Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue

Berkman Sahiner, Heang-Ping Chan, Datong Wei, Nicholas Petrick, Mark A. Helvie, Dorit D. Adler, and Mitchell M. Goodsitt
The University of Michigan, Department of Radiology, Ann Arbor, Michigan 48109-0030

(Received 5 July 1995; resubmitted 14 May 1996; accepted for publication 3 July 1996)

We investigated a new approach to feature selection, and demonstrated its application in the task of differentiating regions of interest (ROIs) on mammograms as either mass or normal tissue. The classifier included a genetic algorithm (GA) for image feature selection, and a linear discriminant classifier or a backpropagation neural network (BPN) for formulation of the classifier outputs. The GA-based feature selection was guided by higher probabilities of survival for fitter combinations of features, where the fitness measure was the area A_z under the receiver operating characteristic (ROC) curve. We studied the effect of different GA parameters on classification accuracy, and compared the results to those obtained with stepwise feature selection. The data set used in this study consisted of 168 ROIs containing biopsy-proven masses and 504 ROIs containing normal tissue. From each ROI, a total of 587 features were extracted, of which 572 were texture features and 15 were morphological features. The GA was trained and tested with several different partitionings of the ROIs into training and testing sets. With the best combination of the GA parameters, the average test A_z value using a linear discriminant classifier reached 0.90, as compared to 0.89 for stepwise feature selection. Test A_z values with a BPN classifier and a more limited feature pool were 0.90 with GA-based feature selection, and 0.89 for stepwise feature selection. The use of a GA in tailoring classifiers with specific design characteristics was also discussed. This study indicates that a GA can provide versatility in the design of linear or nonlinear classifiers without a trade-off in the effectiveness of the selected features.

Key words: mammography, computer-aided diagnosis, genetic algorithms, feature selection

I. INTRODUCTION

Computer-aided diagnosis (CAD) for detection and classification of breast abnormalities on mammograms is an active area of research.[1] Clinical studies have shown that 10% to 30% of breast cancers visible on mammograms in retrospective studies were initially missed by radiologists,[2,3] and that only 15% to 30% of the patients who have undergone biopsy due to a suspicious finding on mammograms are found to have breast cancer.[4,5] CAD methods have the potential of reducing the false-negative rate while improving the positive predictive values of the mammographic abnormalities.

Masses are important indicators of malignancy on mammograms. In recent years, considerable effort has been devoted to the development of computerized methods for detection and classification of masses.[6–12] In all of these investigations, the detection or classification task relies on the use of features extracted from the digitized mammograms. The extracted features represent properties of pixels (or groups of pixels) which contain characteristic information of the masses. In this paper, we report our development of a computerized method for classification of regions of interest (ROIs) on mammograms as either masses or normal tissue, with particular emphasis on a genetic algorithm for feature selection.

Feature selection is a very important step in classification,[8,10,11,13–16] because the inclusion of inappropriate features often adversely affects classifier performance, especially when the training set is not sufficiently large. The methods employed for feature selection vary. In some approaches,[7,9] very few features were used, and the process of feature selection was not clearly described. It is reasonable to assume that the features were selected on the basis of some prior knowledge from clinical experience. Wu *et al.*[13] selected 14 features from a total of 43 for classification of malignant and benign masses, and observed an improvement in classification accuracy when the reduced feature space was used instead of the entire feature space. The criterion for selection was the difference of the average values of individual features between the two classes. Goldberg *et al.*[14] first selected five features from a total of 26 based on the ability of the individual features to discriminate between malignant and benign masses. Subsequently, based on their pairwise discriminatory ability, three final features were selected from the remaining five features. In the study by Chitre *et al.*,[15] the criterion for texture feature selection was the combination of a classification error and a clustering technique using individual features independently. In our previous studies, we employed a stepwise feature selection procedure in linear discriminant analysis (LDA),[10,11] in which a feature is included or excluded at each step based on a chosen statistical criterion. The LDA takes into account the correlation between the features and the joint probability distri-

 0094-2405/96/23(10)/1671/14/$10.00

bution of the feature vectors in the multidimensional feature space.

Many feature selection methods have been explored in CAD. However, the best method which can provide the highest accuracy for a given application is still in question. This is partly because feature selection is theoretically a difficult problem.[17] It is well known, for example, that the two independent features that yield the highest classification accuracy in a feature set may not constitute the best pair of features together.[18] In the training process in CAD, the classifier can be designed so that the probability of training error will not increase when the number of selected features increases. However, when both training and testing are desired, the problem becomes more complicated due to overfitting. Test results can deteriorate when the number of selected features increases,[13] especially when the number of training cases is small. It is imperative to select a smaller subset of features to overcome the so-called "curse of dimensionality"[19,20] (decrease in classification accuracy of the test set with an increasing number of features) if the ratio of the number of training cases to the number of available features is not sufficiently large. Several recipes for feature selection are mentioned in the literature,[19,21] but none of these, except for an exhaustive search procedure, is optimal.

Genetic algorithms (GAs), first introduced by Holland in the early seventies,[22] are becoming increasingly popular in solving optimization and machine learning problems.[23,24] The fundamental principle underlying GAs is based on natural selection. To solve an optimization task, a GA maintains a population of bit strings, which are referred to as chromosomes. Each chromosome corresponds to a possible solution of the problem. In each generation of the GA, the population is probabilistically modified, generating new chromosomes which may have a better chance of solving the optimization problem. GAs have been applied to complex optimization problems such as the control of a gas-pipeline system,[25] design of jet engine turbines,[26] training of a backpropagation neural network,[27] feature selection for an artificial neural network,[28] and automated detection of lung nodules.[29] GAs usually yield nonoptimal, but near-optimal solutions. They are thus well-suited for feature selection problems in large feature spaces, where the optimal solution is practically impossible to compute, and a near-optimal solution is the best alternative.

In this paper, we studied the ability of a GA to select features from a large feature space. Our goal was to introduce a more effective and versatile feature selection mechanism. The effectiveness and the versatility of the GA was demonstrated by its application to the problem of classification of masses and normal tissue on mammograms. The feature space included local and global multiresolution texture features[30] as well as morphological features.[31] The rest of the paper is organized as follows. In the next section, we briefly discuss important components of a GA. In Sec. III, we describe our image database, background correction method, extraction of texture and morphological features, and the GA implementation for feature selection. In Sec. IV, we evaluate the dependence of the classification results on different GA parameters. Section V contains a discussion of these results. Finally, Sec. VI concludes the investigation and provides a scope for further research.

II. GENETIC ALGORITHMS

In natural evolution, the basic problem of each population is to find beneficial adaptations to a complex environment. The characteristics that each individual has gained or inherited are carried in its chromosomes and each individual reproduces more or less in proportion to its fitness within the environment. Crossover and mutation provide the possibility of evolution toward better-fit individuals.

Genetic algorithms[22–24] apply the principles of natural selection to machine learning. To solve an optimization problem, a GA requires five components, which are analogous to components of natural selection. These components are described below.

A. Encoding

Encoding is a way of representing the decision variables of the optimization problem in a string of binary digits called chromosomes. If there are v decision variables in an optimization problem and each decision variable is encoded as an n-digit binary number, then a chromosome is a string of $n \times v$ binary digits. Each chromosome is a possible solution to the optimization problem.

B. Initial population

The initial population is a set of chromosomes offered as an initial solution or as a starting point in the search for better chromosomes. The initial population must be large and diverse enough to allow evolution toward better individuals. In general, the population is initialized at random to a bit string of 0's and 1's. However, more directed methods for finding the initial population can sometimes be used to improve convergence time.

C. Fitness function

The fitness function rates chromosomes (i.e., possible solutions) in terms of how good they are in solving the optimization problem. It thus plays the role of the environment. The fitness function returns a single value for each chromosome, which is then used to determine the probability that this chromosome will be selected as a parent to generate new chromosomes. The fitness function is the primary GA component in which a traditional GA is tailored to a specific problem.

D. Genetic operators

Genetic operators are applied probabilistically to chromosomes of a generation to produce a new generation of chromosomes. Three basic operators are parent selection, crossover, and mutation. The parent selection operation mimics the natural selection process by selecting which chromosomes will be used to create a new generation, where the fittest chromosomes reproduce most often. The crossover op-

eration refers to the exchange of substrings of two chromosomes to generate two new offspring. After parents are selected, and crossover generates two new chromosomes, the operation of mutation is applied to each bit in the string. Mutation simply alters the binary value of the bit when a random value generated for the bit is less than a predefined mutation rate.

