Phys. Med. Biol. **43** (1998) 2853–2871. Printed in the UK PII: S0031-9155(98)88223-7

Design of a high-sensitivity classifier based on a genetic algorithm: application to computer-aided diagnosis

Berkman Sahiner†, Heang-Ping Chan, Nicholas Petrick, Mark A Helvie and Mitchell M Goodsitt

Department of Radiology, University of Michigan, Ann Arbor, USA

Received 6 October 1997

Abstract. A genetic algorithm (GA) based feature selection method was developed for the design of high-sensitivity classifiers, which were tailored to yield high sensitivity with high specificity. The fitness function of the GA was based on the receiver operating characteristic (ROC) partial area index, which is defined as the average specificity above a given sensitivity threshold. The designed GA evolved towards the selection of feature combinations which yielded high specificity in the high-sensitivity region of the ROC curve, regardless of the performance at low sensitivity. This is a desirable quality of a classifier used for breast lesion characterization, since the focus in breast lesion characterization is to diagnose correctly as many benign lesions as possible without missing malignancies. The high-sensitivity classifier, formulated as the Fisher's linear discriminant using GA-selected feature variables, was employed to classify 255 biopsy-proven mammographic masses as malignant or benign. The mammograms were digitized at a pixel size of 0.1 mm × 0.1 mm, and regions of interest (ROIs) containing the biopsied masses were extracted by an experienced radiologist. A recently developed image transformation technique, referred to as the rubber-band straightening transform, was applied to the ROIs. Texture features extracted from the spatial grey-level dependence and run-length statistics matrices of the transformed ROIs were used to distinguish malignant and benign masses. The classification accuracy of the high-sensitivity classifier was compared with that of linear discriminant analysis with stepwise feature selection ($\mathrm{LDA}_{\mathrm{sfs}}$). With proper GA training, the ROC partial area of the high-sensitivity classifier above a true-positive fraction of 0.95 was significantly larger than that of $\mathrm{LDA}_{\mathrm{sfs}}$, although the latter provided a higher total area (A_z) under the ROC curve. By setting an appropriate decision threshold, the high-sensitivity classifier and $\mathrm{LDA}_{\mathrm{sfs}}$ correctly identified 61% and 34% of the benign masses respectively without missing any malignant masses. Our results show that the choice of the feature selection technique is important in computer-aided diagnosis, and that the GA may be a useful tool for designing classifiers for lesion characterization.

1. Introduction

Due to its high sensitivity, mammography is usually the first radiological examination used for the early detection of malignant breast lesions. However, the positive predictive value (PPV) of mammographic diagnosis (ratio of the number of malignancies to the total number of biopsy recommendations) is not high. Biopsies performed for mammographically suspicious non-palpable breast masses had PPVs of 20 to 30% in three studies (Hermann *et al* 1987, Hall *et al* 1988, Jacobson and Edeiken 1990). To reduce health-care costs and patient morbidity, it is desirable to increase the PPV of mammographic diagnosis

† Address for correspondence: Department of Radiology, University of Michigan, 1500 E. Medical Center Drive, CGC B2102, Ann Arbor, MI 48109-0904, USA. E-mail address: berki@umich.edu

while maintaining its sensitivity of cancer detection. Computerized mammographic analysis methods can potentially aid radiologists in achieving this goal.

In recent years, several researchers have developed new techniques for the classification of mammographic masses based on computer-extracted features (Brzakovic *et al* 1990, Kilday *et al* 1993, Huo *et al* 1995, Pohlman *et al* 1996, Rangayyan *et al* 1996, Sahiner *et al* 1996a, 1997, 1998). Kilday *et al* (1993) classified masses using morphological features and patient age. Brzakovic *et al* (1990) classified suspected lesions using their shape and intensity variations. Huo *et al* (1995) developed a technique to quantify the degree of spiculation of a lesion, and classified masses as malignant and benign using these spiculation measures. Pohlman *et al* (1996) developed a region growing algorithm for tumour segmentation, and used features describing the tumour shape for classification. Rangayyan *et al* (1996) used an edge acutance measure extracted from the grey-scale intensity along the normal direction to the mass shape, as well as moments to classify masses. We have developed the rubber-band straightening transform (RBST) for facilitating the extraction of effective texture features, and used the texture features extracted from the transformed image for classification (Sahiner *et al* 1996a, 1997, 1998).

A common characteristic of the above approaches is that the lesion is first segmented from the surrounding tissue, and then features are extracted from the shape and grey-level characteristics of the lesion and the surrounding tissue. The extracted features usually represent a mathematical description of characteristics that are helpful for distinguishing malignant and benign lesions. When several features are extracted for classification, it may be difficult to predict which features or feature combinations will result in more accurate classification. For example, it is known that the borders of malignant masses tend to be more irregular than those of benign masses; therefore, it is expected that the normalized radial lengths (Kilday *et al* 1993) carry useful information about the probability of malignancy of a mass. However, since the normalized radial lengths, and especially the features extracted from them (for example variance and entropy), do not exactly measure irregularity but instead merge information from a combination of border characteristics, it is difficult to predict which feature combination will yield the highest classification accuracy when used in a statistical classifier. It is known that the inclusion of inappropriate features may adversely affect classifier performance, especially when the training set is not sufficiently large (Raudys and Jain 1991, Sahiner *et al* 1996c). Therefore, in many situations, one must face the task of selecting a subset of effective features for classification.

One systematic method for feature selection is linear discriminant analysis with stepwise feature selection (LDA_{sfs}), which has been applied to feature selection problems in computer-aided diagnosis (Chan *et al* 1995, Wei *et al* 1995). LDA_{sfs} is an iterative procedure, where one feature is entered into or removed from the selected feature pool at each step by analysing its effect on a selection criterion. The nature of the stepwise selection procedure makes it imperative that the selection criterion be a statistical distance measure between the two groups to be classified. The Wilks lambda and the Mahalanobis distance are commonly used measures. Genetic algorithm (GA) based feature selection, which is capable of using any numerically computed criterion for its fitness function, is a slower but more versatile method than stepwise feature selection. We have demonstrated that when the GA fitness criterion is related to the area A_z under the receiver operating characteristic (ROC) curve, GA-based feature selection yields slightly more effective features than LDA_{sfs} (Sahiner *et al* 1996c).

In the task of lesion characterization, the cost of missing a malignancy is very high. Therefore, the performance of a classifier in the high-sensitivity (high true-positive fraction) region of the ROC curve is more important than the overall area A_z under the ROC curve. In other words, if a classifier is to be designed for breast lesion characterization, the specificity at high levels of sensitivity is much more important than the specificity at low levels of sensitivity. Recently, Jiang *et al* (1996) developed a method for describing an ROC partial area index that may be useful as a performance measure in lesion characterization problems. Since a feature (or feature combination) that can provide a large overall A_z (or a large Wilks lambda and Mahalanobis distance) may not provide a large partial ROC area, it is important to develop a feature selection method for the design of high-sensitivity classifiers. The partial ROC area is potentially a good feature selection criterion for this application. The flexibility of a GA in the selection of its fitness function allows this index to be incorporated for feature selection.

In this study, we developed a methodology to design high-sensitivity classifiers. The design process was illustrated by the task of classifying masses on digitized mammograms as malignant or benign. A GA-based algorithm with the ROC partial area index as the feature selection criterion, in combination with Fisher's linear discriminant, was used for the design of this classifier. Texture features extracted from RBST images (Sahiner *et al* 1998) were used for classification. The performance of the high-sensitivity classifier was compared with the performance achieved by LDA_{sfs} using the Wilks lambda as the feature selection criterion.

2. Materials and methods

2.1. Data set

The mammograms used in this study were selected from the files of patients at the Radiology Department of the University of Michigan who had undergone biopsy. The mammograms were acquired with dedicated mammographic systems with 0.3 mm focal spots, molybdenum anodes, 0.03 mm thick molybdenum filters and 5:1 reciprocating grids. For recording the images, a Kodak MinR/MRE screen/film system with extended cycle processing was used. The criterion for inclusion of a mammogram in the data set was that the mammogram contained a biopsy-proven mass, and that approximately equal numbers of malignant and benign masses were present in the data set.

