values. After an experienced radiologist compared a subset of randomly selected microcalcification clusters extracted at different thresholds, an SNR threshold of 2.0 was chosen for all cases. An example of a malignant cluster and the microcalcifications extracted at an SNR threshold of 2.0 is shown in Fig. 2.

The feature descriptors determined from the extracted microcalcifications are listed in Table I. The size of a microcalcification (SA) is estimated as the number of pixels in the

TABLE I. The 21 morphological features extracted from a microcalcification cluster.

	Average	Standard deviation	Coefficient of variation	Maximum
Area	AVSA	SDSA	CVSA	MXSA
Mean density	AVMD	SDMD	CVMD	MXMD
Eccentricity	AVEC	SDEC	CVEC	MXEC
Moment ratio	AVMR	SDMR	CVMR	MXMR
Axis ratio	AVAR	SDAR	CVAR	MXAR
No. of microcalcifications in cluster	NUMS			

signal region. The mean density (MD) is the average of the pixel values above the background level within the signal region. The second moments are calculated as

$$M_{xx}=\sum_i g_i(x_i-M_x)^2/M_0, \tag{1}$$

$$M_{yy}=\sum_i g_i(y_i-M_y)^2/M_0, \tag{2}$$

$$M_{xy}=\sum_i g_i(x_i-M_x)(y_i-M_y)/M_0, \tag{3}$$

where g_i is the pixel value above the background, and (x_i,y_i) are the coordinates of the ith pixel. The moments M_0, M_x and M_y are defined as follows:

$$M_0=\sum_i g_i, \tag{4}$$

$$M_x=\sum_i g_i x_i/M_0, \tag{5}$$

$$M_y=\sum_i g_i y_i/M_0. \tag{6}$$

The summations are over all pixels within the signal region. The lengths of the major axis, $2a$, and the minor axis, $2b$, of the effective ellipse that characterizes the second moments are given by

$$2a=\sqrt{2[M_{xx}+M_{yy}+\sqrt{(M_{xx}-M_{yy})^2+4M_{xy}^2}]}, \tag{7}$$

$$2b=\sqrt{2[M_{xx}+M_{yy}-\sqrt{(M_{xx}-M_{yy})^2+4M_{xy}^2}]}. \tag{8}$$

The eccentricity (EC) of the effective ellipse can be derived from the major and minor axes as

$$\epsilon=\frac{\sqrt{a^2-b^2}}{a}. \tag{9}$$

The moment ratio (MR) is defined as the ratio of M_{xx} to M_{yy}, with the larger second moment in the denominator. The axis ratio (AR) is the ratio of the major axis to the minor axis of the effective eclipse.

To quantify the variation of the visibility and shape descriptors in a cluster, the maximum (MX), the average (AV) and the standard deviation (SD) of each feature for the individual microcalcifications in the cluster are calculated. The coefficient of variation (CV), which is the ratio of the SD to AV, is used as a descriptor of the variability of a certain

feature within a cluster. Twenty cluster features are therefore derived from the five features (size, mean density, moment ratio, axis ratio, and eccentricity) of the individual microcalcifications. Another feature describing the number of microcalcifications in a cluster (NUMS) is also added, resulting in a 21-dimensional morphological feature space.

C. Texture feature space

Our texture feature extraction method has been described in detail previously.[31] Briefly, texture features are extracted from a 1024×1024 pixel region of interest (ROI) that contains the cluster of microcalcifications. Most of the clusters in this data set can be contained within the ROI. For the few clusters that are substantially larger than a single ROI, additional ROIs containing the remaining parts of the cluster are extracted and processed in the same way as the other ROIs. The texture feature values extracted from the different ROIs of the same cluster are averaged and the average values are used as the feature values for that cluster.

For a given ROI, background correction is first performed to reduce the low frequency gray-level variation due to the density of the overlapping breast tissue and the x-ray exposure conditions. The gray level at a given pixel of the low frequency background is estimated as the average of the distance-weighted gray levels of four pixels at the intersections of the normals from the given pixel to the four edges of the ROI.[39] The estimated background image was subtracted from the original ROI to obtain a background-corrected image. An example of the background correction procedure is shown in Fig. 3.

As discussed in our previous study,[31] it was found that the texture features derived from the SGLD matrix of the ROI provided useful texture information for classification of microcalcification clusters. The SGLD matrix element, $p_\theta,_d(i,j)$, is the joint probability of the occurrence of gray levels i and j for pixel pairs which are separated by a distance d and at a direction θ.[40] The SGLD matrices were constructed from the pixel pairs in a subregion of 512×512 pixels centered approximately at the center of the cluster in the background-corrected ROI so that any potential edge effects caused by background correction will not affect the texture extraction. We analyzed the texture features in four directions: $\theta=0°$, 45°, 90°, and 135° at each pixel pair distance d. The pixel pair distance was varied from 4 to 40 pixels in increments of 4 pixels. Therefore, a total of 40 SGLD matrices were derived from each ROI. The SGLD matrix depends on the bin width (or gray-level interval) used in accumulating the histogram. Based on our previous study, a bin width of four gray levels was chosen for constructing the SGLD matrices. This is equivalent to reducing the gray-level resolution (or bit depth) of the 12-bit image to 10 bits by eliminating the 2 least significant bits.

From each of the SGLD matrices, we derived 13 texture measures including correlation, entropy, energy (angular second moment), inertia, inverse difference moment, sum average, sum entropy, sum variance, difference average, difference entropy, difference variance, information measure of correlation 1, and information measure of correlation 2. The formulation of these texture measures could be found in the literature.[31,40] As found in our previous study,[41] we did not observe a significant dependence of the discriminatory power of the texture features on the direction of the pixel pairs for mammographic textures. However, since the actual distance between the pixel pairs in the diagonal direction was a factor