E. Working parameters

A set of parameters, which includes the number of chromosomes in each generation, the crossover rate, the mutation rate, and the stopping criterion, is predefined to guide the GA. The crossover and mutation rates, assigned as real numbers between 0 and 1, are used as thresholds to determine whether the operators will be applied or not. The stopping criterion is predefined as the number of generations the algorithm is to be run or as a tolerance value for the fitness function.

Two forces, exploration and exploitation, interact in the search for better-fit chromosomes. Exploitation occurs in the form of parent selection. Chromosomes with higher fitness exploit this fitness by reproducing more often. Exploration occurs in the form of mutation and crossover, which allow the offspring to achieve a higher fitness than their parents. Crossover is the key to exploration, whereas mutation provides background variation and occasionally introduces beneficial genes into the chromosomes. For a successful GA, exploration and exploitation have to be in good balance. With too much exploitation, the GA may be stuck with copies of the same chromosome after a few generations, whereas with too much exploration, good genes may never be able to accumulate in the genetic pool.

GAs are ideal for sampling large search spaces and locating the regions of enhanced opportunity. Although GAs yield near-optimal solutions rather than optimal ones, obtaining such near-optimal solutions are usually the best that one can do in many complex optimization problems involving large numbers of parameters.

III. METHODS

A. Data set

The mammograms used in this study were randomly selected from the files of patients who had undergone biopsy in the Department of Radiology at the University of Michigan. The criterion for inclusion of a mammogram in the data set was that the mammogram contained a biopsy-proven mass. To avoid the effect of repetitive grid lines on the image texture, mammograms that contained these grid lines caused by the stationary grid of some older mammographic units were excluded. The data set included 168 mammograms, with a mixture of benign ($n=85$) and malignant ($n=83$) masses. The visibility of the masses was ranked by an experienced breast radiologist on a scale of 1 to 10, where a ranking of 1 corresponded to the most visible category. The distribution of the visibility ranking of the masses is shown in Fig. 1. It

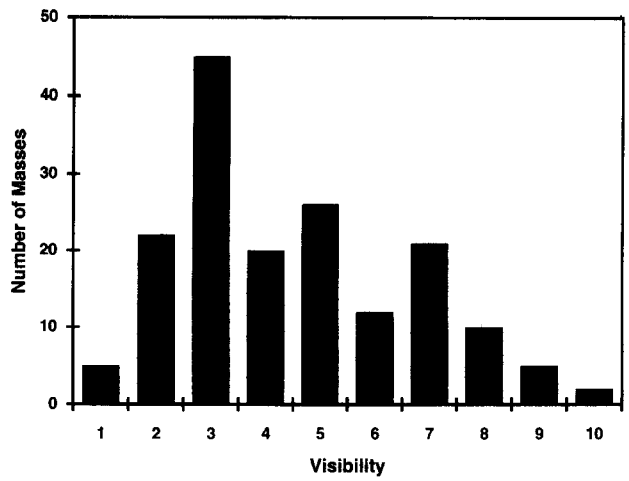


FIG. 1. The distribution of the visibility ranking of the masses in the data set.

can be observed that the visibility of the masses in our data set ranged from subtle to obvious.

The mammograms were digitized with a LUMISYS DIS-1000 laser scanner at a pixel resolution of 100 μm$\times$100 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly and inversely proportional to the optical density (OD) within the range of 0.1- to 2.8-OD units, with a slope of -0.001-OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually. The OD range of the digitizer was 0 to 3.5.

Four different ROIs, each with 256$\times$256 pixels, were selected from each mammogram. One of the selected ROIs contained the true mass as identified by an experienced radiologist and verified by biopsy. In addition to the ROI that contained the true mass location, the radiologist in the study was asked to select three presumably normal ROIs from the mammogram. The first of these three ROIs contained primarily dense tissue which could mimic a mass lesion, the second ROI contained mixed dense/fatty tissue, and the third contained mainly fatty tissue. An example of each of these ROIs is shown in Fig. 2.

B. Background correction

Breast masses are superimposed on structured background tissue in the ROIs. In most cases, this background tissue is not uniform over our 256$\times$256 pixel ROI. For example, one side of the ROI may contain denser tissue than the other side, or, when the mass is close to the outer edge of the breast, one corner of the ROI may contain a nonbreast region. This nonuniformity may affect texture and morphological features that are extracted from the ROI. To reduce this effect, we developed a correction method that estimated the low-frequency background level based on the image intensities in a band of pixels surrounding the ROI. The background level at each pixel on the edge of the ROI was first estimated by gray-level averaging in a rectangular region surrounding the pixel. The background level of a pixel inside the ROI was then estimated by interpolation using the background pixel

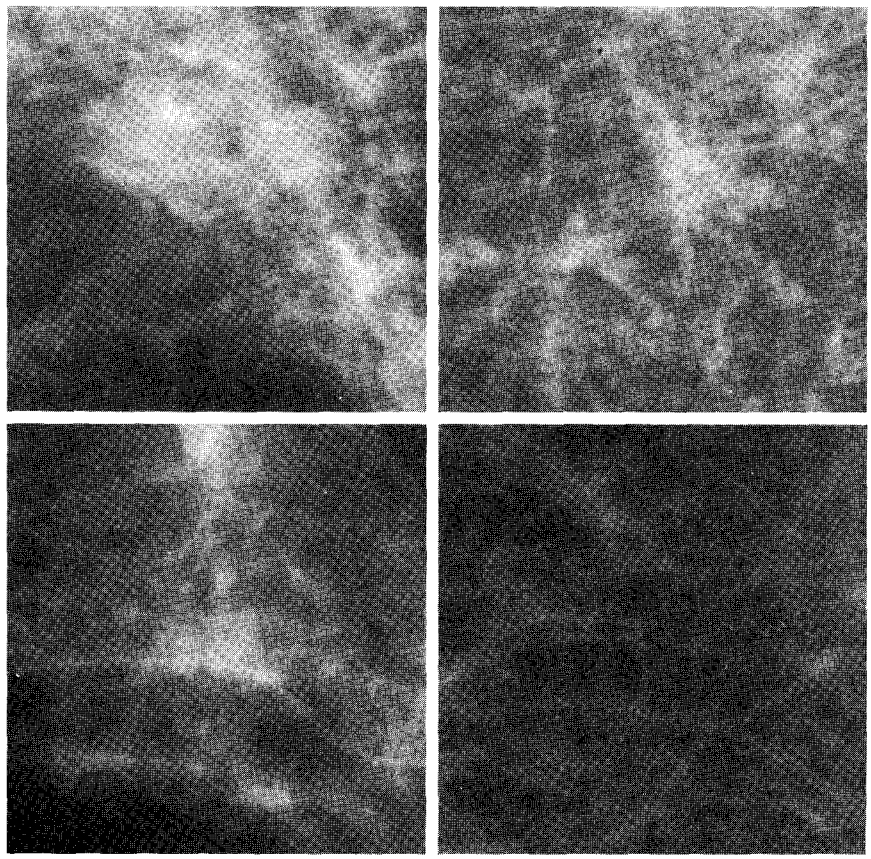


FIG. 2. An example of the mass and normal ROIs selected from one of the mammograms used in this study. The four ROIs are upper left—mass; upper right—mixed dense/fatty tissue; lower left—dense tissue; lower right—fatty tissue.

values on the edges. A more detailed description of this background correction method can be found in the literature.[10,32]

C. Feature extraction

1. Texture features

The texture features used in this study were calculated from spatial gray-level dependence (SGLD) matrices. The (i,j)th element of an SGLD matrix is the joint probability that gray levels i and j occur in a direction θ at a distance of d pixels apart in an image. We computed global texture features, which represent the average texture measures throughout the entire ROI, and local texture features, which represent (i) the texture measure of a denser subregion inside the ROI which is likely to contain the mass, and (ii) the texture difference between this subregion and other peripheral regions in the ROI which contain normal breast tissue. The method used for the computation of SGLD matrices and multiresolution texture analysis are explained in full detail elsewhere.[30] A brief description is given below.