Our data set consisted of 255 mammograms from 104 patients. For most of the patients we had two mammograms in the data set, which were the craniocaudal and the mediolateral oblique views. However, for some of the patients, extra views such as lateral and oblique views were included in the data set. There were 128 mammograms with benign masses, of which 8 were spiculated based upon radiologist interpretation, and 127 mammograms with malignant masses, of which 62 were spiculated. Of the 104 patients evaluated in this study, 48 had malignant masses. The probability of malignancy of the biopsied mass on each mammogram was ranked by a Mammography Quality Standards Act (MQSA) approved radiologist experienced in mammographic interpretation on a scale of 1 to 10. A ranking of 1 corresponded to the masses with the most benign mammographic appearance, and a ranking of 10 corresponded to the masses with the most malignant mammographic appearance. The distribution of the malignancy ranking of the masses is shown in figure 1. The true pathology of the masses was determined by biopsy and histological analysis.

The mammograms in the data set were digitized with a Lumisys DIS-1000 laser scanner at a pixel resolution of 0.1 mm × 0.1 mm and 4096 grey levels. The digitizer was calibrated so that grey-level values were linearly 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, with the OD range extending to 3.5.

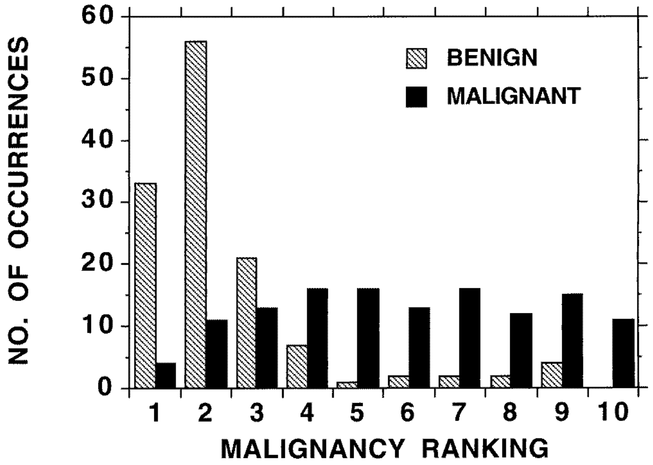


Figure 1. The distribution of the malignancy ranking of the masses in our data set, as determined by a radiologist experienced in mammographic interpretation: 1, very likely benign; 10, very likely malignant.

The pixel values were linearly converted before they were stored on the computer so that a high pixel value represented a low optical density.

The location of the biopsied mass was identified by the radiologist, and a region of interest (ROI) containing the biopsied mass was extracted for computerized analysis. The size of the ROI was allowed to vary according to the lesion size. The extracted ROIs contained a non-uniform background, which depended on the overlapping breast structures and the location of the lesion on the mammogram. The non-uniform background is not related to mass malignancy, but may affect the segmentation and feature extraction results used in our computerized analysis. To reduce the background non-uniformity, an automated background correction technique was applied to each ROI as the very first step in our analysis. Details and examples of our background correction technique can be found in the literature (Sahiner *et al* 1996b).

2.2. The rubber-band straightening transform (RBST)

In this study, the classification of malignant and benign masses was based on the textural differences of their mammographic appearance. We have previously designed a rubber-band straightening transform (RBST) which was found to facilitate the extraction of texture features from the region surrounding a mammographic mass. The image transformation performed by the RBST is depicted in figure 2, and a block diagram of different stages of the RBST is given in figure 3. A detailed discussion of the transform can be found in the literature (Sahiner *et al* 1996a, 1997, 1998). For completeness, a brief description is given below.

The RBST transforms a band of pixels surrounding a mass onto the Cartesian plane. The four basic steps in the RBST are mass segmentation, edge enumeration, computation of normals and interpolation. A modified K-means clustering algorithm (Sahiner *et al* 1995) was used for segmentation. The parameters of the segmentation algorithm were chosen so that the segmented region was slightly smaller than the actual size of the mass. After clustering, one to several objects would be segmented in the ROI. If more than one object was segmented, the largest connected object was selected. The selected object

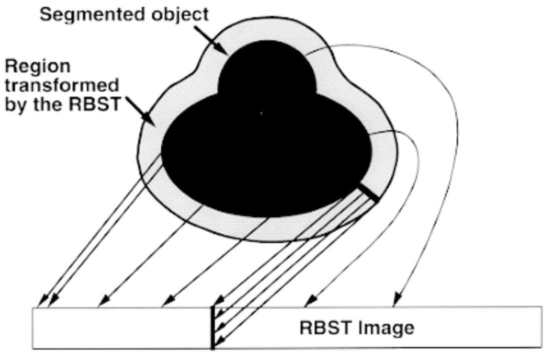


Figure 2. The formation of the RBST image.

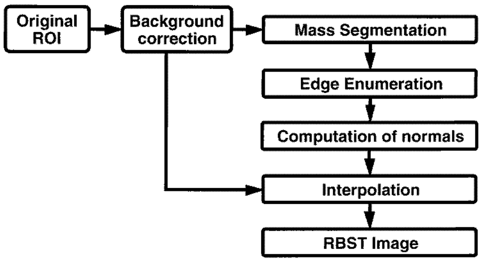


Figure 3. Block diagram of the stages of RBST image computation.

was then filled, grown in a local neighbourhood, and eroded and dilated with morphological operators. The implementation details of these steps have been described elsewhere (Sahiner *et al* 1998). After the outline of the mass was obtained, an edge enumeration algorithm assigned a pixel number to each border pixel of the mass, such that neighbouring pixels were assigned consecutive numbers. The computation of normals depended on the output of the edge enumeration algorithm. The normal $L(i)$ at border pixel i was determined as the normal to the line joining border pixels $i - K$ and $i + K$. The choice of the constant K represents a trade-off between a noisy estimate of the normal direction (small K) and an estimate that misses fine variations in the normal direction (large K). In order to determine the constant K to be used in this study, we selected a small subset of images from our database, and plotted the normal direction obtained by using different values of K superimposed on the segmented image. By performing a visual comparison of the computed normal direction to what was perceived to be the true normal direction, it was empirically found that $K = 12$ resulted in a satisfactory normal estimation. In the interpolation step, the value of the pixel in row j, column i of the RBST image was found as follows. Let $p(i, j)$ denote the location in the original image at a distance j along $L(i)$ from border pixel i. The two closest pixels in the original ROI to location $p(i, j)$ were identified, and the (i, j)th pixel value of the RBST image was defined as the distance-weighted average of these two pixel values.

The width of the band transformed by the RBST was chosen as 40 pixels in this study, which corresponded to 4 mm on the mammogram. An example of the background-corrected ROI, the segmented and morphologically filtered mass shape, and the RBST image are shown in figure 4.

