Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers

Heang-Ping Chan[a] and Berkman Sahiner
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109-0030

Robert F. Wagner
Center for Devices and Radiology Health, FDA, Rockville, Maryland 20852

Nicholas Petrick
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109-0030

(Received 14 June 1999; accepted for publication 16 September 1999)

Classifier design is one of the key steps in the development of computer-aided diagnosis (CAD) algorithms. A classifier is designed with case samples drawn from the patient population. Generally, the sample size available for classifier design is limited, which introduces variance and bias into the performance of the trained classifier, relative to that obtained with an infinite sample size. For CAD applications, a commonly used performance index for a classifier is the area, A_z, under the receiver operating characteristic (ROC) curve. We have conducted a computer simulation study to investigate the dependence of the mean performance, in terms of A_z, on design sample size for a linear discriminant and two nonlinear classifiers, the quadratic discriminant and the backpropagation neural network (ANN). The performances of the classifiers were compared for four types of class distributions that have specific properties: multivariate normal distributions with equal covariance matrices and unequal means, unequal covariance matrices and unequal means, and unequal covariance matrices and equal means, and a feature space where the two classes were uniformly distributed in disjoint checkerboard regions. We evaluated the performances of the classifiers in feature spaces of dimensionality ranging from 3 to 15, and design sample sizes from 20 to 800 per class. The dependence of the resubstitution and hold-out performance on design (training) sample size (N_t) was investigated. For multivariate normal class distributions with equal covariance matrices, the linear discriminant is the optimal classifier. It was found that its A_z-versus-$1/N_t$ curves can be closely approximated by linear dependences over the range of sample sizes studied. In the feature spaces with unequal covariance matrices where the quadratic discriminant is optimal, the linear discriminant is inferior to the quadratic discriminant or the ANN when the design sample size is large. However, when the design sample is small, a relatively simple classifier, such as the linear discriminant or an ANN with very few hidden nodes, may be preferred because performance bias increases with the complexity of the classifier. In the regime where the classifier performance is dominated by the $1/N_t$ term, the performance in the limit of infinite sample size can be estimated as the intercept $(1/N_t=0)$ of a linear regression of A_z versus $1/N_t$. The understanding of the performance of the classifiers under the constraint of a finite design sample size is expected to facilitate the selection of a proper classifier for a given classification task and the design of an efficient resampling scheme.
[S0094-2405(99)00212-6]

Key words: computer-aided diagnosis, classifier design, linear classifier, quadratic classifier, neural network, sample size, feature space dimensionality, ROC analysis

I. INTRODUCTION

With the advent of digital imaging modalities, computer-aided diagnosis (CAD) is becoming an important area of research in medical imaging. A CAD algorithm can detect abnormalities and classify disease or normal cases based on image and/or patient information, and thus provide a second opinion to the radiologist in the detection or diagnostic decision making process.

Design of classifiers that can accurately distinguish normal and abnormal features is a critical step in the development of CAD algorithms. It has been shown that the performance of a classifier for unknown cases depends on the sample size used for training.[1] When a finite design (training) sample size is used, the performance is pessimistically biased in comparison to that obtained from an infinitely large design sample. In order to design a classifier with a performance generalizable to the population at large, one has to use a sufficient number of case samples that are representative of the population. However, the availability of case samples is often limited in medical imaging research. It is therefore important to study the sample-size dependence of different classifiers and determine the most efficient way of training a classifier, under the constraint of a finite sample size.

We note that the concept of generalizability may be used in several technical senses when assessing the performance of a classifier: one with respect to mean classifier performance, the other with respect to the variance of classifier performance. In many classifier design problems, one is most interested in investigating if the mean performance of a classifier estimated from a given set of finite design samples can be generalized to classification performance with unknown test samples drawn from the same population of cases. The generalizability in this regard can be observed from the biases of the mean performances in the finite design set and in the test set in comparison to the optimal performance estimated from an infinite design set. The bias in the mean performance of different classifiers under various input conditions is the subject of investigation in this study. We will discuss further other interpretation of generalizability in the Discussion section of this paper.