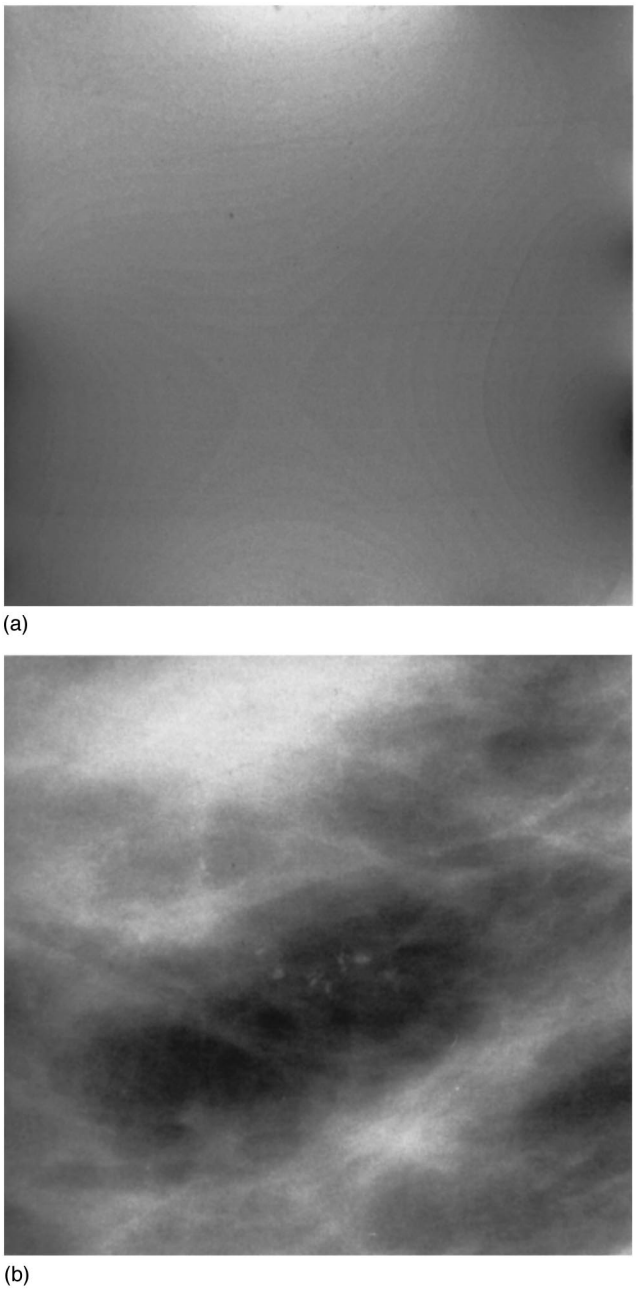

FIG. 3. An example of background correction for the ROIs before texture feature extraction. The ROI from the original image is shown in Fig. 2(a). (a) The estimated low frequency background gray level, and (b) the ROI after background correction. The background gray-level variation due to the varying x-ray penetration in the breast tissue is reduced. The contouring in the background image is a display artifact that does not exist in the calculated image file. For display purpose, the background-corrected ROI is contrast-enhanced to improve the visibility of the microcalcifications and the detailed structures.

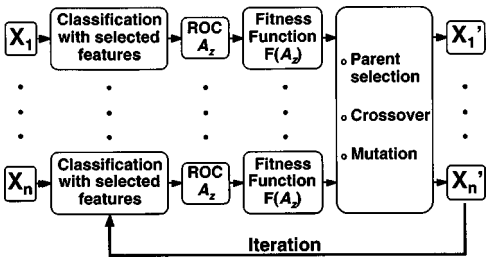

FIG. 4. A schematic diagram of the genetic algorithm designed for feature selection used in this study. $X_1,\ldots,X_n$ represents the set of parent chromosomes and $X_1',\ldots,X_n'$ represents the set of offspring chromosomes.

of $\sqrt{2}$ greater than that in the axial direction, we averaged the feature values in the axial directions (0° and 90°) and in the diagonal directions (45° and 135°) separately for each texture feature derived from the SGLD matrix at a given pixel pair distance. The average texture features at the ten pixel pair distances and two directions formed a 260-dimensional texture feature space.

D. Feature selection

Feature selection is one of the most important steps in classifier design because the presence of ineffective features often degrades the performance of a classifier on test samples. This is partly caused by the ''curse of dimensionality'' problem that the classifier is inadequately trained in a large-dimension feature space when only a finite number of training samples is available.[42–45] We compared two feature selection methods to extract useful features from the morphological, texture, and the combined feature spaces. One is a genetic algorithm approach, and the other is the commonly used stepwise linear discriminant analysis method.

1. Genetic algorithm for feature selection

The genetic algorithm (GA) methodology was first introduced by Holland in the early 1970s.[32,33] A GA solves an optimization problem based on the principles of natural selection. In natural selection, a population evolves by finding beneficial adaptations to a complex environment. The characteristics of a population are carried onto the next generation by its chromosomes. New characteristics are introduced into a chromosome by crossover and mutation. The probability of survival or reproduction of an individual depends more or less on its fitness to the environment. The population therefore evolves toward better-fit individuals.

The application of GA to feature selection has been described in the literature.[46,47] We have demonstrated previously that a GA could select effective features for classification of masses and normal breast tissue from a very large-dimension feature space.[34] The GA was adapted to the current problem for classification of malignant and benign microcalcifications. A brief outline is given as follows. Each feature in a given feature space is treated as a gene and is encoded by a binary digit (bit) in a chromosome. A ''1'' represents the presence of the feature and a ''0'' represents the absence of the feature. The number of genes (bits) on a chromosome is equal to the dimensionality (k) of the feature space, but only the features that are encoded as ''1'' are actually present in the subset of selected features. A chromosome therefore represents a possible solution to the feature selection problem.

The implementation of GA for feature selection is illustrated in the block diagram shown in Fig. 4. To allow for diversity, a large number, n, of chromosomes, $X_1,\ldots,X_n$, is chosen as the population. The number of chromosomes is kept constant in each generation. At the initiation of the GA, each bit on a chromosome is initialized randomly with a small but equal probability, P_{init}, to be ''1.'' The selected feature subset on a chromosome is used as the input feature variables to a classifier, which was chosen to be the Fischer's linear discriminant in this study.

The available samples in the dataset are randomly partitioned into a training set and a test set. The training set is used to formulate a linear discriminant function with each of the selected feature subsets. The effectiveness of each of the linear discriminants for classification is evaluated with the test set. The classification accuracy is determined as the area, A_z, under the ROC curve. To reduce biases in the classifiers due to case selection, training and testing are performed a large number of times, each with a different random partitioning of the data set. In this study, we chose to partition the dataset 80 times and the 80 test A_z values were averaged and used for determination of the fitness of the chromosome.