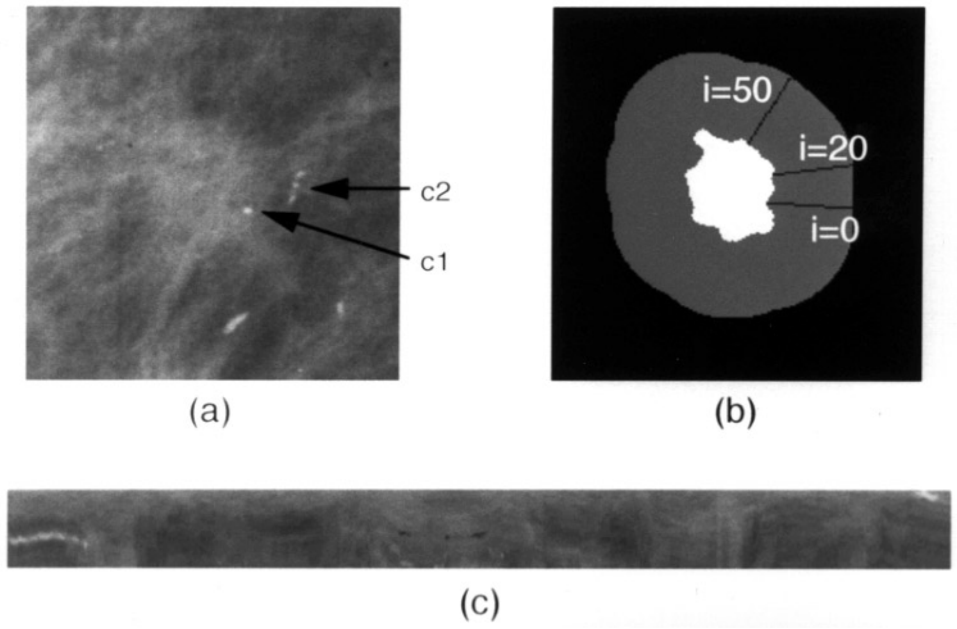


Figure 4. (a) The original mammographic ROI. (b) The segmented and morphologically filtered mass shape (white), and the 40-pixel-wide band around it (grey). For the purpose of illustration, the normals computed at $i = 0$, 20 and 50 are also shown. (c) The RBST image. Notice that due to the position of the first normal location ($i = 0$), the calcifications c1 and c2 on the original ROI appear at the right and the left of the RBST image respectively. The pathological analysis indicated that this was an invasive ductal and intraductal carcinoma.

2.3. Texture features

The texture features used for the classification of the malignant and benign masses were spatial grey-level dependence (SGLD) and run length statistics (RLS) features. These features were extracted from SGLD and RLS matrices, which were constructed from the RBST images as described below.

2.3.1. SGLD features. The (i, j)th element of the SGLD matrix $p_{\theta,d}(i, j)$ represents the probability that grey levels i and j occur at an angle θ and a distance d with respect to each other. The use of SGLD matrices for feature extraction was motivated by the assumption that texture information is contained in the average spatial relationships between the grey-level tones in the image (Haralick *et al* 1973). The features extracted from SGLD matrices of mammographic ROIs have been shown to be useful in classification of mass and normal tissue, and malignant and benign masses or microcalcifications in computer-aided diagnosis (CAD) (Chan *et al* 1995, 1997a, Wei *et al* 1995, Sahiner *et al* 1996b, 1998).

In this study, four different directions ($\theta = 0°$, 45°, 90° and 135°) and ten different pixel pair distances ($d = 1$, 2, 3, 4, 6, 8, 10, 12, 16 and 20) were used for the construction of SGLD matrices from RBST images. The total number of SGLD matrices was therefore 40. Based on our previous studies (Chan *et al* 1995), a bit depth of eight bits was used in the SGLD matrix construction.

A number of SGLD features, which describe the shape of the SGLD matrices, can be extracted from each SGLD matrix. In this study, we extracted eight such features, which were also used in our previous studies (Chan *et al* 1995, Wei *et al* 1995, Sahiner *et al* 1998). These texture features were correlation, difference entropy, energy, entropy, inertia, inverse difference moment, sum average and sum entropy. This resulted in the computation of 320 SGLD features per RBST image. These features characterize information such as homogeneity, contrast and structural linearity in the images. However, it is difficult to establish a one-to-one correspondence between these qualitative image characteristics and the extracted texture features (Haralick *et al* 1973). The definitions of the SGLD features used in this study can be found in the literature (Haralick *et al* 1973, Chan *et al* 1995, Wei *et al* 1995).

2.3.2. RLS features. The pixels along a given line in an image occasionally contain runs of consecutive pixels that all have the same grey level. A grey-level run is defined as a set of consecutive, collinear pixels in a given direction which have the same grey-level value. A run length is the number of pixels in a grey-level run. The RLS matrix for a given image describes the run length statistics in a given direction for each grey-level value in the image. The (i, j)th element of the RLS matrix $r\theta(i, j)$ represents the number of times that runs of length j in the direction θ consisting of pixels with a grey level i exist in the image (Weszka *et al* 1976).

The RLS matrices in this study were extracted from the vertical and horizontal gradient magnitudes of the RBST images. The vertical and horizontal gradients were obtained by filtering the RBST images with horizontally and vertically oriented Sobel filters (Jain 1989) respectively. Examples of the gradient magnitude images are shown in figure 5. The RLS matrices were obtained from each gradient magnitude image in two directions, $\theta = 0^\circ$ and $\theta = 90^\circ$. Therefore, a total of four RLS matrices were obtained for each RBST image.

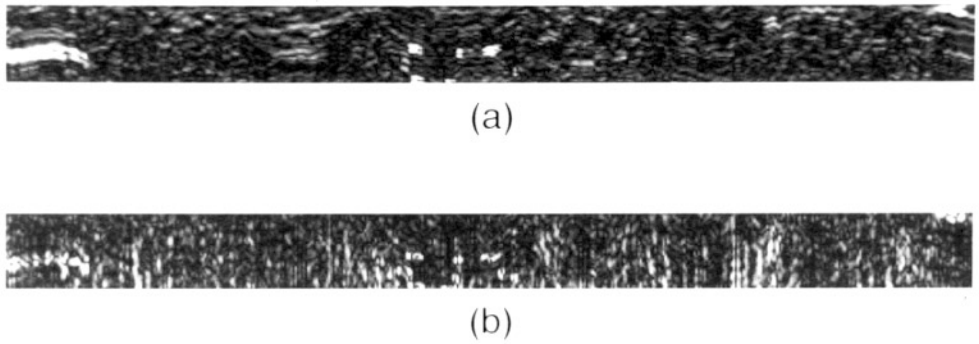


Figure 5. Gradient magnitude images for the RBST image in figure 4: (a) horizontal gradient magnitude image and (b) vertical gradient magnitude image.

Based on our previous study, a bit depth of 5 was used for the computation of RLS matrices (Sahiner *et al* 1998). Five RLS features, namely short runs emphasis, long runs emphasis, grey-level non-uniformity, run length non-uniformity and run percentage were extracted from each RLS matrix. This resulted in the computation of 20 RLS features per RBST image. The definitions of these features can be found in the literature (Galloway 1975). It is possible to describe the general aspects of the relationship between the image characteristics and the RLS feature values. For example, run percentage is low for images with long linear structures, and grey-level non-uniformity is low for images where runs are equally distributed throughout the grey levels (Galloway 1975). However, it is again difficult to establish a one-to-one correspondence between these texture features and visual image features.

2.4. Fisher's linear discriminant and LDA_{sfs}

For a two-class problem, Fisher's linear discriminant projects the multidimensional feature space onto the real line in such a way that the ratio of between-class sum of squares to within-class sum of squares is maximized after the projection (Duda and Hart 1973). This is the optimal classifier if the features for the two classes have a multivariate Gaussian distribution with equal covariance matrices (Lachenbruch 1975). It has been shown to be a reasonably good classifier even when the feature distributions for the two classes are non-Gaussian (Duda and Hart 1973). Linear discriminant analysis (LDA) is a class of statistical techniques based on Fisher's linear discriminant.

When the training data size is limited, the inclusion of inappropriate features in a classifier may reduce the test accuracy due to overtraining. Therefore, when a large number of features are available for a classification task, it is necessary to select a subset of the most effective features from the feature pool. $\mathrm{LDA}_{\mathrm{sfs}}$ is a commonly used feature selection method (Lachenbruch 1975). In this study, the performance of a GA-based high-sensitivity feature selection method was compared with that of stepwise feature selection.

Wilks' lambda, which is defined as the ratio of within-group sum of squares to the total sum of squares (Lachenbruch 1975), was used as the selection criterion for the stepwise feature selection method. The stepwise feature selection algorithm starts with no selected features at step 0. At step s of the algorithm, the available features are entered into the selected feature pool one at a time during feature entry, and those already selected are removed one at a time during feature removal. The significance of the change in the Wilks' lambda, as determined by F-statistics, when a new feature is entered into the selected feature pool is compared with a threshold F_{in}. The feature with the highest significance is entered to the selected feature pool only if the significance is higher than F_{in}. Likewise, the significance of the change in the Wilks' lambda when a selected feature is removed from the feature pool is compared with a threshold F_{out}. The feature with the least significance is removed from the selected feature pool only if the significance is lower than F_{out}. This completes step s of the algorithm. The algorithm terminates when no more features can satisfy the criteria for either being added to or removed from the selected feature pool.