A number of investigators have studied the finite-sample-size problem[1–9] Fukunaga[1,3] derived a general formulation for the bias and variance of a function, f, which is to be estimated from the available samples. When f is a nonlinear function of the mean vectors and covariance matrices of two feature distributions, it has been shown that a bias results from the nonlinear propagation of the finite-sample variances in the estimates of the mean vectors and covariance matrices of the distributions through this function. For multivariate-normal data, these variances are proportional to $1/N_t$, where N_t is the design sample size, and this dependence propagates into the lowest-order terms in the bias. The bias is independent of the test sample size, N_{test}. All measures of classifier performance that count the fraction of times the decision value for an abnormal case exceeds that for a normal case (independent of underlying distribution), and various measures of error for normally distributed decision functions, are nonlinear functions of the parameters of the underlying distributions. They are thus subject to this effect. Fukunaga and Hayes[3] analyzed the finite sample effects on the probability of misclassification (PMC) of a classifier and suggested a technique that makes use of the linear dependence of PMC on $1/N_t$ to estimate the performance at $N_t \rightarrow \infty$ with a finite sample set.

For the evaluation of medical diagnostic systems, the most commonly used performance index is the area under the receiver operating characteristic (ROC) curve, A_z. We have derived analytically that, for linear discriminant classifiers, the classifier performance in terms of A_z can be approximated by a linear function in $1/N_t$, under conditions when higher order terms in N_t can be neglected. We have been investigating the dependence of A_z on sample size by simulation studies.[7–9] Wagner *et al.*[10,11] have also analyzed the effects of design and test sample sizes on the variance components of the classifier performance. Although these behaviors depend strongly on the class distributions and the properties of the classifier, the studies will provide some insight into the sample size requirements for the design of different classifiers. This work may eventually lead to the selection of an efficient resampling scheme for classifier design, as well as the development of a statistical test of the sample size requirements and the generalizability of the trained classifier.

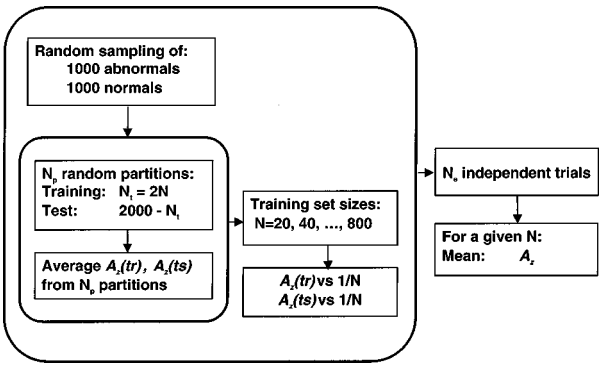


FIG. 1. The sampling and evaluation scheme of the simulation study.

In this paper, we will describe the simulation studies and analyze the effects of sample size on classifier performance. Several commonly used classifiers, including the linear discriminant, the quadratic discriminant, and the back-propagation neural network will be studied and compared under different input conditions. Feature distributions with markedly different characteristics will be used to represent a variety of situations that may be encountered in classification problems for many detection or diagnostic tasks.

II. MATERIALS AND METHODS

We performed simulation studies to evaluate the effects of sample size on classifier design. Normal and abnormal case samples were randomly drawn from known probability distributions of the two classes. These samples were then used to design classifiers for differentiation of normal and abnormal cases. The simulation approach assures that any number of case samples can be obtained from populations with known statistical properties. It thus allows evaluation of the dependence of classifier performance on design sample size and comparison of the performance with theoretically predicted optimal classification based on the chosen probability distributions.