The fitness function for the ith chromosome, $F(i)$, is formulated as

$$F(i)=\left[\frac{f(i)-f_{\min}}{f_{\max}-f_{\min}}\right]^2, \quad i=1,...,n, \tag{10}$$

where

$$f(i)=\overline{A_z(i)}-\alpha N(i),$$

$\overline{A_z(i)}$ is the average test A_z for the ith chromosome over the 80 random partitions of the data set, $f_{\min}$ and $f_{\max}$ are the minimum and maximum $f(i)$ among the n chromosomes, $N(i)$ is the number of features in the ith chromosome, and α is a penalty factor, whose magnitude is less than $1/k$, to suppress chromosomes with a large number of selected features. The value of the fitness function $F(i)$ ranges from 0 to 1. The probability of the ith chromosome being selected as a parent, $P_s(i)$, is proportional to its fitness function:

$$P_s(i)=F(i)/\sum_{i=1}^{n} F(i), \quad i=1,...,n. \tag{11}$$

A random sampling based on the probabilities, $P_s(i)$, will allow chromosomes with higher value of fitness to be selected more frequently.

For every pair of selected parent chromosomes, X_i and X_j, a random decision is made to determine if crossover should take place. A uniform random number in (0,1] is generated. If the random number is greater than P_c, the probability of crossover, then no crossover will occur; otherwise, a random crossover site is selected on the pair of chromosomes. Each chromosome is split into two strings at this site and one of the strings will be exchanged with the corre-

sponding string from the other chromosome. Crossover results in two new chromosomes of the same length.

After crossover, another chance of introducing new features is obtained by mutation. Mutation is applied to each gene on every chromosome. For each bit, a uniform random number in (0,1] is generated. If the random number is greater than P_m, the probability of mutation, then no mutation will occur; otherwise, the bit is complemented. The processes of parent selection, crossover, and mutation result in a new generation of n chromosomes, $X_1',\ldots,X_n'$, which will again be evaluated with the 80 training and test set partitions as described above. The chromosomes are allowed to evolve over a preselected number of generations. The best subset of features is chosen to be the chromosome that provides the highest average A_z during the evolution process.

In this study, 500 chromosomes were used in the population. Each chromosome has 281 gene locations. P_{init} was chosen to be 0.01 so that each chromosome started with two to three features on the average. We varied P_c from 0.7 to 0.9, P_m from 0.001 to 0.005, and α from 0 to 0.001. These ranges of parameters were chosen based on our previous experience with other feature selection problems using GA.[34]

2. Stepwise linear discriminant analysis

The stepwise linear discriminant analysis (LDA) is a commonly used method for selection of useful feature variables from a large feature space. Detailed descriptions of this method can be found in the literature.[35] The procedure is briefly outlined below. The stepwise LDA uses a forward selection and backward removal strategy. When a feature is entered into or removed from the model, its effect on the separation of the two classes can be analyzed by several criteria. We use the Wilks' lambda criterion which minimizes the ratio of the within-group sum of squares to the total sum of squares of the two class distributions; the significance of the change in the Wilks' lambda is estimated by F-statistics. In the forward selection step, the features are entered one at a time. The feature variable that causes the most significant change in the Wilks' lambda will be included in the feature set if its F value is greater than the F-to-enter (F_{in}) threshold. In the feature removal step, the features already in the model are eliminated one at a time. The feature variable that causes the least significant change in the Wilks' lambda will be excluded from the feature set if its F value is below the F-to-remove (F_{out}) threshold. The stepwise procedure terminates when the F values for all features not in the model are smaller than the F_{in} threshold and the F values for all features in the model are greater than the F_{out} threshold. The number of selected features will decrease if either the F_{in} threshold or the F_{out} threshold is increased. Therefore, the number of features to be selected can be adjusted by varying the F_{in} and F_{out} values.

E. Classifier

The training and testing procedure described above was used for the purpose of feature selection only. After the best subset of features as determined by either the GA or the stepwise LDA procedure was found, we performed the classification as follows.

The linear discriminant analysis[36] procedure in the SPSS software package[35] was used to classify the malignant and benign microcalcification clusters. We used a cross-validation resampling scheme for training and testing the classifier. The data set of 145 samples was randomly partitioned into a training set and a test set by an approximately 3:1 ratio. The partitioning was constrained so that ROIs from the same patient were always grouped into the same set. The training set was used to determine the coefficients (or weights) of the feature variables in the linear discriminant function. The performance of the trained classifier was evaluated with the test set. In order to reduce the effect of case selection, the random partitioning was performed 50 times. The results were then averaged over the 50 partitions.

The classification accuracy of the LDA was evaluated by ROC methodology. The output discriminant score from the LDA classifier was used as the decision variable in the ROC analysis. The LABROC program,[37] which assumes binormal distributions of the decision variable for the two classes and fits an ROC curve to the classifier output based on maximum-likelihood estimation, was used to estimate the ROC curve of the classifier. The ROC curve represents the relationship between the true-positive fraction (TPF) and the false-positive fraction (FPF) as the decision threshold varies. The area under the ROC curve and the standard deviation of the A_z were provided by the LABROC program for each partition of training and test sets. The average performance of the classifier was estimated as the average of the 50 test A_z values from the 50 random partitions.

To obtain a single distribution of the discriminant scores for the test samples, we performed a leave-one-case-out resampling scheme for training and testing the classifier. In this scheme, one of the 78 cases was left out at a time and the clusters from the other 77 cases were used for formulation of the linear discriminant function. The resulting LDA classifier was used to classify the clusters from the left-out case. The procedure was performed 78 times so that every case was left out once to be the test case. The test discriminant scores from all the clusters were accumulated in a distribution which was then analyzed by the LABROC program. Using the distributions of discriminant scores for the test samples from the leave-one-case-out resampling scheme, the CLABROC program could be used to test the statistical significance of the differences between ROC curves[48] obtained from different conditions. The two-tailed p value for the difference in the areas under the ROC curves was estimated.

III. RESULTS

The variations of best feature set size and classifier performance in terms of A_z with the GA parameters were tabulated in Table II(a)–(c) for the morphological, the texture, and the combined feature spaces, respectively. The number of generations that the chromosomes evolved was fixed at 75

TABLE II. Dependence of feature selection and classifier performance on GA parameters: (a) morphological feature space, (b) texture feature space, and (c) combined feature space. The number of generations that the GA evolved was fixed at 75. The best result for each feature space is identified with an asterisk.