2.5. Genetic algorithms for feature selection

Genetic algorithms solve optimization problems by mimicking the natural selection process. A GA follows the evolution of a population of chromosomes which are encoded so that each chromosome corresponds to a possible solution of the optimization problem. The chromosomes consist of genes, which are components of the solution. The goal of a GA is to search for better combinations of the genes, i.e. new chromosomes which are better solutions to the optimization problem. This goal is achieved by evolution. A new generation of chromosomes is produced from the current population by means of parent selection, crossover and mutation. The probability that a chromosome is selected as a parent is related to its ability to solve the optimization problem, i.e. its fitness. Chromosomes which are better solutions to the optimization problem are given a higher chance to reproduce than those which are worse solutions to the problem, similar to the principle of natural selection. The fitness of a chromosome is computed using a fitness function, which is designed on the basis of the optimization criterion for the problem. The probability that a chromosome is selected as a parent is equal to its normalized fitness, which is defined as the fitness of the chromosome divided by the sum of fitnesses for all chromosomes. The chromosomes of the selected parents are allowed to randomly cross over and mutate, introducing new genes and new chromosomes into the population. This process generates a new population of chromosomes, which tends to evolve towards a better solution.

GAs had been applied to the problem of feature selection (Brill *et al* 1992, Sahiner *et al* 1996c). The most natural way of encoding a chromosome for this problem is as follows (Sahiner *et al* 1996c). Each gene in a chromosome is a bit, which takes a value of either 1 or 0. Each gene location in a chromosome corresponds to a particular feature. If the bit value at a gene location is 1, the corresponding feature is selected for the solution of the classification problem. Otherwise, the corresponding feature is not selected. Each chromosome thus defines a set of selected features. A statistical classifier, such as Fisher's linear classifier or a neural network classifier, is then employed for classification based on the selected feature set. The fitness function reflects the success of the selected feature set for solving the classification problem. The design of the fitness function for a high-sensitivity classifier is described in the next section. The GA training method and the choice of GA parameters are summarized next.

2.5.1. GA training. The GA in this study was trained using a leave-one-case-out paradigm. In this paradigm, all ROIs except those from a particular patient were defined as the training set, and the ROIs from that particular patient were defined as the test set. For each chromosome of the GA, the coefficients of Fisher's linear discriminant function were determined using the features of the training set. The trained discriminant function was then used to classify the test cases using the features of the test cases as the input. In a given generation of the GA, all patients were visited in a round-robin manner, so that test scores were obtained for each ROI in the entire data set. The fitness of the chromosome was computed based on the classification accuracy for the test cases, as described in the next section.

2.5.2. GA parameters. The fundamental parameters of a GA are the number of chromosomes, the chromosome length, the crossover rate, the mutation rate and the stopping criterion. In a GA, the population must contain a large number of chromosomes to provide the variability that offers the opportunity to evolve towards the optimal solution. This requirement and computing speed considerations are trade-offs for selecting the number of chromosomes in a given application. The length of a chromosome is determined by the encoding mechanism which translates the optimization problem into a GA. With the encoding mechanism described earlier in this subsection, the length of each chromosome is equal to the total number of features. The fitness function is the most important component of the GA, and its design is described in the next section. Pairs of chromosomes are probabilistically selected as parents based on their fitness. A selected pair may exchange genes to generate two offspring. The crossover rate determines the probability that parents will exchange genes. After crossover, the binary value of each bit may probabilistically be altered (from 1 to 0, or vice versa), i.e. mutated. The mutation rate determines the probability that genes will undergo mutation. The increase in the fitness of the chromosomes starts to stagnate after a number of generations. The stopping criterion determines when the evolution is terminated. In this study, the GA evolution was terminated after a fixed number of iterations. The appropriateness of this stopping criterion is discussed in section 4. After the termination, the chromosome with the highest fitness value provided the set of selected features.

Table 1 shows the values of each of these parameters, selected based on our previous work. More detailed discussion of these operators and parameters can be found in the literature (Sahiner *et al* 1996c).

Table 1. GA parameters used in this study.

Crossover rate	0.9
Mutation rate	0.0025
Chromosome length	340
Number of chromosomes	200
Stopping criterion	200 iterations

2.6. Design of a high-sensitivity classifier

A widely accepted method for comparing the performance of two classifiers is to consider their ROC curves. The area A_z under the ROC curve is a commonly used index for this comparison. However, for applications where the performance at high sensitivity (or high true-positive fraction) is important, for example breast lesion characterization in CAD, this index may be inadequate. Jiang *et al* (1996) explored this issue, and defined an ROC partial area index that will be denoted as A_{TPF_0} in this paper.

The partial area index A_{TPF_0} summarizes the average specificity above a sensitivity of TPF_0 (figure 6), and can be expressed as (Jiang *et al* 1996)

$$A_{\mathrm{TPF}_0} = 1 - \frac{1}{1-\mathrm{TPF}_0}\int_{\mathrm{TPF}_0}^{1} \mathrm{FPF}(\mathrm{TPF})\,\mathrm{d}(\mathrm{TPF}) \qquad (1)$$

which is the ratio of the partial area under the actual ROC curve to the partial area of the perfect ROC curve. The maximum value for A_{TPF_0} is thus 1. The A_{TPF_0} value for a classifier that operates purely on random guessing is $(1-\mathrm{TPF}_0)/2$, which is the area under the chance diagonal normalized to $1-\mathrm{TPF}_0$.

When the conventional binormal model is employed for the computation of the ROC curve, the curve is completely defined by two parameters, a and b, which are determined from the rating data using maximum likelihood estimation. The constant b represents the estimated standard deviation of the actually negative cases, normalized by the estimated standard deviation of the actually positive cases, and the constant a represents the estimated difference between the means of actually positive and negative cases, normalized again by the estimated standard deviation of the actually positive cases. Using the binormality assumption, the partial area index A_{TPF_0} can be expressed as (McClish 1989, Jiang *et al* 1996)

$$A_{\mathrm{TPF}_0} = 1 - \frac{1}{1-\mathrm{TPF}_0}\int_{c_0}^{\infty} \Phi\left(\frac{u-a}{b}\right)\phi(u)\,\mathrm{d}u \qquad (2)$$

where

$$\phi(u) = \frac{1}{\sqrt{2\pi}}\exp(-u^2/2)$$

and

$$\Phi(u) = \int_{-\infty}^{u} \phi(x)\,\mathrm{d}x.$$

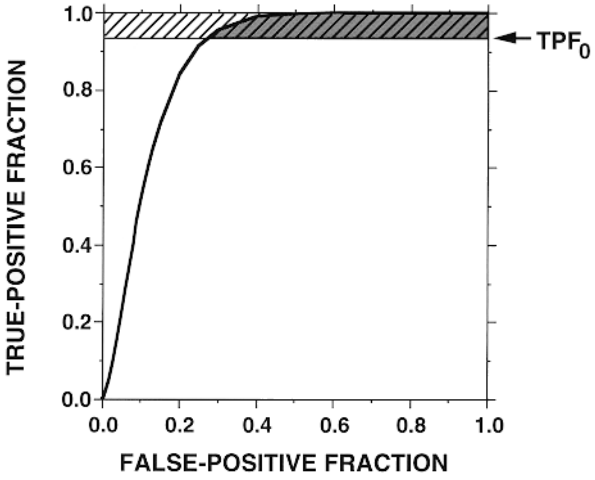


Figure 6. The partial area index A_{TPF_0} is defined as the ratio of the partial area under the ROC curve above a given sensitivity (grey area) to the partial area of the perfect ROC curve (hatched region) above the same sensitivity.