A. Simulation study

The sampling and evaluation scheme of the simulation study is shown in Fig. 1. In this study, we considered only the situation in which equal numbers ($=N_{\text{total}}/2$) of normal and abnormal cases randomly drawn from the class distributions were available in our data set. A resampling strategy similar to the technique suggested by Fukunaga and Hayes was devised to generate the A_z-vs-$1/N_t$ curve. Subsets of $N_{t_1}, N_{t_2}, \ldots, N_{t_j}$ design samples were randomly drawn from the available sample set, again under the constraint that the numbers of normal and abnormal samples were equal in each subset, i.e., $N_{t_i,\text{normal}} = N_{t_i,\text{abnormal}} = N_{t_i}/2 \, (i = 1,...,j)$. A classifier was designed by using each subset of samples. The random sampling of a given subset from the available set of N_{total} samples was performed without replacement, whereas the random sampling of different subsets always started from

the same set of N_{total} samples. Therefore, after drawing a given design subset N_{t_i}, the remaining samples, $N_{\text{total}}-N_{t_i}$ were independent of the design samples and used as the test samples. For simplicity, the number of design samples per class is denoted as N in the following discussion.

In general, there are two methods, resubstitution and hold-out, for testing classifier performance. In the resubstitution method, the design sample set is resubstituted into the trained classifier to test its performance, whereas in the hold-out method, an independent test set is used. It has been shown[1] that, for a Bayes classifier, if the classifier is trained with a finite number of design samples, the resubstitution estimate of the classifier performance is optimistically biased whereas the hold-out estimate is pessimisticaly biased in comparison to that achievable with an infinite design sample set. The mean performance obtained from the former estimation provides an upper bound and that from the latter provides a lower bound on the true classifier performance. When the design sample size is limited, it is important to evaluate the hold-out performance to avoid an overly optimistic prediction of the classifier performance. In the limit of very large sample size, the upper and lower bounds converge towards the unbiased estimate.

In this study, we evaluated the performance of the classifier using both the resubstitution and the hold-out methods as a function of finite design sample size N_t. In order to reduce the variances in the estimates of A_z, we randomly resampled without replacement each N_{t_i} from the same N_{total} samples N_p times, trained and tested the classifier, and estimated the average A_z from the N_p individual A_z's as shown in Fig. 1. The resubstitution or hold-out A_z-vs-$1/N_t$ curve was plotted from the j points and the unbiased estimate of A_z in the limit of $N_t \rightarrow \infty$ could be extrapolated from either curve.

This method of estimating classifier performance at large N_t by generating a few data points at finite sample sizes is similar to the Fukunaga and Hayes technique. However, we did not assume that the j points were in the linear region of the A_z-vs-$1/N_t$ curve and we used resampling to reduce the variances. In fact, one of the goals of this study was to investigate the range of design sample size in which the performance curve was approximately linear for various classifiers and probability distributions of the class populations. Therefore, we used a much larger total number of samples ($N_{\text{total}}=2000$) in our simulation study than was generally available for classifier design. We could then choose N_{t_i} over a wide range and study the behavior of the entire A_z-vs-$1/N_t$ curve.

To estimate the population mean of A_z at each N_{t_i}, we repeated the above experiment N_e times, each with 2000 independently drawn samples from the population. The population mean of A_z was estimated by averaging the A_z values obtained from the N_e experiments. We did not analyze the variances in this study because of the complication in the correlation among the N_p values of A_z introduced by resampling. A detailed analysis of the variances and its modeling was performed in a separate study by Wagner *et al.*[10,11] in which a different study design was used.

By varying the number of design samples per class, N, over a large range from 20 to 800, the regime where the $1/N_t$ dependence dominated could be observed from the A_z (population mean)-vs-$1/N_t$ (or $1/N$) curves. It is important to note that, although the number of test samples, $N_{\text{test}_i}=2000-N_{t_i}$, varied from point to point on both the resubstitution and the hold-out curves, the bias in A_z is independent of N_{test_i}.[1] The shape of the A_z-vs-$1/N$ curve is independent of N_{test_i} after N_{t_i} is fixed. However, the variance of a given A_z does depend on the test sample size.

For simplicity, we will refer to these estimates of A_z (population mean) as $A_z(\text{tr})$ for the resubstitution and as $A_z(\text{ts})$ for the hold-out performance in the following discussions.