(a)

P_c	P_m	α	No. of features	A_z (Training)	A_z (Test)
0.7	0.001	0	6	0.84	0.79
0.8			3	0.77	0.76
0.9			4	0.80	0.77
0.7	0.003		7	0.82	0.78
0.8			6	0.82	0.79
0.9			6	0.84	0.79
0.7	0.001	0.0005	3	0.77	0.76
0.8			4	0.80	0.77
0.9			3	0.77	0.76
0.7	0.003		6	0.84	0.79*
0.8			6	0.84	0.79
0.9			6	0.82	0.79
0.7	0.001	0.0010	3	0.77	0.76
0.8			4	0.80	0.77
0.9			3	0.77	0.76
0.7	0.003		6	0.84	0.79
0.8			7	0.84	0.79
0.9			4	0.80	0.77

(b)

P_c	P_m	α	No. of features	A_z (Training)	A_z (Test)
0.7	0.001	0	7	0.87	0.82
0.8			8	0.88	0.84
0.9			8	0.88	0.84
0.7	0.003		17	0.91	0.82
0.8			9	0.88	0.79
0.9			10	0.88	0.79
0.7	0.001	0.0005	9	0.88	0.85*
0.8			7	0.86	0.82
0.9			8	0.87	0.84
0.7	0.003		13	0.90	0.81
0.8			10	0.87	0.81
0.9			12	0.88	0.81
0.7	0.001	0.0010	7	0.87	0.83
0.8			9	0.88	0.83
0.9			8	0.88	0.83
0.7	0.003		10	0.88	0.83
0.8			21	0.94	0.82
0.9			12	0.88	0.80

(c)

P_c	P_m	α	No. of features	A_z (Training)	A_z (Test)
0.7	0.001	0	13	0.93	0.88
0.8			12	0.92	0.88
0.9			12	0.92	0.89
0.7	0.003		12	0.91	0.86
0.8			16	0.94	0.88
0.9			17	0.95	0.88
0.7	0.001	0.0003	12	0.92	0.87
0.8			12	0.92	0.86
0.9			12	0.93	0.88
0.7	0.003		13	0.93	0.87
0.8			13	0.93	0.88
0.9			12	0.94	0.89*
0.7	0.005		12	0.89	0.80
0.7	0.001	0.0010	11	0.92	0.87
0.8			10	0.91	0.87
0.9			11	0.91	0.86
0.7	0.003		10	0.91	0.86
0.8			14	0.93	0.87
0.9			13	0.92	0.87
0.7	0.005		11	0.89	0.81
0.8			12	0.88	0.82
0.9			12	0.89	0.81

TABLE III. Dependence of feature selection and classifier performance on F_{out} and F_{in} thresholds using stepwise linear discriminant analysis: (a) morphological feature space, (b) texture feature space, and (c) combined feature space. The best result for each feature space is identified with an asterisk. When the test A_z is comparable, the feature set with fewer number of features is considered to be better.

(a)

F_{out}	F_{in}	No. of features	A_z (Training)	A_z (Test)
2.7	3.8	2	0.76	0.76
1.7	2.8	4	0.79	0.76
1.7	1.8	6	0.83	0.79*
1.0	1.4			
1.0	1.2	7	0.84	0.79
0.8	1.0	9	0.85	0.79
0.6	0.8			
0.4	0.6	10	0.85	0.79
0.2	0.4	12	0.86	0.78
0.1	0.2			

(b)

F_{out}	F_{in}	No. of features	A_z (Training)	A_z (Test)
2.7	3.8	4	0.82	0.80
1.7	2.8			
1.0	1.4	8	0.88	0.83
1.0	1.2	10	0.89	0.82
0.8	1.0	11	0.89	0.83
0.6	0.8	14	0.91	0.85*
0.4	0.6	17	0.92	0.84
0.2	0.4	18	0.92	0.81
0.1	0.2	16	0.90	0.80

(c)

F_{out}	F_{in}	No. of features	A_z (Training)	A_z (Test)
3.0	3.2	6	0.84	0.80
2.9	3.2			
2.8	3.1			
2.0	3.1			
3.0	3.1	10	0.88	0.83
2.9	3.0			
2.7	2.8			
2.0	2.3	11	0.90	0.86
2.0	2.2			
1.9	2.0			
1.7	1.8			
1.3	1.5	14	0.92	0.86
1.0	1.2	19	0.95	0.86
1.0	1.1	23	0.96	0.87*
0.8	1.2	28	0.97	0.86

in these tables. The training and test A_z values were obtained from averaging results of the 50 partitions of the data sets using the selected feature sets.

The results of feature selection using the stepwise LDA procedure with a range of F_{in} and F_{out} thresholds were tabulated in Table III(a)–(c). The thresholds were varied so that the number of selected features varied over a wide range. Often different choices of F_{in} and F_{out} values could result in the same selected feature set as shown in the tables by the number of features in the set. The average A_z values obtained from the 50 partitions of the data set using the selected feature sets were listed. The best feature sets selected in the different feature spaces are shown in Table IV.

TABLE IV. The best feature sets selected by the GA and stepwise LDA methods (indicated by asterisk in Tables II and III) in the three feature spaces. The number of generations for chromosome evolution in the GA algorithm to reach the selected feature sets is listed. The abbreviations for the texture features are: correlation (CORE), energy (ENER), entropy (ENTR), difference average (DFAV), difference entropy (DFEN), difference variance (DFVR), inertia (INER), inverse difference moment (INVD), information measure of correlation 1 (ICO1), information measure of correlation 2 (ICO2), sum average (SMAV), sum entropy (SMEN), sum variance (SMVR). After an abbreviation, the letter ''A'' indicates diagonal features and the number indicates the pixel distance. The abbreviations for the morphological features can be found in Table I.

GA			Stepwise LDA		
Morphological generation 39	Texture generation 64	Combined generation 169	Morphological	Texture	Combined
CMVD	DFAVA_8	DFAVA_4	AVMD	DFAV_12	CORE_40
CVMR	DFEN_16	DFEN_28	CVMD	DFEN_4	COREA_16
CVSA	DFVRA_24	DFVRA_36	CVMR	DFEN_8	COREA_40
MXMR	DFVR_24	DFVR_12	CVSA	DFENA_12	DFAVA_8
MXSA	DFVR_4	DFVR_20	MXMR	DFENA_24	DFEN_4
SDMD	DFVR_8	ICO1A_20	MXSA	DFVR_24	DFEN_8
	ICO1A_12	ICO1A_32		DFVR_40	DFENA_36
	ICO2A_28	SMEN_16		ICO1_16	DFVR_20
	ICO2_40	SMEN_36		ICO1A_8	ICO1A_28
		AVAR		ICO2_40	ICO2_24
		CVMD		INER_8	ICO2_36
		CVSA		INVD_16	INER_12
		MXEC		INVD_4	INERA_16
		NUMS		INVDA_8	INVDA_36
		SDMD			SMEN_40
					SMENA_4
					AVAR
					CVMD
					CVSA
					MXAR
					MXEC
					NUMS
					SDMD

Table V compares the training and test A_z values from the best feature set in each feature space for the two feature selection methods. The GA parameters that selected the feature set with best classification performance in each feature space after 75 generations (Table II) were used to run the GA again for 500 generations. The A_z values obtained with the best GA selected feature sets after 75 generations are listed together with those obtained after 500 generations. The A_z values obtained with the leave-one-case-out scheme are also shown in Table V. The differences between the corresponding A_z values from the two resampling schemes are within 0.01. The two feature selection methods provided feature sets that had similar test A_z values in the morphological and texture feature spaces. In the combined feature space, there was a slight improvement in the test A_z value obtained with the GA selected features. Although the difference in the A_z

TABLE V. Classification accuracy of linear discriminant classifier in the different feature spaces using feature sets selected by the GA and the stepwise LDA procedure.