Our goal in this study was to train a GA to select features which would yield high specificity in the high-sensitivity region of the ROC curve. Therefore, the fitness of a chromosome was defined as a monotonic function of A_{TPF_0}, such that the maximization of A_{TPF_0} would maximize the fitness function

$$\text{fitness} = \left(\frac{A_{\mathrm{TPF}_0} - A_{\min}}{A_{\max} - A_{\min}} \right)^n \qquad (3)$$

where $A_{\max}$ and $A_{\min}$ were the maximum and minimum values of A_{TPF_0} among all chromosomes in a generation, and n was a power parameter whose effect on GA feature selection was investigated, as discussed in section 3. From equation (3), it is seen that as the power parameter becomes larger the difference in the fitness, and thus the probability of being chosen as parents, between the chromosomes are more amplified. The choice of n is a tradeoff between the goal of promoting chromosomes with high fitness values and the need to retain segments of good genes in other chromosomes.

For a given chromosome, the parameters a and b that are required for the computation of A_{TPF_0} were determined from the distribution of test scores using the LABROC program of Metz *et al* (1998). The partial area index A_{TPF_0} was then computed by numerically integrating equation (2). The classifiers thus designed will be referred to as GA-based high-sensitivity classifiers in the following discussions.

In this study, the significance of the difference in A_{TPF_0} of different classifiers was determined using a recently developed statistical test (Jiang *et al* 1996). The test is analogous to statistical tests involving the area A_z under the entire ROC curve, and is implemented using the covariance estimates of a and b values for the two curves.

3. Results

To demonstrate the training of high-sensitivity classifiers using GA, we chose two levels of sensitivity thresholds, $\mathrm{TPF}_0 = 0.50$ and $\mathrm{TPF}_0 = 0.95$ in equation (1). The classification results of these classifiers were compared with those of $\mathrm{LDA}_{\mathrm{sfs}}$. GA-based feature selection

Table 2. The number of features, the area A_z under the ROC curve, the partial area above the true positive fraction of 0.5 ($A_{0.50}$), and that above 0.95 ($A_{0.95}$) for various values of F_{in} and F_{out} in the stepwise feature selection method.

F_{in}	F_{out}	Number of selected features	A_z	$A_{0.50}$	$A_{0.95}$
3.8	2.7	9	0.84	0.71	0.22
2.6	2.4	13	0.85	0.72	0.27
2.2	2.0	14	0.86	0.73	0.25
1.8	1.6	26	0.89	0.80	0.38
1.4	1.2	41	0.92	0.83	0.47
1.0	1.0	49	0.92	0.83	0.46

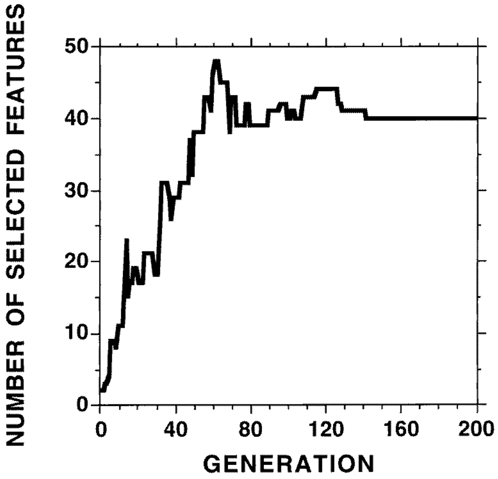


Figure 7. The evolution of the number of selected features for a GA training session ($n = 4$, $\text{TPF}_0 = 0.95$).

was also performed with no emphasis on high sensitivity ($\text{TPF}_0 = 0$). The classifier designed with the features thus selected will be referred to as an ordinary GA-based classifier. Its performance was compared with those of the GA-based high-sensitivity classifiers and LDA_{sfs}.

In LDA_{sfs}, the optimal values of the F_{in} and F_{out} thresholds are not known *a priori*. We therefore varied these thresholds to obtain the feature subset with the best test performance. Table 2 shows the number of selected features, the area A_z under the ROC curve, the partial area above the true positive fraction of 0.5 ($A_{0.50}$), and that above 0.95 ($A_{0.95}$) as these F thresholds are varied. By comparing the A_z values and the performance at the high-sensitivity portion of the ROC curve, the combination $F_{\text{in}} = 1.4$, $F_{\text{out}} = 1.2$ was found to provide the best feature subset.

High-sensitivity classifiers with $\text{TPF}_0 = 0.50$ and $\text{TPF}_0 = 0.95$ were trained with three different values of the power parameter, n ($n = 1$, 2 and 4). Figure 7 shows the evolution of the number of selected features, and figure 8 shows the total area under the ROC curve (A_z) and the partial area above the true positive fraction of 0.95 ($A_{0.95}$) for a typical GA training ($n = 4$, $\text{TPF}_0 = 0.95$).

The ROC curve of the best LDA_{sfs} classifier and those of GA-based classifiers ($\text{TPF}_0 = 0.50$ and $\text{TPF}_0 = 0.95$) with $n = 1$, 2 and 4 are compared in figures 9–11

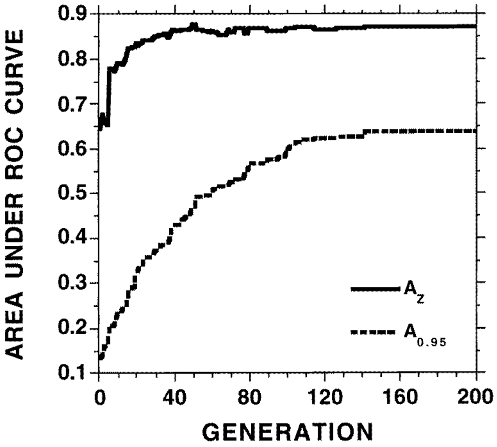
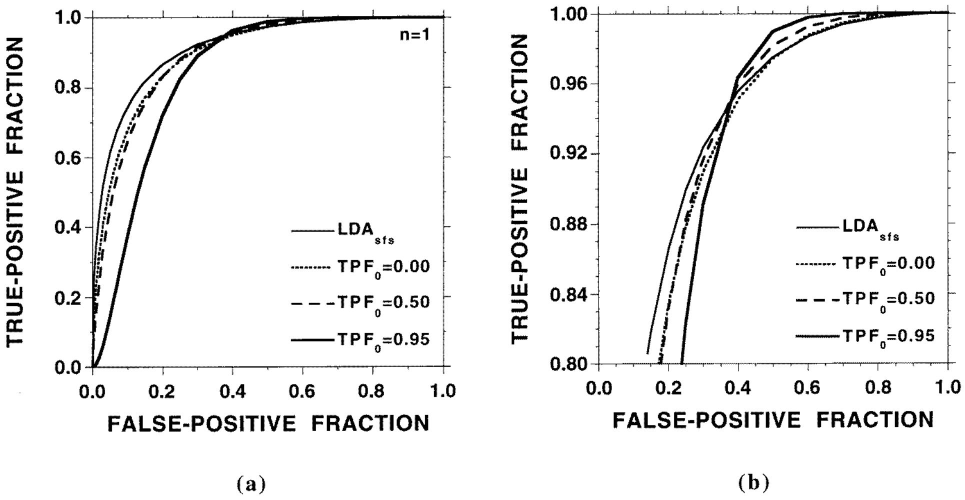


Figure 8. The evolution of the area A_z and the partial area $A_{0.95}$ under the ROC curve for the GA training session of figure 7 ($n = 4$, $\mathrm{TPF}_0 = 0.95$).

respectively. It is observed from figures 10 and 11 that for $n = 2$ or 4, the designed high-sensitivity classifiers seem to be superior to the best $\mathrm{LDA}_{\mathrm{sfs}}$ classifier for large values of true positives. When $n = 1$, the ROC curves of the GA-based high-sensitivity classifiers are still higher than that of the $\mathrm{LDA}_{\mathrm{sfs}}$ classifier when TPF is very close to 1; however, the difference between the curves is small. To quantify the improvement obtained by the GA-based high-sensitivity classifier, we performed statistical significance tests (Jiang *et al* 1996) on the partial area above a true-positive threshold of 0.95 ($A_{0.95}$) as described in the previous section. With $n = 4$, the difference between the partial areas of the GA-based high-sensitivity classifiers and $\mathrm{LDA}_{\mathrm{sfs}}$ above a true-positive threshold of 0.95 was statistically significant with two-tailed p-levels of 0.006 and 0.02 for the classifiers trained with $\mathrm{TPF}_0 = 0.95$ and $\mathrm{TPF}_0 = 0.5$ respectively. For $n = 2$, the corresponding p-levels were 0.01 and 0.07 respectively. For $n = 1$, the difference did not achieve statistical significance ($p = 0.14$ for $\mathrm{TPF}_0 = 0.95$ and $p = 0.49$ for $\mathrm{TPF}_0 = 0.5$). The difference of the partial area index over a true-positive threshold of 0.5 ($A_{0.50}$) did not achieve statistical significance when the high-sensitivity classifiers trained with $\mathrm{TPF}_0 = 0.5$ were compared with $\mathrm{LDA}_{\mathrm{sfs}}$ for any of the power parameters studied ($n = 1$, 2 and 4).