B. Class distributions

1. Multivariate normal distributions

For three of the four types of class distributions, we assumed that the normal and abnormal classes followed multivariate normal distributions in the feature space. The dimensionality of the feature space, k, was varied from 3 to 15. The characteristics of the multivariate normal distributions can be completely specified by the multivariate mean vector of the rth class, denoted as $\mu_r\,(r=1,2)$ and its covariance matrix, denoted as Σ_r. The separation of the normal and abnormal classes is measured by the Bhattacharyya distance, B, defined as[1,12]

$$B=\frac{1}{8}\Delta+\frac{1}{2}\ln\frac{\det[(\Sigma_1+\Sigma_2)/2]}{\sqrt{\det\Sigma_1}\sqrt{\det\Sigma_2}}, \tag{1}$$

where $\det\Sigma_r$ denotes the determinant of Σ_r, and Δ is the squared Mahalanobis distance,[12] defined as

$$\Delta=(\mu_2-\mu_1)^T\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}(\mu_2-\mu_1). \tag{2}$$

The Mahalanobis distance is the Euclidean distance between the means of the two distributions, normalized by the square root of the average of their covariance matrices. It can therefore be considered to be a measure of the signal-to-noise ratio (SNR) between the abnormal and the normal distributions. The second term of B is the contribution from the difference in the covariance matrices of the two class distributions. If the covariance matrices are equal, the second term will be zero and the Bhattacharyya distance will be equal to 1/8 of the squared Mahalanobis distance.

In the current study, three types of multivariate normal class distributions were considered. In the following discussion, we shall refer to the use of simultaneous diagonalization for the two covariance matrices of the class distributions. This operation leaves the normal-based decision functions unchanged because the distance measures that arise in these decision functions are invariant to any non-singular linear transformation.[1]

(1) Equal covariance matrices and unequal means: In this case, the covariance matrices of the normal and abnormal class distributions can be simultaneously diagonalized

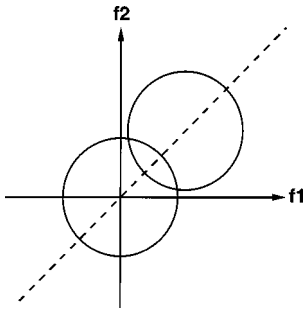


FIG. 2. A schematic illustration of the two class distributions with equal covariance matrices and unequal means in a 2D feature space. The circles represent contours of equal probability in each distribution.

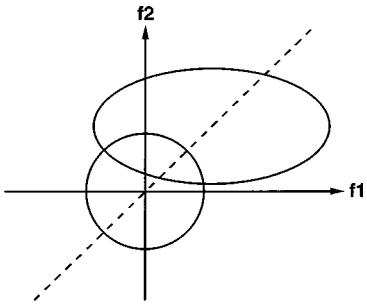


FIG. 3. A schematic illustration of the two class distributions with unequal covariance matrices and unequal means in a 2D feature space. The closed curves represent contours of equal probability in each distribution.

and the variances of the individual feature components can be scaled to unity. Therefore, without loss of generality, the covariance matrices of the two classes could be assumed to be equal to identity matrices, $\Sigma_1=\Sigma_2=I$. The mean feature vector for the first class was assumed to be zero, $\mu_1=\mathbf{0}$, and the mean feature vector for the second class, $\mu_2=\mathbf{M}$ with all components of $\mathbf{M}$ equal to a constant m. The magnitude of m could be adjusted to obtain a desired separation of the two classes. For the purpose of this simulation study, we chose m such that the squared Mahalanobis distance was 3, i.e., the Bhattacharyya distance was 3/8, for feature spaces of any dimensionality. As discussed below, this separation corresponds to a theoretical A_z of 0.89, which is in the performance range of many classification problems in CAD applications. An example of the two class distributions in a 2D feature space is shown schematically in Fig. 2.