	Training A_z			Text A_z		
Feature selection	Morphological	Texture	Combined	Morphological	Texture	Combined
Cross-validation						
GA (75 generations)	0.84±0.04	0.88±0.03	0.94±0.02	0.79±0.07	0.85±0.07	0.89±0.05
GA (500 generations)	0.84±0.04	0.88±0.03	0.96±0.02	0.79±0.07	0.85±0.07	0.90±0.05
Stepwise LDA	0.83±0.04	0.91±0.03	0.96±0.02	0.79±0.07	0.85±0.06	0.87±0.06
Leave-one-case-out						
GA (75 generations)	0.83±0.03	0.88±0.03	0.94±0.02	0.79±0.04	0.84±0.03	0.89±0.03
GA (500 generations)	0.83±0.03	0.88±0.03	0.95±0.02	0.79±0.04	0.84±0.03	0.89±0.03
Stepwise LDA	0.83±0.03	0.91±0.02	0.96±0.02	0.79±0.04	0.85±0.03	0.87±0.03

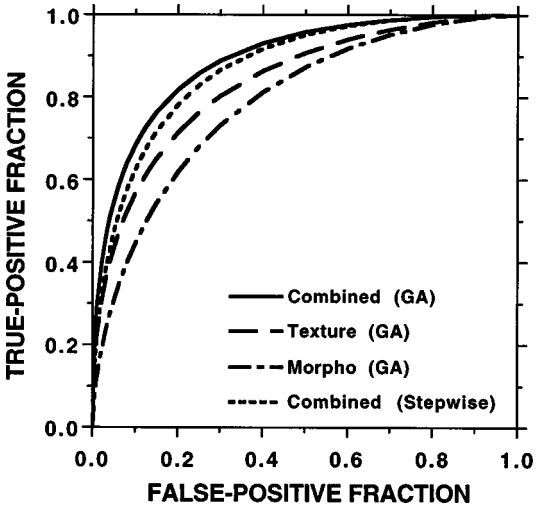

FIG. 5. Comparison of ROC curves of the LDA classifier performance using the best GA selected feature sets in the three feature spaces. In addition, the ROC curve obtained from the best feature set selected by the stepwise LDA procedure in the combined feature space is shown. The classification was performed with a leave-one-case-out resampling scheme.

values from the leave-one-case-out scheme between the two feature selection methods did not achieve statistical significance ($p = 0.2$), as estimated by CLABROC, the differences in the paired A_z values from the 50 partitions demonstrated a consistent trend (40 out of 50 partitions) that the A_z from the GA selected features were higher than those obtained by the stepwise LDA. This trend was also observed in our previous study in which mass and normal tissue were classified.[34]

The ROC curves for the test samples using the feature sets selected by the GA were plotted in Fig. 5. The classification accuracy in the combined feature space was significantly higher than those in the morphological ($p = 0.002$) or the texture feature space ($p = 0.04$) alone. The ROC curve using the feature set selected by the stepwise procedure in the combined feature space was also plotted for comparison. The distribution of the discriminant scores for the test samples using the feature set selected by the GA in the combined feature space is shown in Fig. 6(a). If a decision threshold is chosen at 0.3, 29 of the 82 (35%) benign samples can be correctly classified without missing any malignant clusters.

Some of the 145 samples are different views of the same microcalcification clusters. In clinical practice, the decision regarding a cluster is based on information from all views. If it is desirable to provide the radiologist a single relative malignancy rating for each cluster, two possible strategies may be used to merge the scores from all views: the average score or the minimum score. The latter strategy corresponds to the use of the highest likelihood of malignancy score for the cluster. There were a total of 81 different clusters (44 benign and 37 malignant) from the 78 cases because 3 of the cases contained both a benign and a malignant cluster. The distributions of the average and the minimum discriminant scores of the 81 clusters in the combined feature space were plotted in Fig. 6(b) and Fig. 6(c), respectively. Using the average scores, ROC analysis provided test A_z values of 0.93 ± 0.03 and 0.89 ± 0.04, respectively, for the GA selected and stepwise LDA selected feature sets. Using the minimum scores, the test A_z values were 0.90 ± 0.03 and 0.85 ± 0.04, respectively. The difference between the A_z values from the two feature selection methods did not achieve statistical significance in either case ($p = 0.07$ and $p = 0.09$, respectively). If a decision threshold is chosen at an average score of 0.2, 22 of the 44 (50%) benign clusters can be correctly identified with 100% correct classification of the malignant clusters. If a decision threshold is set at a minimum score of 0.2, 14 of the