Wavelet transform[33] using the four-coefficient Daubechies wavelet filter was applied to each ROI to decompose the image into a low-pass image and three high-pass subband images. For extracting global multiresolution texture features, we used the original image (scale=1) and the low-pass images at scales 2 and 4 to formulate SGLD matrices at $d=1$ in the transformed images. The distance of $d=1$ at these scales was equivalent to distances of 1, 2, and 4 in the original image. The wavelet coefficients at scale 8 were obtained with wavelet filtering but without down-sampling. The coefficients at scale 8 were used to formulate SGLD matrices at $d=2,3,4,\ldots,12$. Since no down-sampling was used at scale 8, these distances between pixel pairs were equivalent to distances of 8,12,16,...,48 in the original image. Thus a total of 14 distances were used. At each distance, four SGLD matrices at $\theta=0°$, 45°, 90°, and 135° were determined. Thirteen texture features were calculated from each SGLD matrix. The features at $\theta=0°$, 90° and at $\theta=45°$, 135° were averaged separately. Thus 26 texture features were computed for each d, resulting in a total of 364 global features.

For extracting local texture features, five subregions were automatically identified in the background-corrected ROI: a 90×90 pixel object subregion that contained the suspicious dense tissue or the mass, and four 64×64 pixel peripheral subregions that were located in the four corners of the ROI. The suspicious object subregion was automatically detected by searching for the highest average gray-level inside the ROI using a 90×90 moving box. For a given d, an SGLD matrix was derived from the object subregion, and a background SGLD matrix was derived from the pixel pairs in the

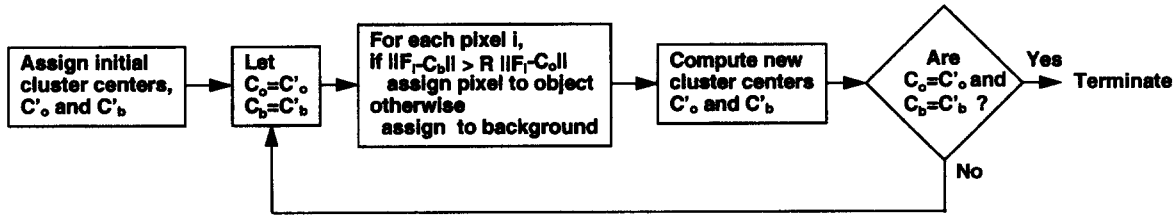


FIG. 3. A schematic of the clustering algorithm.

four peripheral subregions. The SGLD matrices were computed at $d=1,2,4$, and 8. Analogous to the global texture feature extraction, for a given d, 26 features were computed from the object SGLD matrix, and 26 features were computed from the background SGLD matrix. The local texture feature space therefore consisted of 104 features extracted from the object subregion, and 104 features calculated as the differences between the corresponding features extracted from the object and peripheral subregions. This resulted in a total of 208 local texture features.

The detail images in the wavelet transform can be expected to contain useful information for texture-based classification of a large class of images. However, in our previous studies, we found that using the SGLD texture features based on the detail images in the wavelet transform domain did not result in proper classification of breast masses and normal breast tissue.[11] Since this study focused on the feature selection aspect of classification, we did not attempt to search for new texture features that are presumably present in the detail images.

2. Morphological features

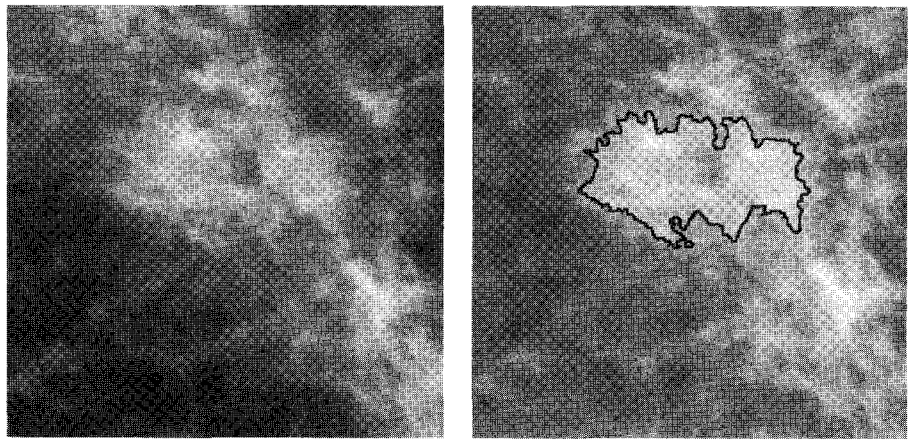
We have developed an automated algorithm for segmentation of an ROI into an object region and background tissue.[31] The morphological features are extracted automatically from the object region after the segmentation is performed.

We used a pixel-by-pixel clustering algorithm followed by binary object detection for ROI segmentation. Pixel-by-pixel clustering algorithms have found widespread use in segmentation of remote sensing data,[34] where multispectral and/or multisource data are obtained for each pixel in the image. Data points for each pixel are regarded as components of a multidimensional feature vector, and pixels with feature vectors of similar characteristics are assigned to the same class using a clustering algorithm. Our data set contains a single data point (the gray level) for each pixel. We derived several filtered images from this single image, and used the original and filtered pixel values as the components of the feature vectors in the clustering algorithm. Inclusion of the filtered images makes it possible to incorporate neighborhood information into the classification of each pixel.

Our clustering algorithm, depicted in Fig. 3, is very similar to the migrating means algorithm.[34] The goal is to classify pixel p_i as either an object or a background pixel. This is achieved by clustering with feature vector $F_i=[f(1),\ldots,f(L)]$ of length L, where L is the total number of images used in clustering. The algorithm starts by choosing initial cluster center vectors, for the object and the background, as described below. Let $C_o=[c_o(1),\ldots,c_o(L)]$ and $C_b=[c_b(1),\ldots,c_b(L)]$ denote these cluster center vectors, respectively. Let $d_o(i)$ denote the Euclidean distance between F_i and C_o, $d_b(i)$ denote the Euclidean distance between F_i and C_b, and R denote a constant distance ratio. If $d_b(i)/d_o(i)>R$, the pixel p_i is temporarily classified as an object pixel; otherwise, it is classified as a background pixel. If $R=1$, the algorithm becomes identical to the migrating means algorithm. After this temporary classification, two new cluster center vectors are computed. The lth component of the new object and background center vectors are the averages of the lth components for pixels temporarily classified as object and background pixels, respectively. If the new cluster centers are different from the previous ones, the procedure of temporary classification is repeated, otherwise, the clustering is completed. In this paper, we used $R=3.75$ so that F_i had to be much closer to C_o than to C_b to be classified as an object pixel. This conservative criterion reduces the chance that a mass region merges with adjacent tissue. However, it also slightly underestimates the mass size so that the detected edge is often within the margin of the mass. The initial center vectors were chosen such that each component of the initial object center vector is 1.1 times the average of that component over the entire ROI, and each component of the initial background center vector is 0.9 times the same average.

After clustering, the ROI may contain several disconnected objects. To obtain a single suspected mass object, we selected the largest connected object among all detected objects. We finally applied region growing to a small region outside the boundary of the suspected object to get a better definition of its borders. To achieve this, we thresholded the original image pixels that were within ten pixels of the object border. The threshold value was chosen experimentally to be the difference between the mean of the pixel values inside the object and half of their standard deviation. Figure 4 shows an example of the result of our segmentation algorithm.

In this paper, we used three filtered images along with the original image to form the feature vectors. The first filtered image was obtained by median filtering with a 5×5 kernel. The second and third filtered images were edge-enhanced images at different resolutions.[31] Each filtered image, as well as the original image was linearly normalized between 0 and S_l, where S_l, $l=1\ldots L$ is a scaling factor. The scaling factors



(a) (b)

FIG. 4. (a) An ROI with an ill-defined mass, (b) mass object extracted automatically by the clustering algorithm and superimposed on the background-corrected ROI.

S_i were chosen experimentally to be $S_1 = S_2 = 1400$ for the original image and the median filtered image, and $S_3 = S_4 = 770$ for the edge-enhanced images. Therefore, the original image and the median filtered image were weighted approximately twice as much as the edge-enhanced images in the clustering algorithm. This bias in favor of the weights of the original image and the median filtered image was necessary because the algorithm showed a tendency to segment only disconnected edges if all images were equally weighted.