(2) Unequal covariance matrices and unequal means: The covariance matrix of the first class was again diagonalized and scaled to be an identity matrix, $\Sigma_1=I$, and the mean feature vector for the first class was assumed to be zero, $\mu_1=\mathbf{0}$. The covariance matrix of the second class, Σ_2, was simultaneously diagonalized to have eigenvalues λ_i, $i=1,...,k$. For this study, we generated the values of λ_i with the simple relationship:

$$\lambda_i=\lambda_{\min}+\frac{(i-1)(\lambda_{\max}-\lambda_{\min})}{(k-1)},\quad i=1,...,k \tag{3}$$

and evaluated one condition where $\lambda_{\min}=1$, and $\lambda_{\max}=2$ for all dimensionalities of the feature spaces. We also assumed that the components of the mean feature vector μ_2 were equal, the values of which were adjusted to achieve a Bhattacharyya distance of 3/8. For the purpose of demonstrating the general trends of the A_z-vs-$1/N$ curves and comparing the relative performance of the different classifiers under the various conditions, the specific choices of these values are not critical. Figure 3 illustrates an example of the two class distributions in a 2D feature space.

(3) Unequal covariance matrices and equal means: The covariance matrix of the first class was the same as that in the first two cases described above. The covariance matrix of the second class was proportional to the identity matrix, $\Sigma_2=\alpha I$, where the proportionality constant α was adjusted to provide a Bhattacharyya distance of 3/8. The mean feature vectors of the two classes were equal, $\mu_1=\mu_2=\mathbf{0}$. In this case, the discriminatory power of the two classes comes entirely from the difference in the covariance matrices. A schematic of the two class distributions in a 2D feature space is shown in Fig. 4.

2. Checkerboard distributions

The fourth type of class distributions was a checkerboard where the normal and abnormal classes were located in alternate square box regions of the feature space. Within each box of the checkerboard, the feature vectors were uniformly distributed. The two classes did not overlap with each other so that they could be perfectly separated by an ''ideal'' classifier with $A_z=1$. We considered a 2×3 checkerboard in a 2D feature space and a $2\times2\times2$ checkerboard in a 3D feature space. The example of a 2×3 checkerboard in a 2D feature space is shown in Fig. 5. Such class distributions may not be common in actual classification problems encountered in CAD. However, it was included in this study to demonstrate the capability and limitations of the different classifiers when the class distributions were not multivariate normal.

C. Classifiers

We studied three types of classifiers: the linear discriminants, the quadratic discriminants, and the back-propagation neural networks. They represent a range of classifiers commonly used in the field of pattern recognition at present.

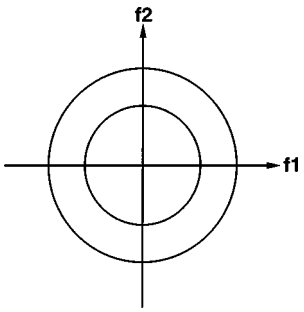


FIG. 4. A schematic illustration of the two class distributions with unequal covariance matrices and equal means in a 2D feature space. The circles represent contours of equal probability in each distribution.

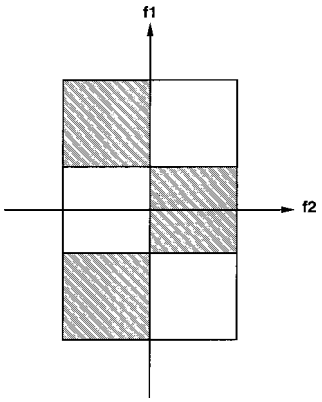


FIG. 5. An example of a 2×3 checkerboard in a 2D feature space.

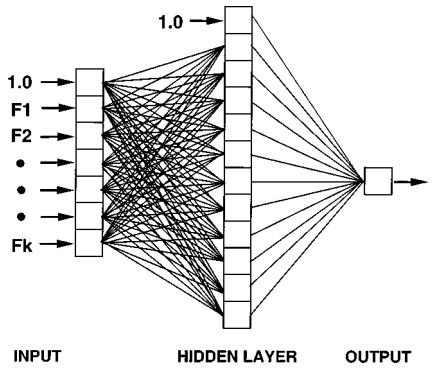


FIG. 6. A schematic diagram of a backpropagation neural network with one hidden layer.