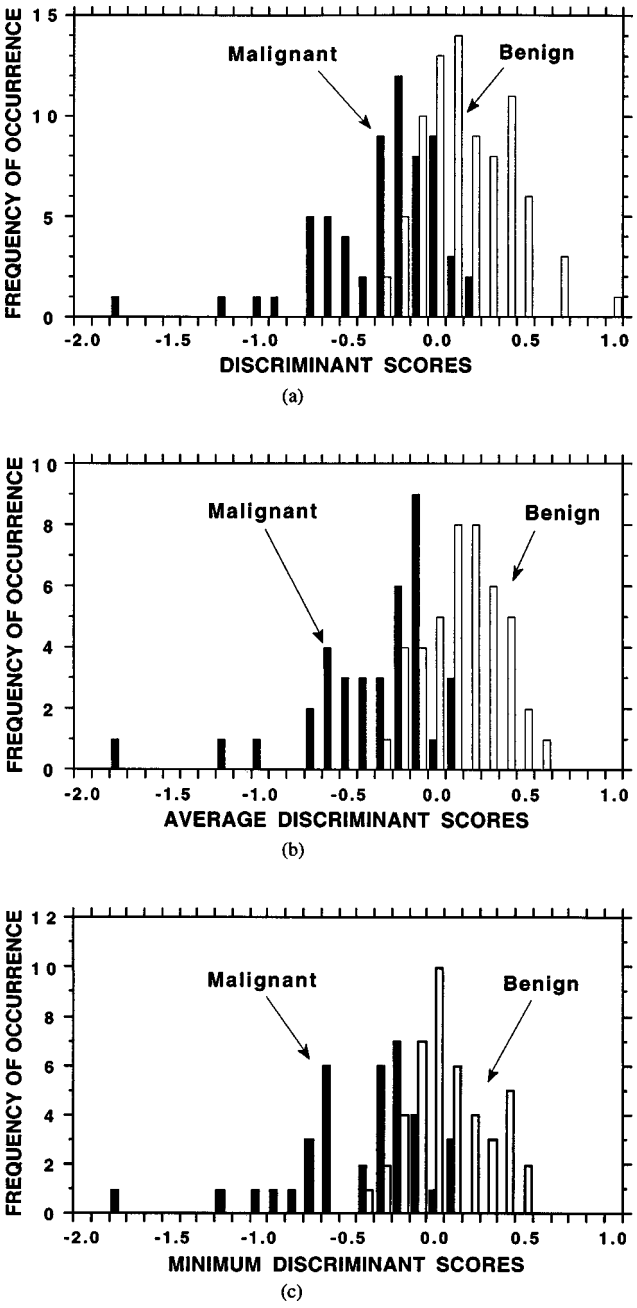

FIG. 6. Distribution of the discriminant scores for the test samples using the best GA selected feature set in the combined texture and morphological feature space. (a) Classification by samples from each film, (b) classification by cluster using the average scores, (c) classification by cluster using the minimum scores.

44 (32%) benign clusters can be identified at 100% sensitivity.

IV. DISCUSSION

The Fischer's linear discriminant is the optimal classifier if the class distributions are multivariate normal with equal covariance matrices.[42] Even if these conditions are not satisfied, as in most classification tasks, the LDA may still be a preferred choice when the number of available training samples is small. Our previous investigation[43,45] of the dependence of classifier performance on design sample size indicated that, in general, the training performance (resubstitution) of a classifier is positively biased whereas the test performance (hold-out) is negatively biased by the sample size. The magnitudes of the biases increase when the dimensionality of the input feature space or the complexity of the classifier increases, or when the design sample size decreases. Therefore, the test performance of a linear classifier is generally better than that of a more complex classifier such as a neural network or a quadratic classifier when the training sample size is small. The training results should not be used for comparison of classifier performance because a classifier can often be overtrained and give a near-perfect classification on training samples while the generalization to any unknown test samples is poor. In this study, we evaluated the effectiveness of using the morphological and the texture features extracted from mammograms for classification of a microcalcification cluster. Although we expanded the data set from our previous study, the current data set was still relatively small. We therefore chose to use a linear discriminant classifier for this classification task. Stepwise feature selection or a GA was used to reduce the dimensionality of the feature space.

In the morphological feature space, the features related to three characteristics, mean density, the moment ratio, and the signal area, were chosen most often. The features related to axis ratio, eccentricity, and the number of microcalcifications in a cluster were chosen only when they were combined with texture features. These results indicate the usefulness of classification in multi-dimensional feature spaces. Some features that are not useful by themselves can become effective features when they are combined with other features. The results also indicate that all six characteristics of the microcalcifications designed for this task have some discriminatory power to distinguish malignant and benign microcalcifications. The morphological features are not as effective as the texture features. This is evident from the smaller A_z values in the morphological feature space. However, when the morphological feature space is combined with the texture feature space, the resulting feature set selected from the combined feature space can significantly improve the classification accuracy, in comparison with those from the individual feature spaces.

The SGLD texture features characterize the shape of the SGLD matrix and generally contain information about the image properties such as homogeneity, contrast, the presence of organized structures, as well as the complexity and gray-level transitions within the image.[40] As an example, the entropy feature measures the uniformity of the SGLD matrix. The entropy value is maximum when all the matrix elements are equal. The entropy value is small when large matrix elements concentrate in a small region of the SGLD matrix while the other matrix elements are relatively small. Therefore, large entropy represents a large but random variation of pixel values in an image without regular structures whereas small entropy represents an image with relatively uniform pixel values if the SGLD matrix peaks along the diagonal and an image with regular texture patterns if it peaks off the diagonal. The ambiguity may be resolved when the sum entropy and difference entropy measures are analyzed. Unlike morphological features, it is difficult, in general, to find the direct relationship between a texture measure and the structures seen on an image,[40] and often a combination of several texture measures extracted at different angles and pixel pair distances are required to describe a texture pattern. It may also be noted that some textures can only be described by second-order statistics and may not be distinguishable by human eyes. The feature selection methods are used to empirically find the combination of features that can most effectively distinguish the malignant and benign lesions.

From Table IV, it can be seen that many of the features in the best feature sets selected by the GA method and the stepwise LDA method are similar. In the morphological feature space, five of the six selected features are the same in the two feature sets. In the combined feature space, six morphological features (out of six and seven morphological features in the two sets, respectively) are the same. For the texture features, there are more variations in the features selected by the two methods. However, the differences are mainly in the pixel distances and the directions of the features, while the major types of the texture features are similar. For example, four types of texture features, energy, entropy, sum average, and sum variance were not selected in either the texture or the combined feature space by both methods. Another four types of texture features, difference average, difference entropy, difference variance, and information measure of correlation 1 were chosen in each case, and information measure of correlation 2 was chosen in three of the four cases. Inertia and inverse difference moment were selected by the stepwise LDA method in both the texture and the combined feature spaces. Sum entropy was selected by both methods in the combined feature space. These results indicate that some features are more effective than the others for distinguishing benign and malignant microcalcifications. The pixel distance and the direction of the texture features may be considered to be higher order effects that have less influence on the discriminatory ability of a given type of texture measure. The smaller differences in their discriminatory ability would subject them to greater variability of being chosen in the feature selection processes. It may also be noted that many of the features are highly correlated. The correlated features can be interchanged in a classifier model without a strong effect on its performance.

The GA solves an optimization problem based on a search guided by the fitness function. Ideally, the values for the P_m,

P_c, and α parameters chosen in the GA only affect the convergence rate but will eventually evolve to the same global maximum. However, when the dimensionality of the feature space is very large and the design samples are sparse, the GA often reaches local maxima corresponding to different feature sets, as can be seen in Table II. Similarly, the stepwise feature selection may reach a different local maximum and choose a feature set different from those chosen by the GA. The different feature sets may provide different or similar performance. The latter is often a result of the correlation among the features, as described above.