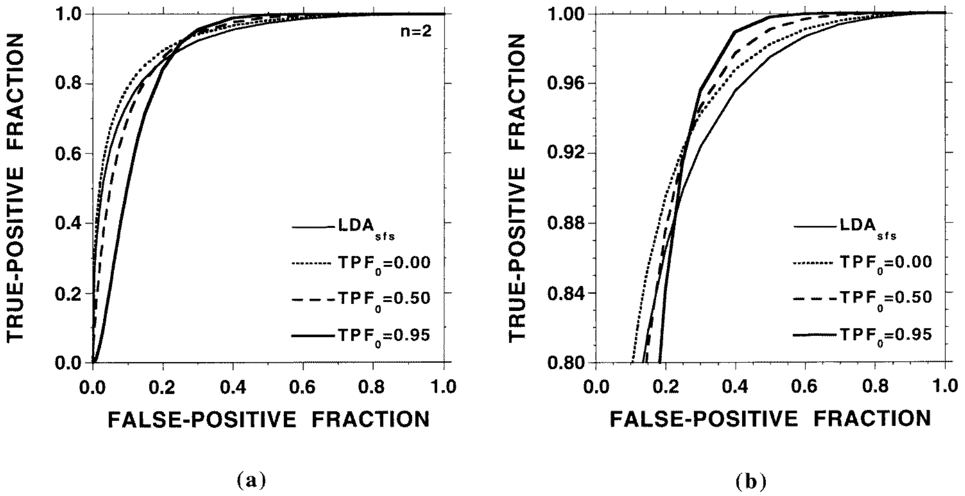
The performance of the high-sensitivity classifiers and the ordinary GA-based classifiers ($\mathrm{TPF}_0 = 0$) are also compared in figures 9–11. It is observed that the difference between the high-sensitivity and the ordinary GA-based classifiers is less than the difference between the high-sensitivity classifiers and the $\mathrm{LDA}_{\mathrm{sfs}}$. With a two-tailed significance test, it was found that the difference between the partial areas of the high-sensitivity and the ordinary GA-based classifiers above a true-positive threshold of 0.95 ($A_{0.95}$) did not achieve statistical significance for any of the power parameter values studied ($n = 1$, 2 and 4) with p-levels ranging between 0.06 and 0.5. Similarly, the difference between the ordinary GA-based classifiers and $\mathrm{LDA}_{\mathrm{sfs}}$ did not achieve statistical significance for any of the power parameter values studied. Table 3 summarizes the A_z, $A_{0.50}$ and $A_{0.95}$ values, as well as the number of features selected by each classifier.

Figures 12 and 13 show the distributions of the classifier outputs for the high-sensitivity classifier ($n = 4$, $\mathrm{TPF}_0 = 0.95$) and the $\mathrm{LDA}_{\mathrm{sfs}}$ respectively. Using the $\mathrm{LDA}_{\mathrm{sfs}}$, the distribution of the malignant masses has a relatively long tail that overlaps with the distribution of the benign masses. With the high-sensitivity classifier, this tail seems to



(a) (b)

Figure 9. The ROC curves of the LDA$_{sfs}$, the ordinary GA-based classifier (TPF$_0$ = 0), and the GA-based high-sensitivity classifiers trained with TPF$_0$ = 0.50 and TPF$_0$ = 0.95 using power parameter $n = 1$: (a) the entire ROC curves, (b) enlargement of the curves for TPF > 0.8.



(a) (b)

Figure 10. The ROC curves of the LDA$_{sfs}$, the ordinary GA-based classifier (TPF$_0$ = 0), and the GA-based high-sensitivity classifiers trained with TPF$_0$ = 0.50 and TPF$_0$ = 0.95 using power parameter $n = 2$: (a) the entire ROC curves, (b) enlargement of the curves for TPF > 0.8.

be shortened, so that more benign masses may be correctly diagnosed without missing malignancies. At 100% sensitivity, the specificity with the appropriate choice of the decision threshold was 61% and 34% for the high-sensitivity classifier and the LDA$_{sfs}$ respectively.

4. Discussion

Figures 10 and 11 demonstrate that when the feature selection is performed with a properly designed fitness function in the GA, the designed classifier can be more effective than

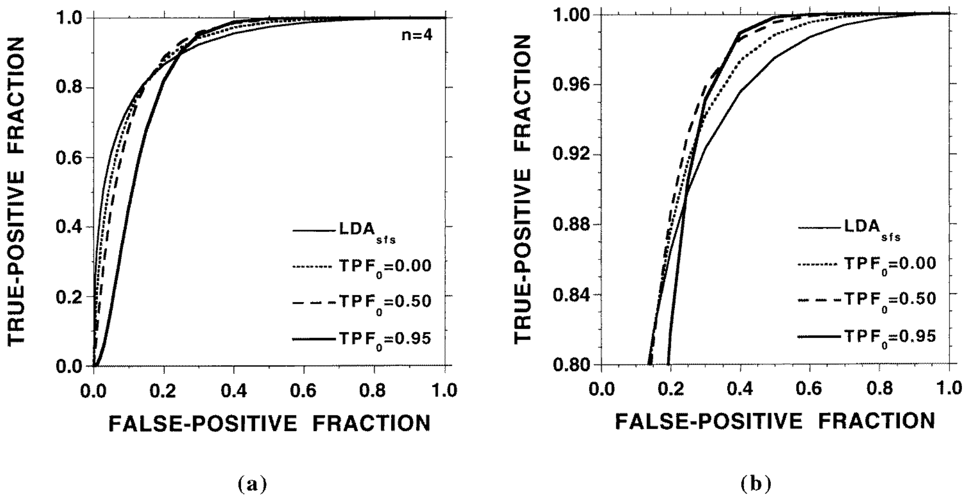


Figure 11. The ROC curves of the LDA$_{\rm sfs}$, the ordinary GA-based classifier (TPF$_0 = 0$), and the GA-based high-sensitivity classifiers trained with TPF$_0 = 0.50$ and TPF$_0 = 0.95$ using power parameter $n = 4$: (a) the entire ROC curves, (b) enlargement of the curves for TPF > 0.8.

Table 3. The number of features, the area A_z under the ROC curve, the partial area above the true positive fraction of 0.5 ($A_{0.50}$), and that above 0.95 ($A_{0.95}$) for the GA parameters studied. For comparison purposes, the results with linear discriminant analysis are also included as the last row.

Power Parameter, n	TPF$_0$ value for GA training	Number of selected features	A_z	$A_{0.50}$	$A_{0.95}$
1	0	62	0.90 ± 0.02	0.81 ± 0.03	0.47 ± 0.07
1	0.5	61	0.89 ± 0.02	0.81 ± 0.03	0.51 ± 0.07
1	0.95	58	0.84 ± 0.02	0.76 ± 0.03	0.55 ± 0.05
2	0	60	0.93 ± 0.02	0.86 ± 0.03	0.51 ± 0.08
2	0.5	48	0.91 ± 0.02	0.85 ± 0.03	0.58 ± 0.07
2	0.95	50	0.88 ± 0.02	0.82 ± 0.03	0.63 ± 0.05
4	0	40	0.92 ± 0.02	0.85 ± 0.03	0.56 ± 0.07
4	0.5	39	0.91 ± 0.02	0.85 ± 0.03	0.62 ± 0.06
4	0.95	40	0.87 ± 0.02	0.81 ± 0.03	0.64 ± 0.05
Linear discriminant analysis		41	0.92 ± 0.02	0.83 ± 0.03	0.47 ± 0.07

LDA$_{\rm sfs}$ in the high-sensitivity region of the ROC curve. From table 3 it is observed that although the A_z value for the properly trained high-sensitivity classifier (e.g. TPF$_0 = 0.5$ or 0.95 and $n = 2$ or 4) may be less than that of the LDA$_{\rm sfs}$, the partial area index $A_{0.95}$ is larger. The statistical analysis in this study showed that the difference between the properly designed high-sensitivity classifiers and the LDA$_{\rm sfs}$ at the high-sensitivity region of the ROC curve can be significant.