After detection of a single suspicious object within each ROI, features were extracted from the object and its margins. We extracted eleven shape features from each object, and four features from the margins of each object. The shape features included the number of edge pixels, area, circularity, rectangularity, contrast, the ratio of the number of edge pixels to the area, and five normalized radial length features. A detailed discussion of the shape features used in this study can be found in Ref. 35. The margin features were computed as follows. First, the mean and the standard deviation of the pixel values inside the object were computed. Next, pixels in a boundary region outside the object but within a distance of 15 pixels from the object border were thresholded. The values of the thresholds were chosen to be the mean minus 0.5, 1, 1.5, and 2 times the standard deviation. The number of pixels in the boundary region which were above the thresholds was defined as the margin features. Thus a total of 15 morphological features were extracted from each ROI.

D. Classifiers

In this paper, we investigated GA-based feature selection for two kinds of classifiers, namely (i) a linear classifier based on Fisher's linear discriminant:[20] and (ii) a multilayer backpropagation neural network (BPN).[36] For each ROI, both classifiers produced a scalar, termed the classifier output, which indicated the likelihood that the ROI contained a real mass.

Fisher's linear discriminant is based on a linear projection of the feature space onto the real line such that the ratio of the between-class sum of squares to within-class sum of squares is maximized after the projection.[20] In our two-class problem, the statistical procedure for formulation of the linear discriminant function is equivalent to multiple linear regression.[37] Fisher's linear discriminant is the optimal classifier if the features are distributed as multivariate Gaussian random variables with equal covariance matrices under each class.[21]

The BPN used in this study consisted of an input layer, an output layer, and a single hidden layer. Each layer in the BPN contained a number of nodes, which were connected to previous and subsequent layers by trainable weights. A single feature was applied to each node in the input layer. The net input to each node in the hidden layer and the output layer was a weighted sum of the node outputs from the previous layer. The output of a node was related to its net input by a sigmoidal function. The output layer contained a single node, whose output indicated the likelihood that the ROI contained breast mass tissue. The BPN was trained using batch processing and the delta-bar-delta rule for improved rate of convergence and stability.[32]

Since our purpose in this study is to design a feature selection algorithm, we did not compare BPN and linear discriminant classifiers. Instead, we compared the classification accuracy obtained by using different feature selection methods, with a fixed classifier for each comparison.

E. GA-based feature selection

In this paper, we used a GA to select features for discrimination of mass and nonmass ROIs. In our GA, the number of bits in a chromosome was equal to the total number of available features, and each bit corresponded to an individual feature extracted from the ROIs. A feature was termed "present" in a chromosome if the value of the bit corre-

sponding to that feature was 1. The population was initialized at random, with a small probability P_{init} of having a 1 at each bit location. This allowed the GA to start with a few selected features and grow to a reasonable number of features as the population evolved. The total number of chromosomes at each generation was kept constant at $M=250$.

At each run of the GA, the image data set of 672 ROIs was divided into a training and a test set, with ROIs belonging to the same film grouped into the same set. The training set was used in the GA for feature selection. After feature selection, a classifier was trained using only the GA-selected features of the training set. The classification accuracy of the procedure was evaluated by applying the classifier to the same set of features of the test group, as described below. For studying the effect of GA parameters on the classification accuracy with the linear discriminant classifier, ten random partitionings of training and test sets were obtained for each set of different GA parameters, and the results were averaged in order to reduce the effect of case selection. For experiments with the BPN, 50 random partitionings were used. For both experiments, the number of mass and nonmass ROIs in each training set was 126 and 378 ($\frac{3}{4}$ of the total), respectively, while the number of mass and nonmass ROIs in each test set was 42 and 126 ($\frac{1}{4}$ of the total), respectively.

Inside the GA, the training set was equally divided into two groups, $S1$ and $S2$. For each chromosome, two classifiers were trained, with $S1$ and $S2$ as the training groups, respectively. Only the features present in the chromosome were used as features in classifier training. The classifier trained on group $S1$ was applied to the group $S2$, and vice versa, for calculation of two sets of *pseudotest* classifier outputs. The accuracy of the pseudotest classifier outputs, and the number of selected features were then used to define the fitness of the individual chromosome. This process was repeated for each of the M chromosomes in each generation.

The main component of the fitness function was the area A_z under the receiver operating characteristic (ROC) curve of the pseudotest sets. A widely accepted procedure for computing the ROC curve assumes that the classifier output follows a normal distribution for each class, and fits the ROC curve to the classifier output using maximum likelihood estimation.[38] We adopted this approach when we studied and compared the classification accuracy of our classifiers with the selected feature sets. However, it is computationally expensive to use this approach in the fitness function calculation inside the GA, because it is required for each chromosome in each generation. Instead, we chose to estimate the ROC curve by varying the decision threshold, and determining the true-positive fraction (TPF) as a function of the false-positive fraction (FPF). The A_z value was estimated by numerical integration using the trapezoidal rule. Since the estimation of the A_z was internal to the GA, it did not affect the A_z values reported in the Sec. IV for a set of selected features. Internal to the GA, the fitness ranking of the chromosomes might be slightly different from that obtained by using the maximum likelihood ROC curve. However, the effect on the final selected feature set should be small, because this slight difference did not completely eliminate the lower-ranking chromosomes. A slightly lower-ranking chromosome was assigned a slightly lower probability of being a parent, but it could still be competitive after mutation and crossover if it contained effective features. This minor inaccuracy in the fitness function computation was a trade-off in order to execute the computation in a reasonable amount of time while using the A_z value in the feature selection procedure.

A second component of the fitness function was a penalty term, analogous to Brill's utility term,[28] which was linearly proportional to the number of features present in the chromosome. The purpose of this penalty term was to control the number of selected features and to prevent overfitting in the test stage of classifier design. In other words, the penalty term was designed to improve the classification accuracy, and not for accelerating the computational speed. The function of the penalty term was comparable to those of the F-to-enter and F-to-remove thresholds in the stepwise feature selection method, described in the next subsection. Similar to these corresponding parameters in stepwise feature selection, increasing the penalty term decreased the number of selected features. We studied the effect of the presence of this penalty term on the test results.

In a given generation, the fitness function $f(m)$ for a chromosome m was computed as follows. First, the two pseudotest A_z values, corresponding to pseudotest sets $S1$ and $S2$, were averaged to yield $\bar{A}_z(m)$. Next, a fitness function $\tilde{f}(m)$ was computed as

$$\tilde{f}(m)=\bar{A}_z(m)-\alpha N(m), \tag{1}$$

where $N(m)$ was the number of 1's (present features) in chromosome m and α was the penalty constant. After $\tilde{f}(m)$ was determined for all chromosomes, the maximum $\tilde{f}_{\max}$ and the minimum $\tilde{f}_{\min}$ of $\tilde{f}(m)$ over the population of M chromosomes were calculated. Finally, $\tilde{f}(m)$ was normalized using $\tilde{f}_{\max}$ and $\tilde{f}_{\min}$ to yield the fitness function $f(m)$,

$$f(m)=\left(\frac{\tilde{f}(m)-\tilde{f}_{\min}}{\tilde{f}_{\max}-\tilde{f}_{\min}}\right)^2, \quad 1\leq m\leq M. \tag{2}$$

The genetic operators were applied as follows. First, parent selection was performed using roulette wheel selection.[23] In this method, each chromosome in a generation occupies an area

$$A(m)=\frac{f(m)}{\sum_{m=1}^{M} f(m)} \tag{3}$$

proportional to its fitness, on a roulette wheel. A parent is selected by spinning the roulette wheel, i.e., by generating a random number $\gamma_i\in(0,1]$ and determining the chromosome m_i that satisfies

$$\sum_{m=1}^{m_i-1} A(m)<\gamma_i\leq\sum_{m=1}^{m_i} A(m), \quad i=1,2. \tag{4}$$

After two parents m_1 and m_2 were selected for generating two offspring, a probabilistic decision was made as to

whether crossover should be applied or not. A random number β with uniform distribution in the interval (0,1] was generated and compared to P_c, the probability of crossover. If $\beta > P_c$, then no crossover was applied, and m_1 and m_2 were accepted into the new generation. Otherwise, a random crossover site was selected inside the chromosomes, and each of the parent chromosomes were split into left and right strings at this location. Crossover was completed by combining the left string of m_1 with the right string of m_2, and vice versa.