For the linear discriminant classifier, the stepwise LDA procedure can select near-optimal features for the classification task. We have shown that the GA could select a feature set comparable to or slightly better than that selected by the stepwise LDA. The number of generations that the GA had to evolve to reach the best selection increased with the dimensionality of the feature space as expected. However, even in a 281-dimensional feature space, it only took 169 generations to find a better feature set than that selected by stepwise LDA. Further search up to 500 generations did not find other feature combinations with better performance. Although the difference in A_z did not achieve statistical significance, probably due to the large standard deviation in A_z when the number of case samples in the ROC analysis was small, the improvements in A_z in this and our previous studies[34] indicate that the GA is a useful feature selection method for classifier design. One of the advantages of GA-based feature selection is that it can search for near-optimal feature sets for any types of linear or nonlinear classifiers, whereas the stepwise LDA procedure is more tailored to linear discriminant classifiers. Furthermore, the fitness function in the GA can be designed such that features with specific characteristics are favored. One of the applications in this direction is to select features to design a classifier with high sensitivity and high specificity for classification of malignant and benign lesions.[49,50] Although the GA requires much longer computation time than the stepwise LDA to search for the best feature set, the flexibility of the GA makes it an increasingly popular alternative for solving machine learning and optimization problems. Since feature selection is performed only during training of a classifier, the speed of a trained classifier for processing test cases is not affected by the choice of the feature selection method. Therefore, the longer computation time of GA is not a problem in practice if the GA can provide a better feature set for a given classification task.

V. CONCLUSIONS

In this study, we evaluated the effectiveness of morphological and texture features extracted from mammograms for classification of malignant and benign microcalcification clusters. We also compared a GA-based feature selection method and a stepwise feature selection procedure based on linear discriminant analysis. It was found that the best feature set was selected from the combined morphological and texture feature space by the GA-based method. A linear discriminant classifier using the best feature set and a properly chosen decision threshold could correctly identify 35% of the benign clusters without missing any malignant clusters. If the average discriminant score from all views of the same cluster was used for classification, the accuracy improved to 50% specificity at 100% sensitivity. Alternatively, if the minimum discriminant score from all views of the same cluster was used, the accuracy would be 32% specificity at 100% sensitivity. This information may be used to reduce unnecessary biopsies, thereby improving the positive predictive value of mammography. Although these results were obtained with a relatively small data set, they demonstrate the potential of using CAD techniques to analyze mammograms and to assist radiologists in making diagnostic decisions. Further studies will be conducted to evaluate the generalizability of our approach in large data sets.

ACKNOWLEDGMENTS

This work is supported by USPHS Grant No. CA 48129 and by U.S. Army Medical Research and Materiel Command Grant No. DAMD 17-96-1-6254. Berkman Sahiner is also supported by a Career Development Award by the U.S. Army Medical Research and Materiel Command (DAMD 17-96-1-6012). Nicholas Petrick is also supported by a grant from The Whitaker Foundation. The content of this publication does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E. Metz, Ph.D. for use of the LABROC and CLABROC programs.

[a]Electronic mail: chanhp@umich.edu

[1]D. D. Adler and M. A. Helvie, ''Mammographic biopsy recommendations,'' Current Opinion in Radiology **4**, 123–129 (1992).
[2]D. B. Kopans, ''The positive predictive value of mammography,'' Am. J. Roentgenol. **158**, 521–526 (1991).
[3]M. Sabel and H. Aichinger, ''Recent developments in breast imaging,'' Phys. Med. Biol. **41**, 315–368 (1996).
[4]F. Shtern, C. Stelling, B. Goldberg, and R. Hawkins, ''Novel technologies in breast imaging: National Cancer Institute perspective,'' Society of Breast Imaging, Orlando, Florida, 153–156 (1995).
[5]L. V. Ackerman and E. E. Gose, ''Breast lesion classification by computer and xeroradiograph,'' Cancer **30**, 1025–1035 (1972).
[6]J. Kilday, F. Palmieri, and M. D. Fox, ''Classifying mammographic lesions using computerized image analysis,'' IEEE Trans. Med. Imaging **12**, 664–669 (1993).
[7]Z. Huo, M. L. Giger, C. J. Vyborny, U. Bick, P. Lu, D. E. Wolverton, and R. A. Schmidt, ''Analysis of spiculation in the computerized classification of mammographic masses,'' Med. Phys. **22**, 1569–1579 (1995).
[8]B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, ''Classification of masses on mammograms using rubber-band straightening transform and feature analysis,'' Proc. SPIE **2710**, 44–50 (1996).
[9]B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, ''Computerized characterization of masses on mammograms: The rubber-band straightening transform and texture analysis,'' Med. Phys. **25**, 516–526 (1998).
[10]S. Pohlman, K. A. Powell, N. A. Obuchowshi, W. A. Chilote, and S. Grundfest-Broniatowski, ''Quantitative classification of breast tumors in digitized mammograms,'' Med. Phys. **23**, 1337–1345 (1996).
[11]W. G. Wee, M. Moskowitz, N.-C. Chang, Y.-C. Ting, and S. Pemmeraju, ''Evaluation of mammographic calcifications using a computer program,'' Radiology **116**, 717–720 (1975).