Comparing figure 9 with figures 10 and 11, it is observed that the selection of the power parameter n in GA training may be important. The classifiers designed with $n = 1$ did not exhibit a major advantage over the LDA$_{\rm sfs}$, as also seen from table 3 and the statistical significance tests. From equation (3), it is seen that as the power parameter becomes larger, the difference in the fitness, and thus the probability of being chosen as

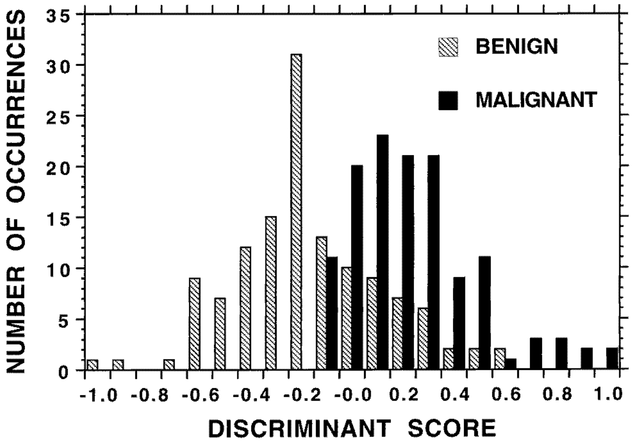


Figure 12. The distribution of the classifier output for the high-sensitivity classifier with $n = 4$, $\mathrm{TPF}_0 = 0.95$. By setting an appropriate threshold on these classifier scores, 61% of masses could correctly be classified as benign without missing any malignancies in this study.

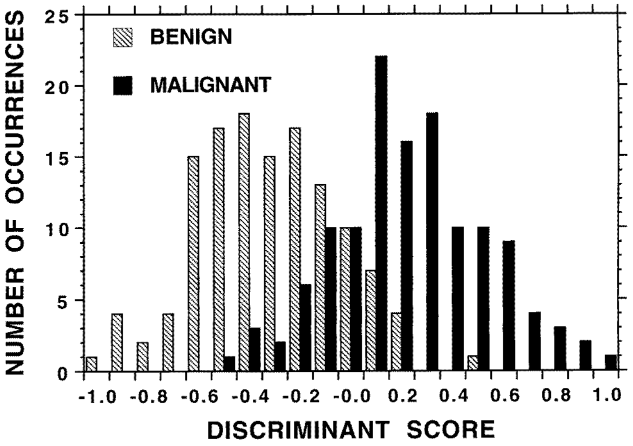


Figure 13. The distribution of the classifier output for $\mathrm{LDA}_{\mathrm{sfs}}$. By setting an appropriate threshold on these classifier scores, 34% of masses could be correctly classified as benign without missing any malignancies in this study.

parents, between the chromosomes are more amplified. Therefore, a larger value of n favours the reproduction of better chromosomes in a generation. Although it is desirable to favour the better chromosomes in any GA algorithm, too much emphasis on better chromosomes might suppress the chance of retaining segments of good genes in other chromosomes in the gene pool. This is best seen by letting n tend to infinity, and observing that only the best single chromosome will reproduce in this case, which reduces the GA to a random search algorithm. In our application, from table 3, it is observed that, for all three sensitivity thresholds ($\mathrm{TPF}_0 = 0.95$, 0.50 and 0), the classifier trained with $n = 1$ has lower performance indices ($A_{0.95}$, $A_{0.50}$ and A_z) than its counterpart trained with $n = 2$ or $n = 4$. Although none of these differences reached statistical significance, the consistently poorer performance of the classifiers trained with $n = 1$ indicates that $n = 1$ may not be a good choice for GA training.

From figures 7 and 8 it is observed that the best fitness and the number of chromosomes did not change between iterations 140 and 200 for the high-sensitivity classifier with $n = 4$ and $\text{TPF}_0 = 0.95$. A similar trend was observed with the other values of n and TPF_0 investigated in this study. Therefore, 200 generations seems to be sufficient for the GA to complete its evolution in this application. In figure 8, the best A_z value was attained around the fiftieth generation, and the A_z value did not change considerably afterwards. However, the $A_{0.95}$ value increased until around 140 generations. This meant that the classification accuracy at high sensitivity continued to increase although the A_z value did not change, i.e. the shape of the ROC curve changed so that the specificity at the high-sensitivity region of the ROC curve increased, while the specificity at the low-sensitivity region of the ROC curve decreased.

Figures 9–11 and the statistical significance tests in section 3 show that although the GA-based high-sensitivity classifiers perform better than the ordinary GA-based classifiers at high sensitivity, the difference between the two classifiers is not statistically significant. Comparison of the LDA_{sfs} and the ordinary GA-based classifiers revealed that neither the difference between the A_z values, nor the difference between the $A_{0.95}$ values were statistically significant ($p > 0.3$). However, the difference between the $A_{0.95}$ values of the LDA_{sfs} and the GA-based high-sensitivity classifiers trained with power parameter $n = 2$ and $n = 4$ was statistically significant (two-tailed p-level <0.05), as described in section 3. Thus, it was necessary to use a high-sensitivity classifier in order to obtain statistically significant improvement over the LDA_{sfs}.

The GA-based high-sensitivity classifiers ($\text{TPF}_0 = 0.95$ and $\text{TPF}_0 = 0.5$) and the ordinary GA-based classifier ($\text{TPF}_0 = 0$) were designed to maximize the partial ROC areas above the chosen true-positive fraction thresholds. From table 3, it is observed that this goal is achieved for the GA-based classifiers with TPF_0 values of 0 and 0.95. For each n, the GA-based classifier with $\text{TPF}_0 = 0$ (ordinary GA-based classifier) yielded the highest A_z value, and the GA-based classifier with $\text{TPF}_0 = 0.95$ yielded the highest $A_{0.95}$ value among the classifiers. For the classifier with $\text{TPF}_0 = 0.5$, the $A_{0.50}$ value was larger than or equal to that of the other GA-based classifiers for $n = 1$ and $n = 4$. However, for $n = 2$, the ordinary GA-based classifier ($\text{TPF}_0 = 0$) had the highest $A_{0.50}$ value, although the difference was not statistically significant ($p > 0.3$). This result is not inconsistent with the GA principles or operation. Since the GA training is based on stochastic search, the GA tends to evolve towards the optimal solution, as evidenced by the comparison of the GA-based classifiers in table 3. However, the optimality of the solution is not guaranteed, and one may encounter situations that the design goal was not totally achieved, as evidenced by the fact that the ordinary GA-based classifier had the highest $A_{0.50}$ value for $n = 2$.

Given the probabilistic nature of GA-based feature selection, it is difficult to predict the conditions under which the GA may select a feature set that provides a better high-sensitivity classifier than LDA_{sfs}. Both our GA-based method and the stepwise feature selection algorithm were designed primarily to select features for classifying classes that have multivariate Gaussian distributions and equal covariance matrices. When these assumptions are not satisfied, the accuracy of feature selection will deteriorate to a different degree for both methods. One possible explanation for the relative success of the GA-based feature selection might be that our data violate the assumptions of multivariate normality and the equality of covariance matrices, and that the GA-based method is less sensitive to these violations.