Finally, mutation was applied to each bit of the chromosomes in the new generation. Again, a random number with uniform distribution in the interval (0,1] was generated, and compared to P_m, the probability of mutation. If P_m was higher, then the bit was complemented. Otherwise, it was left unchanged. We studied the effects of P_c and P_m on the final classification accuracy.

The GA was permitted to evolve for a fixed number of generations. After the evolution was completed, the chromosome with the highest fitness value provided the set of selected features. The entire training set $S1 \cup S2$ was then used in the final multiple linear regression to determine the weight of each selected feature in the classifier. During testing, the values of the selected features of each ROI in the test set were applied as inputs to the trained classifier to calculate the classifier output for that ROI.

To evaluate the classification performance, the classifier output was used as the decision variable, and a test ROC curve was estimated using the LABROC1 program.[39] The LABROC1 program assumes binormal distributions of the decision variable for the normal and abnormal cases, and fits the ROC curve based on maximum likelihood estimation. The area under the fitted ROC curve, A_z, was used as an index of classification accuracy.

F. Stepwise feature selection

For the purpose of comparison with GA-based feature selection, we also studied the classification accuracy of the same classifiers using a well-established feature selection method, called feature selection with stepwise linear discriminant analysis,[21] or stepwise feature selection in short.[40] At each step of the stepwise selection procedure, one feature is entered into or removed from the selected feature pool by analyzing its effect on a selection criterion. In this study, we employed the Wilks' lambda as our selection criterion, which is defined as the ratio of the within-group sum of squares to the total sum of squares of the two classes.[37] The number of features selected by this method are controlled by two parameters, called F-to-enter and F-to-remove. At each step, the stepwise feature selection algorithm first determines the significance of the change, based on F statistics, in Wilks' lambda when a variable is entered into the selected feature pool. If the significance is above the threshold determined by the F-to-enter parameter, then the selected feature pool is augmented with the most significant variable. Next, the algorithm computes the significance of the change in Wilks' lambda when each variable is removed from the selected feature pool. If the significance is below the threshold determined by the F-to-remove parameter, then the least significant variable is removed from the selected feature pool. Increasing either the F-to-enter or the F-to-remove value decreases the number of selected features. Similar to GA-based feature selection, stepwise feature selection is a heuristic procedure. For this reason, the optimal values of F-to-enter and F-to-remove parameters are not known in advance. One has to experiment with these parameters and increase or decrease the number of selected features to obtain the best test performance. A detailed description of the stepwise feature selection procedure and its application to our problems[10,11] can be found in the literature.[21,40]

IV. RESULTS

In the next two subsections, we present the results for evaluation of the effects of various parameters, and for classification with GA-based feature selection using linear discriminant and BPN classifiers, respectively. Since training a linear discriminant classifier was considerably faster than training a BPN, the effects of GA parameters were studied with a linear discriminant classifier. Feature selection for a BPN classifier was performed on a subset of the entire feature set to accelerate training. For both classifiers, a comparison with stepwise feature selection was provided.

A. Feature selection for a linear discriminant classifier

1. Effect of penalty term and number of generations

To determine a reasonable number of generations for the GA to evolve, we selected several combinations of crossover probability (P_c) and mutation probability (P_m), and monitored the growth of the number of selected features. The initial probability of feature presence was fixed at $P_{\text{init}}=0.002$. The GA was allowed to evolve with two different α values of the penalty term in the fitness function of Eq. (1). We observed that the crossover probability P_c did not have a major effect on the number of selected features. However, both α in the penalty term and the mutation probability P_m affected the number of selected features. Figures 5 and 6 plot the average number of selected features over ten training sets versus the generation number for $\alpha=0$ and $\alpha=1/2000$, respectively. The average number of selected features is plotted for $P_m=0.001$ and $P_m=0.003$ in each figure. The crossover probability is kept constant at $P_c=0.7$. The test A_z value obtained up to a given generation is plotted against the generation number in Figs. 7 and 8 for the same conditions ($\alpha=0$ and $\alpha=1/2000$), respectively. The average A_z value over ten test sets is shown.

It is observed that while the average test A_z value does not increase after the 25th generation, the number of selected features keeps increasing beyond the 60th generation for all combinations of GA parameters studied. Since the main component of the fitness function in the GA is the A_z value rather than the number of features, more features may be added into the selected feature pool as long as the area under the ROC curve does not deteriorate. Comparing Figs. 5 and

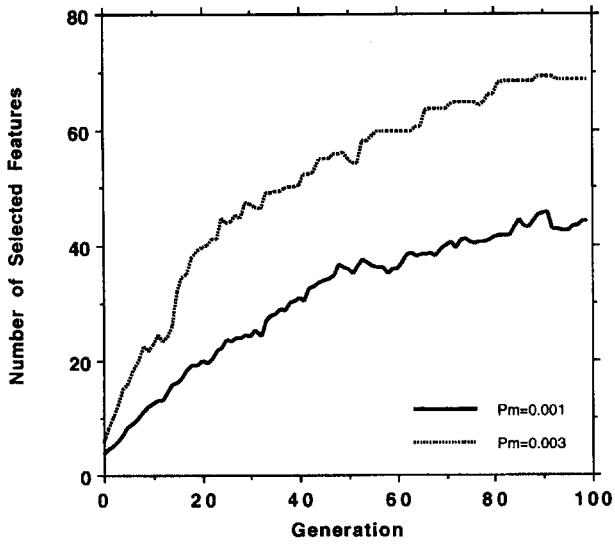


FIG. 5. Evolution of the number of selected features for $\alpha=0$.

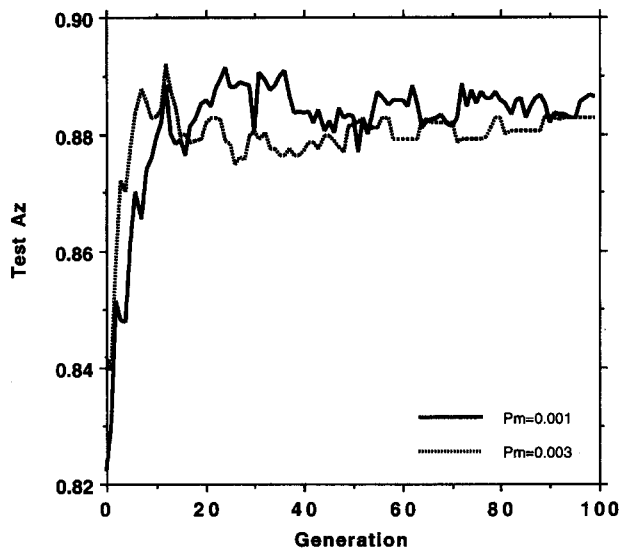


FIG. 7. Evolution of the average test A_z for $\alpha=0$.

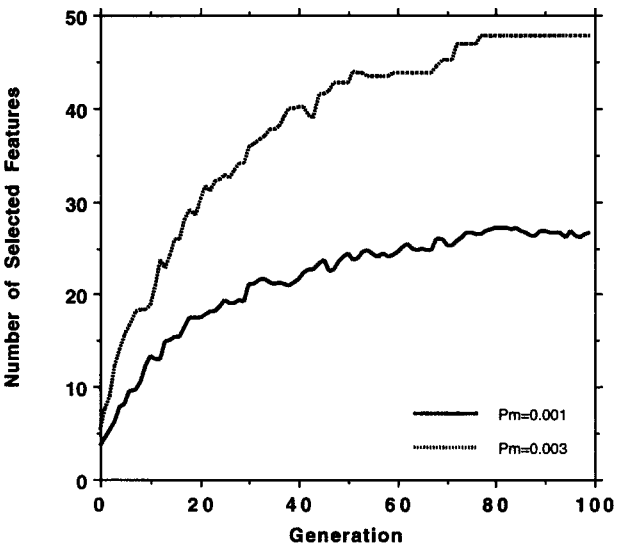


FIG. 6. Evolution of the number of selected features for $\alpha=1/2000$.