[12]S. H. Fox, U. M. Pujare, W. G. Wee, M. Moskowitz, and R. V. P. Hutter, ''A computer analysis of mammographic microcalcifications: global approach.,'' Proceedings of the IEEE 5th International Conference on Pattern Recognition., IEEE, New York, 624–631 (1980).
[13]H. P. Chan, L. T. Niklason, D. M. Ikeda, and D. D. Adler, ''Computer-aided diagnosis in mammography: Detection and characterization of microcalcifications,'' Med. Phys. **19**, 831 (1992).
[14]H. P. Chan, D. Wei, L. T. Niklason, M. A. Helvie, K. L. Lam, M. M. Goodsitt, and D. D. Adler, ''Computer-aided classification of malignant/benign microcalcifications in mammography,'' Med. Phys. **21**, 875 (1994).
[15]H. P. Chan, D. Wei, K. L. Lam, S.-C. B. Lo, B. Sahiner, M. A. Helvie, and D. D. Adler, ''Computerized detection and classification of microcalcifications on mammograms,'' Proc. SPIE **2434**, 612–620 (1995).
[16]H. P. Chan, B. Sahiner, K. L. Lam, D. Wei, M. A. Helvie, and D. D. Adler, ''Classification of malignant and benign microcalcifications on mammograms using an artificial neural network,'' Proc. of World Congress on Neural Networks II, 889–892 (1995).
[17]H. P. Chan, D. Wei, K. L. Lam, B. Sahiner, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, ''Classification of malignant and benign microcalcifications by texture analysis,'' Med. Phys. **22**, 938 (1995).
[18]H. P. Chan, B. Sahiner, D. Wei, M. A. Helvie, D. D. Adler, and K. L. Lam, ''Computer-aided diagnosis in mammography: Effect of feature classifier on characterization of microcalcifications,'' Radiology **197(P)**, 425 (1995).
[19]L. Shen, R. M. Rangayyan, and J. E. L. Desautels, ''Application of shape analysis to mammographic calcifications,'' IEEE Trans. Med. Imaging **13**, 263–274 (1994).
[20]Y. Wu, M. T. Freedman, A. Hasegawa, R. A. Zuurbier, S. C. B. Lo, and S. K. Mun, ''Classification of microcalcifications in radiographs of pathologic specimens for the diagnosis of breast cancer,'' Academic Radiology **2**, 199–204 (1995).
[21]Y. Jiang, R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vyborny, and K. Doi, ''Malignant and benign clustered microcalcifications: automated feature analysis and classification,'' Radiology **198**, 671–678 (1996).
[22]D. L. Thiele, C. Kimme-Smith, T. D. Johnson, M. McCombs, and L. W. Bassett, ''Using tissue texture surrounding calcification clusters to predict benign vs malignant outcomes,'' Med. Phys. **23**, 549–555 (1996).
[23]A. P. Dhawan, Y. Chitre, C. Kaiser-Bonasso, and M. Moskowitz, ''Analysis of mammographic microcalcifications using gray-level image structure features,'' IEEE Trans. Med. Imaging **15**, 246–259 (1996).
[24]L. V. Ackerman, A. N. Mucciardi, E. E. Gose, and F. S. Alcorn, ''Classification of benign and malignant breast tumors on the basis of 36 radiographic properties,'' Cancer **31**, 342 (1973).
[25]A. G. Gale, E. J. Roebuck, P. Riley, and B. S. Worthington, ''Computer aids to mammographic diagnosis,'' Br. J. Radiol. **60**, 887–891 (1987).
[26]D. J. Getty, R. M. Pickett, C. J. D'Orsi, and J. A. Swets, ''Enhanced interpretation of diagnostic images,'' Invest. Radiol. **23**, 240 (1988).
[27]C. J. D'Orsi, D. J. Getty, J. A. Swets, R. M. Pickett, S. E. Seltzer, and B. J. McNeil, ''Reading and decision aids for improved accuracy and standardization of mammographic diagnosis,'' Radiology **184**, 619–622 (1992).
[28]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 **187**, 81–87 (1993).
[29]J. A. Baker, P. J. Kornguth, J. Y. Lo, M. E. Williford, and C. E. Floyd, ''Breast cancer: Prediction with artificial neural network based on BI-RADS standardization lexicon,'' Radiology **196**, 817–822 (1995).
[30]J. Y. Lo, J. A. Baker, P. J. Kornguth, J. D. Iglehart, and C. E. Floyd, ''Predicting breast cancer invasion with artificial neural networks on the basis of mammographic features,'' Radiology **203**, 159–163 (1997).
[31]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 microcalcifications on mammograms: texture analysis using an artificial neural network,'' Phys. Med. Biol. **42**, 549–567 (1997).
[32]J. H. Holland, *Adaptation in Natural and Artificial Systems* (University of Michigan Press, Ann Arbor, MI, 1975).
[33]D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, New York, 1989).
[34]B. Sahiner, H. P. Chan, N. Petrick, D. Wei, 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 on mammograms,'' Med. Phys. **23**, 1671–1684 (1996).
[35]M. J. Norusis, *SPSS for Windows Release 6 Professional Statistics* (SPSS Inc., Chicago, IL, 1993).
[36]P. A. Lachenbruch, *Discriminant Analysis* (Hafner, New York, 1975), Chaps. 1, 3.
[37]C. E. Metz, J. H. Shen, and B. A. Herman, ''New methods for estimating a binormal ROC curve from continuously-distributed test results,'' Annual Meeting of the American Statistical Association, Anaheim, CA (1990).
[38]H. P. Chan, S. C. B. 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).
[39]B. Sahiner, H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, ''Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images,'' IEEE Trans. Med. Imaging **15**, 598–610 (1996).
[40]R. M. Haralick, K. Shanmugam, and I. Dinstein, ''Texture features for image classification,'' IEEE Trans. Syst. Man Cybern. **SMC-3**, 610–621 (1973).
[41]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 discriminant analysis in texture feature space,'' Phys. Med. Biol. **40**, 857–876 (1995).
[42]K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. (Academic, New York, 1990), Chap. 3.
[43]H. P. Chan, B. Sahiner, R. F. Wagner, N. Petrick, and J. Mossoba, ''Effects of sample size on classifier design: Quadratic and neural network classifiers,'' Proc. SPIE **3034**, 1102–1113 (1997).
[44]H. P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, ''Classifier design for computer-aided diagnosis in mammography: Effects of finite sample size,'' Med. Phys. **24**, 1034–1035 (1997).
[45]R. F. Wagner, H. P. Chan, B. Sahiner, N. Petrick, and J. T. Mossoba, ''Finite-sample effects and resampling plans: Applications to linear classifiers in computer-aided diagnosis,'' Proc. SPIE **3034**, 467–477 (1997).
[46]F. J. Ferri, P. Pudil, M. Hatef, and J. Kittler, ''Comparative study of techniques for large-scale feature selection,'' Pattern Recognition in Practice IV, 403–413 (1994).
[47]W. Siedlecki and J. Sklansky, ''A note on genetic algorithm for large-scale feature selection,'' Pattern Recogn. Lett. **10**, 335–347 (1989).
[48]C. E. Metz, P. L. Wang, and H. B. Kronman, ''A new approach for testing the significance for differences between ROC curves measured from correlated data,'' in *Information Processing in Medical Imaging*, edited by F. Deconinck (The Hague, Martinus Nijhoff, 1984), pp. 432–445.
[49]B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, ''Classification of malignant and benign breast masses: Development of a high-sensitivity classifier using a genetic algorithm,'' Radiology **201**, 256–257 (1996).
[50]B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, ''Design of a high-sensitivity classifier based on genetic algorithm: Application to computer-aided diagnosis,'' Phys. Med. Biol. **43**, 2853–2871 (1998).