In this study, our focus was to develop a methodology for the design of high-sensitivity classifiers for applications in CAD. For the specific application of discriminating malignant and benign breast lesions, our data set was limited and the features selected by the GA may not be the optimal set of features for the general population. The same is true for the $\mathrm{LDA}_{\mathrm{sfs}}$. Considering that the data set contained only 255 masses, the number of features selected both by the GA and the $\mathrm{LDA}_{\mathrm{sfs}}$ was large. As a result, if a classifier trained in this study is applied without modification to the population at large, the classification accuracy is likely to be poorer than that obtained in this paper. However, the methodology developed in this study is general. When a sufficiently large data set becomes available, the GA-based high-sensitivity feature selection algorithm can be reapplied, and a more robust feature set can be determined. The number of training cases required for generalizable classifier design and feature selection has been the subject of recent studies (Raudys and Jain 1991, Wagner *et al* 1997, Chan *et al* 1997b), and is currently under investigation.

An important consideration concerning the use of GAs for optimization is the speed of computation. Depending on the number of final features selected, the GA-based feature selection implemented in this study (340 features, 200 chromosomes, 200 generations and leave-one-case-out GA training) took between 24 and 60 h on an AlphaStation 500 (400 Mhz Alpha chip), whereas the stepwise feature selection performed on a PC compatible computer with a 90 MHz Pentium processor took less than 10 min. Therefore, GA-based feature selection implemented in this study may not be practical for studies where the feature selection has to be performed many times. The high-sensitivity classifier design method developed in this study may be more appropriate if the speed of computation is of secondary importance to the classification accuracy of the designed classifier. For example, the GA-based high-sensitivity classifier can be trained only once when a final set of features is desired for a large data set as discussed above.

5. Conclusion

We have developed a GA-based method to design a high-sensitivity classifier for CAD applications. The usefulness of the method was demonstrated by the problem of classifying masses on digitized mammograms. Texture features extracted from RBST images were used to distinguish malignant and benign masses. The accuracy of the high-sensitivity classifier was shown to be significantly higher than that of $\mathrm{LDA}_{\mathrm{sfs}}$ above a true-positive fraction of 0.95. By using an appropriate decision threshold on the high-sensitivity classifier scores, 61% of the benign masses could correctly be identified without missing any malignant masses. The GA may therefore be a useful tool in the design of high-sensitivity classifiers for different classification problems in CAD or other applications.

Acknowledgments

This work is supported by a USPHS grant CA 48129, a Career Development Award (BS) from the USAMRMC (DAMD 17-96-1-6012), and a USAMRMC grant DAMD 17-96-1-6254. No official endorsement of any equipment or product of any companies mentioned in the publication should be inferred. The authors are grateful to Charles E Metz, PhD, for providing the LABROC and CLABROC programs.

References

Brill F, Brown D and Martin W 1992 Fast genetic selection of features for neural network classifiers *IEEE Trans. Neural Networks* **3** 324–8

Brzakovic D, Luo X M and Brzakovic P 1990 An approach to automated detection of tumors in mammograms *IEEE Trans. Med. Imaging* **9** 233–41

Chan H-P, Sahiner B, Petrick N, Helvie M A, Lam K L, Adler D D and Goodsitt M M 1997a Computerized classification of malignant and benign microcalcifications on mammograms: texture analysis using an artificial neural network *Phys. Med. Biol.* **42** 549–67

Chan H-P, Sahiner B, Wagner R F, Petrick N and Mossoba J 1997b Effects of sample size on classifier design: quadratic and neural network classifiers *Proc. SPIE* **3034** 1102–13

Chan H-P, Wei D, Helvie M A, Sahiner B, Adler D D, Goodsitt M M and Petrick N 1995 Computer-aided classification of mammographic masses and normal tissue: linear discriminant analysis in texture feature space *Phys. Med. Biol.* **40** 857–76

Duda R O and Hart P E 1973 *Pattern Classification and Scene Analysis* (New York: Wiley)

Galloway M M 1975 Texture classification using grey level run lengths *Comput. Graphics Image Process.* **4** 172–9

Hall F M, Storella J M, Silverstone D Z and Wyshak G 1988 Nonpalpable breast lesions: recommendations for biopsy based on suspicion of carcinoma at mammography *Radiology* **167** 353–8

Haralick R M, Shanmugam K and Dinstein I 1973 Texture features for image classification *IEEE Trans. Systems Man Cybernetics* **3** 610–21

Hermann G, Janus C, Schwartz I S, Krivisky B, Bier S and Rabinowitz J G 1987 Nonpalpable breast lesions: accuracy of prebiopsy mammographic diagnosis *Radiology* **165** 323–6

Huo Z, Giger M L, Vyborny C J, Bick U, Lu P, Wolverton D E and Schmidt R A 1995 Analysis of spiculation in the computerized classification of mammographic masses *Med. Phys.* **22** 1569–79

Jacobson H G and Edeiken J 1990 Biopsy of occult breast lesions: analysis of 1261 abnormalities *J. Am. Med. Assoc.* **263** 2341–3

Jain A K 1989 *Fundamentals of Digital Image Processing* (New Jersey: Prentice-Hall)

Jiang Y, Metz C E and Nishikawa R M 1996 A receiver operating characteristic partial area index for highly sensitive diagnostic tests *Radiology* **201** 745–50

Kilday J, Palmieri F and Fox M D 1993 Classifying mammographic lesions using computerized image analysis *IEEE Trans. Med. Imaging* **12** 664–9

Lachenbruch P A 1975 *Discriminant Analysis* (New York: Hafner)

McClish D K 1989 Analyzing a portion of the ROC curve *Med. Decision Making* **9** 190–5

Metz C E, Herman B A and Shen J H 1998 Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data *Stat. Med.* **17** 1033–53

Pohlman S, Powell K A, Obuchowski N A, Chilcote W A and Broniatowski S G 1996 Quantitative classification of breast tumors in digitized mammograms *Med. Phys.* **23** 1337–45

Rangayyan R M, El-Faramawy N, Desautels J E L and Alim O A 1996 Discrimination between benign and malignant breast tumors using a region-based measure of edge profile acutance *Digital Mammography '96* ed K Doi, M L Giger, R M Nishikawa and R A Schmidt (Amsterdam: Elsevier) pp 213–18

Raudys S J and Jain A K 1991 Small sample size effects in statistical pattern recognition: recommendations for practitioners *IEEE Trans. Pattern Anal. Machine Intell.* **13** 252–64

Sahiner B, Chan H-P, Petrick N, Goodsitt M M and Helvie M A 1997 Characterization of masses on mammograms: significance of the use of the rubber-band straightening transform *Proc. SPIE* **3034** 491–500

Sahiner B, Chan H-P, Petrick N, Helvie M A and Goodsitt M M 1998 Computerized characterization of masses on mammograms: the rubber band straightening transform and texture analysis *Med. Phys.* **25** 516–26

Sahiner B, Chan H-P, Petrick N, Helvie M A, Goodsitt M M and Adler D D 1996a Classification of masses on mammograms using a rubber-band straightening transform and feature analysis *Proc. SPIE* **2710** 44–50

Sahiner B, Chan H-P, Petrick N, Wei D, Helvie M A, Adler D D and Goodsitt M M 1995 Classification of mass and normal breast tissue: an artificial neural network with morphological features *Proc. World Congress on Neural Networks* vol 2 (New Jersey: INNS Press) pp 876–9

——1996b 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

——1996c Image feature selection by a genetic algorithm: application to classification of mass and normal breast tissue *Med. Phys.* **23** 1671–84

Wagner R F, Chan H-P, Mossoba J, Sahiner B and Petrick N 1997 Finite-sample effects and resampling plans: application to linear classifiers in computer-aided diagnosis *Proc. SPIE* **3034** 467–77

Wei D, Chan H-P, Helvie M A, Sahiner B, Petrick N, Adler D D and Goodsitt M M 1995 Classification of mass and normal breast tissue on digital mammograms: multiresolution texture analysis *Med. Phys.* **22** 1501–13

Weszka J S, Dyer C R and Rosenfeld A 1976 A comparative study of texture measures for terrain classification *IEEE Trans. Syst. Man Cybernetics* **6** 269–85