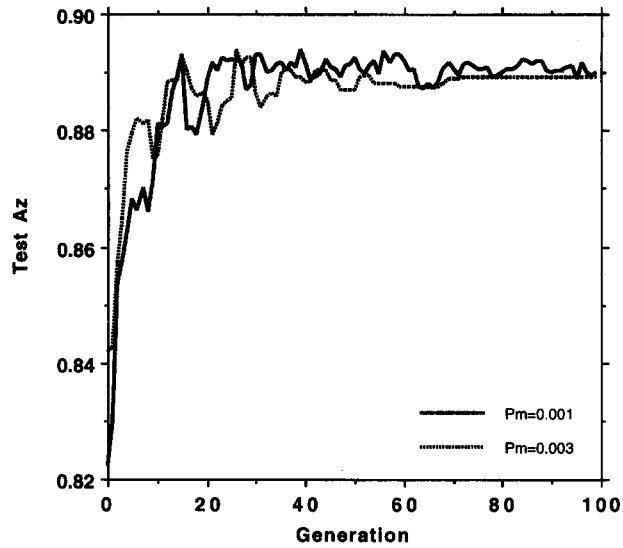


FIG. 8. Evolution of the average test A_z for $\alpha=1/2000$.

6, it can be observed that the penalty term suppressed the number of selected features. The number of selected features eventually leveled off at about the 80th generation when the penalty term was nonzero (Fig. 6).

The average test A_z values at the end of 100 generations were 0.89 for the combinations studied in Fig. 8, and 0.88 for the combinations studied in Fig. 7. The maximum and minimum values of individual test scores for the ten partitions studied were 0.92 and 0.86 for Fig. 8, and 0.92 and 0.85 for Fig. 7. The standard deviation of the individual A_z values, as determined by the LABROC1 program, varied between 0.02 and 0.04.

Since our goal is to select a small number of features while maintaining a high classification accuracy, we performed subsequent GA experiments with $\alpha=1/2000$. Due to computation time constraints, we set the maximum number of generations to be 25 in the following experiments.

2. Effect of initial probability of feature presence (P_{init})

We evaluated the effect of P_{init} on feature selection when the crossover probability P_c and the mutation probability P_m were held constant. The average test A_z values for $P_c=0.9$ and $P_m=0.001$ are tabulated in Table I. It is observed that the performance of the GA reaches a broad maximum when P_{init} is in the range of 0.0005 to 0.020, i.e., when the average number of features in the initial chromosomes is approximately in the range of 0.3 to 12. When P_{init} is out of this range, the average test A_z decreases slightly.

3. Effect of probability of mutation and crossover

The effects of the crossover probability P_c and the mutation probability P_m on the classification accuracy are summarized in Tables II and III, respectively. In Table II, the

TABLE I. The effect of P_{init} on GA performance for $P_m=0.001$, $P_c=0.9$.

P_{init}	Average test A_z	Avg. Num. of features
0	0.88	18.5
0.0005	0.89	17.2
0.001	0.88	20.8
0.002	0.90	20.1
0.005	0.89	18.0
0.010	0.89	23.2
0.020	0.89	22.9
0.050	0.88	32.7

TABLE III. The effect of P_m on GA performance for $P_{\text{init}}=0.002$, $P_c=0.9$.

P_m	Average Test A_z	Avg. Num. of features
0.0005	0.89	16.0
0.001	0.90	20.1
0.003	0.89	32.3
0.005	0.88	33.1
0.007	0.89	33.4
0.009	0.88	33.9

control parameters were fixed at $P_{\text{init}}=0.002$, and $P_m=0.001$, while in Table III, they were fixed at $P_{\text{init}}=0.002$, and $P_c=0.9$. For fixed values of P_{init} and P_m, the average test A_z appears to increase with increasing P_c, while the number of selected features remains relatively constant. On the other hand, for fixed values of P_{init} and P_c, the average test A_z increases initially with increasing P_m, reaching a maximum at $P_m=0.001$, and then decreases slightly as P_m increases beyond 0.003.. Although the variation of the classification accuracy with respect to P_m is not significant, it appears that a reasonable range of choice for P_m is such that the average number of mutations per chromosome per generation is less than 1.5 (0.003× the number of genes per chromosome). Within the range studied, the number of selected features increases with increasing P_m, which may be the reason for the slight deterioration in performance for large P_m.

4. Comparison with LDA classifier and random feature selection

We used a commercial statistics package, SPSS,[40] for LDA classification. The feature selection and formulation of the discriminant function were performed on each of the ten training sets, and the discriminant functions were tested on the corresponding test sets. Using minimization of Wilks' lambda as the feature selection criterion, we varied the two threshold values for F statistics (F-to-enter and F-to-remove) in the SPSS package so that the average test A_z value over the ten partitionings was maximized. The number of selected features and the test results for the ten partitionings are tabulated in Table IV. We chose the best GA classification results (the last line in Table II) for comparison with those of the LDA. The corresponding test A_z values and the number of selected features for each partitioning of the data set are tabulated in Table IV.

TABLE II. The effect of P_c on GA performance for $P_{\text{init}}=0.002$, $P_m=0.001$.

P_c	Average test A_z	Avg. Num. of features
0.1	0.87	18.4
0.3	0.89	18.6
0.5	0.89	17.8
0.7	0.89	18.3
0.9	0.90	20.1

For comparison with these two near-optimal feature selection methods, we performed multiple linear regression training and testing on 20 randomly selected features out of the available 587 features. The test A_z values based on these 20 randomly selected features are also given in Table IV.

B. Feature selection for BPN

Since training a BPN is considerably slower than training a linear discriminant classifier, we modified our training strategy for this classifier. The basic differences between the experiments in this subsection on BPN and the previous subsection on linear discriminant classifier were: (1) In order to handle a smaller feature pool with BPN, we used a single distance for texture features. Based on our previous study of the effects of pixel distance on classification,[10] we selected a pixel distance of $d=20$. The global texture features computed at this pixel distance, plus the morphological features previously described in Sec. III C, constituted the feature pool in this subsection. Therefore, there were a total of 41 features (26 texture and 15 morphological) for the feature selection algorithms to choose from. (2) In order not to repeat the feature selection process several times with several different training sets, the entire data set was used in the feature selection step of the classification procedure. After feature selection was completed, the classifier was trained and tested with 50 different partitionings of the data set into training and test groups. As in the case of linear discriminant classifier, the number of mass and nonmass ROIs in each training set was 126 and 378 ($\frac{3}{4}$ of the total), respectively, while the number of mass and nonmass ROIs in each test set was 42 and 126 ($\frac{1}{4}$ of the total), respectively.

The parameters of the BPN and the GA used in this subsection were as follows. The BPN had a variable number of input nodes, four hidden layer nodes, and a single output node. The BPN was trained for 400 iterations for each chromosome in each generation. The GA was allowed to evolve for a total number of 75 generations. Results of the previous subsections suggest that there is a wide range of choice for the parameters P_{init} and P_m. It appears that a reasonable choice for P_{init} is such that the average number of selected features at generation 0 is in the range of 0.3 to 12, and a reasonable choice for P_m is such that the average number of mutations per chromosome per generation is less than 1.5. For this reason, these parameters of the GA were selected as $P_m=0.02$, and $P_{\text{init}}=0.02$. Since a large probability of crossover seemed to result in the selection of more effective fea-

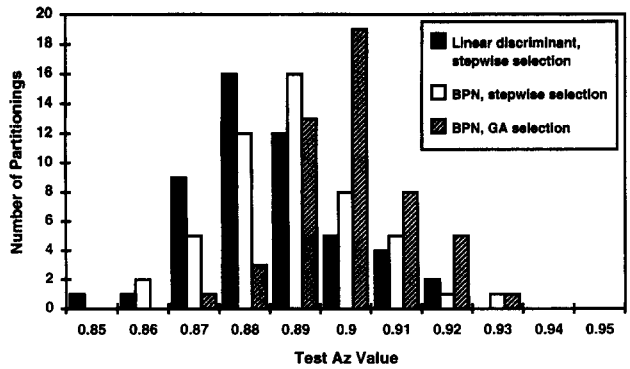


FIG. 9. The distribution of the test A_z values for the linear classifier with stepwise feature selection, BPN classifier with stepwise feature selection, and BPN classifier with GA feature selection.

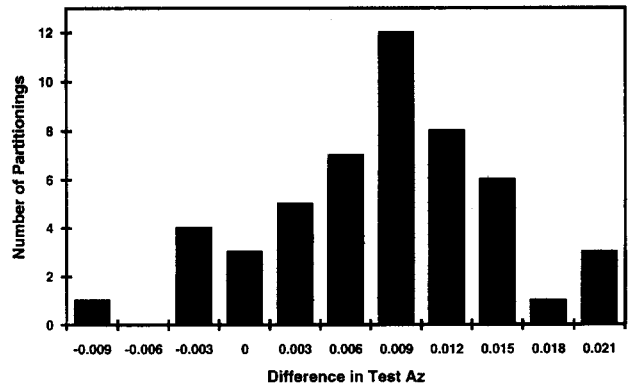


FIG. 10. The distribution of the pairwise difference of the test A_z values of the BPN classifiers with GA and stepwise feature selection.

tures in the previous subsections, the value of P_c was chosen as 0.9. A penalty term was applied to the fitness function with $\alpha = 1/2000$.

The final GA-selected pool of variables contained 16 features. After feature selection using the GA, the performance of the BPN classifier with the selected features was tested with 50 training and test groups as described above. The average training and test A_z values over 50 partitionings were 0.92 and 0.90, respectively.

To compare our GA-based feature selection method for a BPN, we also used the same data set and the 41 features described above with stepwise feature selection. The entire data set was used for feature selection. The final selected pool of variables contained 19 features. The same 50 partitionings used for the GA experiments were used to train and test both a linear discriminant classifier and a BPN with the stepwise-selected features. The average training and test A_z values over 50 partitionings were 0.92 and 0.89 with the linear classifier, and 0.92 and 0.89 with the BPN classifier. The distribution of the test A_z values for the linear classifier, as well as the BPN classifier with features selected using stepwise and the GA-based feature selection are shown in Fig. 9. The distribution of the pairwise difference of the test A_z of the BPN classifiers with stepwise and GA-based feature selection methods is shown in Fig. 10.

V. DISCUSSION

Our goal in this paper was the development of an effective feature selection algorithm given a large number of features extracted from an image data set. Table IV and Figs. 9 and 10 indicate that GA feature selection might be a viable alternative to stepwise feature selection.

The average number of features selected by stepwise and GA-based feature selection methods for a linear discriminant classifier were 19.3 and 20.1, respectively, in Table IV. In the same table, we compared these methods to random feature selection with the number of selected features equal to 20. Both methods performed better than random feature selection. The difference between the average A_z obtained by GA-based feature selection and random feature selection was more than two times the standard deviation of each A_z distribution.

We observed that each time the GA was trained with a different training set, a different set of features was selected. This was also true for stepwise feature selection. The basic reason for this was the limited size of the data set. If training sets that could represent the entire population were available, the selected set of features could be expected to be more consistent among different training sets. With the limited data set used in this study, each time a set of cases was left out as the test data, the statistical characteristics of the training feature set changed. Furthermore, many of the features were highly correlated, with correlation coefficients close to 1 or -1. Therefore, these correlated features could be interchanged. Only ten features were selected three or more times for the experiments in Table IV. Out of these ten features, six were texture and four were morphological features. This indicates that morphological and texture features are both important for the classification of the ROIs.

The high correlation between the features in the feature space used in this study is probably a cause of the surprisingly good classification result ($A_z = 0.82$) obtained with the randomly selected features. This may also indicate that many of the features in the feature space are very effective for this

TABLE IV. Test A_z values of a linear discriminant classifier using stepwise LDA, GA-based feature selection, and 20 randomly selected features.

	Stepwise LDA		GA		Random
Test group	A_z	Num. of features	A_z	Num. of features	A_z
1	0.87	19	0.90	20	0.80
2	0.91	15	0.89	24	0.86
3	0.92	25	0.93	24	0.86
4	0.88	22	0.88	20	0.81
5	0.86	23	0.84	23	0.78
6	0.92	19	0.93	20	0.83
7	0.92	15	0.91	17	0.87
8	0.84	21	0.88	19	0.75
9	0.86	14	0.88	18	0.77
10	0.88	20	0.92	16	0.82
Average	0.89	19.3	0.90	20.1	0.82

classification task. Therefore, even when only 20 features are randomly drawn, we have a high probability of drawing effective features and obtaining a classification result that is much higher than that would be obtained by chance.

Our results indicate that the classification results with GA-based feature selection are better than their counterparts with stepwise feature selection. This is most easily seen from Fig. 9, which compares the distribution of the A_z values for a BPN classifier with GA-based feature selection to that with stepwise feature selection. It can be observed that the two distributions are shifted with respect to each other, with the distribution using GA-based feature selection exhibiting higher A_z values. However, we could not perform a paired t-test to evaluate the statistical significance of the differences for the results listed in Table IV or those shown in Fig. 9. The paired t-test requires independence among the samples whereas our test (or training) sets in the different partitionings overlapped with each other. We have used the CLABROC program[41] to test the statistical significance of the difference between the corresponding pair of ROC curves for each partitioning. The difference did not achieve statistical significance for the individual pairs because the number of cases in each partitioned data set is small and thus the standard deviation of A_z is large (0.02 to 0.04). However, it should be noted that the improvement in A_z with GA-based feature selection, although small, is consistently observed over the different partitionings of the data set, over both the linear discriminant classifier (Table IV) and the BPN classifier (Figs. 9 and 10), as well as over different data sets.[42] The small improvement in A_z may be attributed to two causes: (1) For the linear discriminant classifier, the stepwise feature selection procedure is already near optimal. It is actually somewhat unexpected that the GA-based feature selection can still provide an observable improvement in A_z. (2) It is well known that BPN performance may not reach the global maximum if there are insufficient training samples. For the BPN classifier in this study, the number of weights to be trained was large compared with the number of input training samples. Therefore, it probably did not reach its optimum when it was used in a GA for feature selection. Again, a consistent improvement in A_z demonstrates that the GA can select more effective features for BPN classifiers.

The main advantage of GA-based feature selection is its flexibility. GA-based feature selection can be applied to any classifier and the fitness function can be tailored to select features with specific characteristics. An example of the former application is to select features for a nonlinear classifier such as a BPN as discussed above. An example of the latter application is to select features for development of a highly sensitive classifier[43] described next.

In both breast cancer detection and classification, the cost of missing a malignant lesion is very high. For this reason, an important measure of classification accuracy is the FPF at high true-positive classification. Since the design of the fitness function of a GA is very flexible, one can target to maximize the partial area above a specified TPF in order to optimize the classifier performance in this region. In a preliminary study with our data set,[43] we designed a GA-based feature selection algorithm in which the fitness of a chromosome was defined as the partial area above a TPF of 0.95. We then compared the FPF at TPFs of 100% and 96% using GA-based and stepwise feature selection for a linear discriminant classifier. At a TFP of 100%, the average FPF over the ten partitionings used in this study were 0.44 for GA-based feature selection, and 0.68 for stepwise feature selection. At a TFP of 96%, the average FPFs were 0.33 for GA-based feature selection, and 0.38 for stepwise feature selection. These encouraging results demonstrate the potential of a GA-based approach to designing classifiers for a wide range of practical problems, which cannot be achieved with a conventional method such as stepwise discriminant analysis.

Stepwise feature selection is computationally faster than GA-based feature selection. For example, in the present study, the stepwise feature selection required 64-s CPU time for each partition (Table IV) on a 90-MHz Pentium-based personal computer. The GA-based feature selection required 519-s CPU time for each partition (Table IV) on a 133-MHz alpha-based workstation, when the evolution involved a total of 250 chromosomes. However, a GA is highly parallelizable. In principle, the fitness of each chromosome can be evaluated on a different processor and the computation time can be improved up to a factor equal to the number of chromosomes. The choice between GA-based or stepwise feature selection will depend on the application. For a linear discriminant classifier, the stepwise feature selection may be near optimal so that the advantage of using a GA may be small. However, for other classifiers, a GA may be more effective because the selected feature set will be optimized to the specific classifier used.

A GA was previously used for the task of feature selection in a classification problem with 30 features and 150 cases.[28] The GA fitness criterion in this application was designed to be a function of the correct classification rate with a nearest-neighbor classifier. After the features were selected, a neural network was employed for final classification. Our approach has two advantages over this application. First, we used a more sophisticated classifier in the fitness function computation stage, hence GA training is more efficient. Second, we used the same classifier at the final classification stage, therefore our results are expected to be more consistent. Our results are also expected to be less biased since we divided our data set into independent training and test groups for GA evaluation, whereas the entire data set was used for training in the other study.[28]

VI. CONCLUSION

We investigated the use of a GA for feature selection, and demonstrated its application by classifying ROIs on mammograms as either containing mass or normal tissue. By comparing stepwise feature selection and GA-based feature selection for two different classifiers (the linear discriminant classifier and the BPN), and by examining the problem of designing classifiers biased to have high sensitivity performance, we have demonstrated the versatility offered by a GA

in the design of classifiers for a variety of classification tasks without a trade-off in the effectiveness of the selected features. Future work in this area includes application of GA-based feature selection to different classification tasks such as differentiation of malignant and benign tissue, and a detailed investigation of the formulation of different fitness measures, such as the partial area at the high-TPF region of the ROC curve, for the design of classifiers in different applications.

ACKNOWLEDGMENTS

This work is supported by the USPHS Grant No. CA 48129 and U.S. Army Grant No. DAMD 17-93-J-3007 (through subgrant No. GU RX 4300-803UM from Georgetown University). The content of this publication does not necessarily reflect the position of the Georgetown University or 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 providing the LABROC1 program.

[1]C. J. Vyborny, "Can computers help radiologists read mammograms?," Radiology **191**, 315–317 (1994).
[2]R. E. Bird, T. W. Wallace, and B. C. Yankaskas, "Analysis of cancers missed at screening mammography," Radiology **184**, 613–617 (1992).
[3]M. G. Wallis, M. T. Walsh, and J. R. Lee, "A review of false negative mammography in a symptomatic population," Clin. Radiol. **44**, 13–15 (1991).
[4]D. B. Kopans, "The positive predictive value of mammography," Am. J. Radiol. **158**, 521–526 (1991).
[5]D. D. Adler and M. A. Helvie, "Mammographic biopsy recommendations," Curr. Opinion Radiol. **4**, 123–129 (1992).
[6]D. Brzakovic, X. M. Luo, and P. Brzakovic, "An approach to automated detection of tumors in mammography," IEEE Trans. Med. Imag. **9**, 233–241 (1990).
[7]F. F. Yin, M. L. Giger, C. J. Vyborny, K. Doi, and R. A. Schmidt, "Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses," Invest. Radiol. **28**, 473–481 (1993).
[8]J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computerized image analysis," IEEE Trans. Med. Imag. **12**, 664–669 (1993).
[9]W. P. Kegelmeyer, J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions," Radiology **191**, 331–337 (1994).
[10]H.-P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," Phys. Med. Biol. **40**, 857–876 (1995).
[11]D. Wei, H.-P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis," Med. Phys. **22**, 1501–1513 (1995).
[12]I. E. Magnin, A. Bremond, F. Cluzeau, and C. L. Odet, "Mammographic texture analysis—An evaluation of risk for developing breast cancer," Opt. Eng. **25**, 780–784 (1986).
[13]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).
[14]V. Goldberg, A. Manduca, D. L. Evert, J. J. Gisvold, and J. F. Greenleaf, "Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence," Med. Phys. **19**, 1475–1481 (1992).
[15]Y. Chitre, A. P. Dhawan, and M. Moskowitz, "Artificial neural network based classification of mammographic microcalcifications using image structure features," Int. J. Pattern Recognition Artificial Intelligence **7**, 1377–1401 (1993).
[16]M. F. McNitt-Gray, H. K. Huang, and J. W. Sayre, "Feature selection in the pattern classification problem in digital chest radiograph segmentation," IEEE Trans. Med. Imag. **14**, 537–547 (1995).
[17]T. M. Cover and J. M. V. Campenhout, "On the possible orderings in the measurement selection problem," IEEE Trans. Syst. Man Cybern. **7**, 657–661 (1977).
[18]T. M. Cover, "The best two independent measurements are not the two best," IEEE Trans. Syst. Man Cybern. **4**, 116–117 (1974).
[19]W. S. Meisel, *Computer-Oriented Approaches to Pattern Recognition* (Academic, New York, 1972).
[20]R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (Wiley, New York, 1973).
[21]P. A. Lachenbruch, *Discriminant Analysis* (Hafner, New York, 1975).
[22]J. H. Holland, *Adaptation in Natural and Artificial Systems* (University of Michigan, Ann Arbor, 1975).
[23]D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, New York, 1989).
[24]S. Forrest, "Genetic algorithms: Principles of natural selection applied to computation," Science **261**, 872–878 (1993).
[25]D. E. Goldberg, *Computer-Aided Gas Pipeline Operation Using Genetic Algorithms and Machine Learning* (Ph.D. Dissertation in Civil Eng. University of Michigan, Ann Arbor, 1983).
[26]J. H. Holland, "Genetic algorithms," Sci. Am. **267**, 66–72 (1992).
[27]C. E. Floyd and G. D. Tourassi, "Computer-aided diagnosis using genetic algorithms and neural networks," in *Proceedings of the World Congress on Neural Networks*, Washington, DC (Lawrance Erlbaum Associates, NJ, 1995), pp. 863–866.
[28]F. Z. Brill, D. E. Brown, and W. N. Martin, "Fast genetic selection of features for neural network classifiers," IEEE Trans. Neural Networks **3**, 324–328 (1992).
[29]H. Fujita, T. Hara, X. Jing, T. Matsumoto, H. Yoshimura, and K. Seki, "Automated detection of lung nodules by using genetic algorithm technique in chest radiography," Radiology **197**, 426–426 (1995).
[30]D. Wei, H.-P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Multiresolution texture analysis for classification of mass and normal breast tissue on digital mammograms," in Proceedings of SPIE Medical Imaging: Image Process. **2434**, (San Diego, CA, 1995), pp. 606–611.
[31]B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsit, "Classification of mass and normal breast tissue: An artificial neural network with morphological features," in *Proceedings of the World Congress on Neural Networks*, Washington, DC (Lawrance Erlbaum Associates, NJ, 1995), pp. 876–879.
[32]B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsit, "Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images," IEEE Trans. Med. Imag. (in press).
[33]I. Daubechies, "Orthonormal bases of compactly supported wavelets," Commun. Pure Appl. Math. **41**, 909–996 (1988).
[34]Y. Hara, R. G. Atkins, S. H. Yueh, R. T. Shin, and J. A. Kong, "Application of neural networks to radar image classification," IEEE Trans. Geosci. Remote Sensing **32**, 100–109 (1994).
[35]N. Petrick, H.-P. Chan, B. Sahiner, D. Wei, M. A. Helvie, M. M. Goodsit, and D. D. Adler, "Automated detection of breast masses on digital mammograms using adaptive density-weighted contrast enhancement filtering," in Proc. SPIE Med. Imag. Image Process. **2434**, (San Diego, CA, 1995), 590–597.
[36]J. A. Freeman and D. M. Skapura, *Neural Networks: Algorithms, Applications and Programming Techniques* (Addison-Wesley, Reading, MA, 1991).
[37]M. M. Tatsuoka, *Multivariate Analysis, Techniques for Educational and Psychological Research* (Macmillan, New York, 1988).
[38]D. D. Dorfman and E. Alf, "Maximum likelihood estimation of parameters of signal detection theory—A direct solution," Psychometrika **33**, 117–124 (1968).
[39]C. E. Metz, J. H. Shen, and B. A. Herman, "New methods for estimating a binormal ROC curve from continuously distributed test results," presented at the 1990 Annual Meeting of the American Statistical Association, Anaheim, CA (1990).

[40]M. J. Norusis, *SPSS Professional Statistics 6.1* (SPSS Inc., Chicago, 1993).

[41]C. E. Metz, P. L. Wang, and H. B. Kronman, "A new approach for testing the significance of differences between ROC curves measured from correlated data," in *Information Processing in Medical Imaging: Proceedings of the 8th Conference*, edited by F. Deconinck (Martinus Nijhoff, Boston, 1984), pp. 432–445.

[42]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 classifiers on characterization of microcalcifications," Radiology **197**, 425–425 (1995).

[43]B. Sahiner, H.-P. Chan, N. Petrick, M. A. Helvie, D. D. Adler, and M. M. Goodsit, "Classification of malignant and benign breast masses: Development of a high-sensitivity classifier using a genetic algorithm," accepted for presentation at the 82nd Annual Meeting of the Radiological Society of N. America, Chicago, IL (1996).