(1) Linear discriminant classifier: The linear discriminant classifier can be derived from the means and the covariance matrices of the class distributions as follows:[1,13]

$$h_l(X)=(\mu_2-\mu_1)^T\overline{\Sigma}^{-1}X+\tfrac{1}{2}(\mu_1^T\overline{\Sigma}^{-1}\mu_1-\mu_2^T\overline{\Sigma}^{-1}\mu_2), \quad (4)$$

where $\overline{\Sigma}=(\Sigma_1+\Sigma_2)/2$, and X is the feature vector to be classified. The means and covariance matrices have to be estimated as the sample means and sample covariance matrices from the available design samples. The sample means and covariance matrices undergo a nonlinear transformation to become the discriminant scores, which in turn are transformed nonlinearly into a measure of the performance. The variances in the estimated parameters propagate into the mean classifier performance and result in a bias through the second derivative of the transformation function.

It is known that, for multivariate normal distributions with equal covariance matrices, the linear discriminant classifier is optimal and the classifier performance in the limit of large design samples is determined by the Mahalanobis distance, given by

$$A_Z=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\sqrt{\Delta/2}}e^{-u^2/2}\,du. \quad (5)$$

For the class distributions with $\Delta=3$ to be used in this study, it can be derived from Eq. (5) that the maximum A_z that the optimal linear discriminant can achieve in the limit of large design samples is 0.89.

(2) Quadratic discriminant classifier: The quadratic discriminant classifier can be expressed as[1]

$$h_q(X)=\frac{1}{2}(X-\mu_1)^T\Sigma_1^{-1}(X-\mu_1)-\frac{1}{2}(X-\mu_2)^T\Sigma_2^{-1}(X-\mu_2)+\frac{1}{2}\ln\frac{\det\Sigma_1}{\det\Sigma_2}. \quad (6)$$

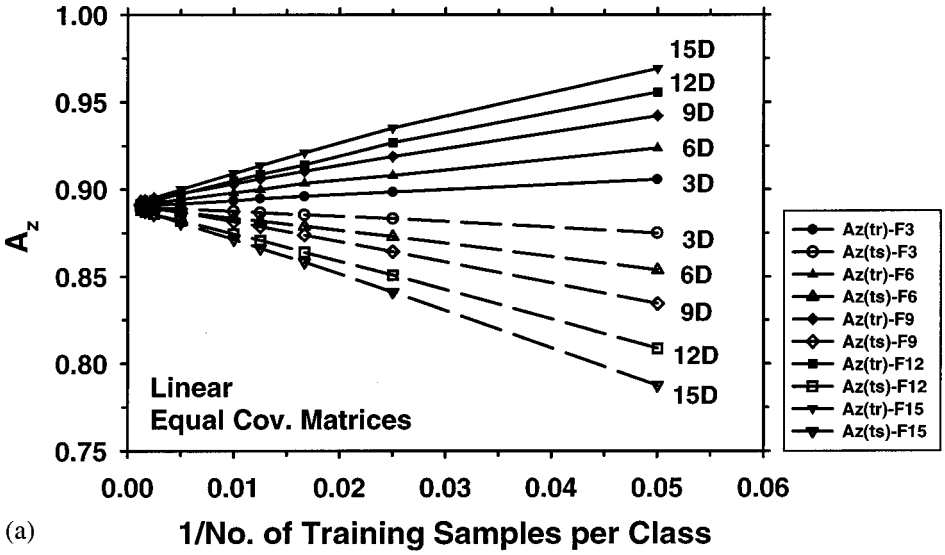
When the class distributions are multivariate normal with unequal covariance matrices, the quadratic discriminant classifier is optimal in the limit of large training samples. The Bhattacharyya distance gives an upper bound on the Bayes error.[1] The general properties of the linear and quadratic classifiers have been described in the literature (for example, Fukunaga[1]).

(3) Back-propagation neural network: Many different architectures and training methods have been developed for artificial neural networks (ANN)[14] in various applications. In this study, we considered only a three-layered neural network trained with a feed-forward back-propagation method. The neural network has k input nodes, n hidden nodes, one output node, and a bias node in both the input and the hidden layers. The ANN architecture is denoted as $k-n-1$. The nodes in the ANN are fully connected and are trained with a minimum sum-of-squares-error criterion. The number of weights to be estimated is equal to $n(k+1)+(n+1)$. A schematic diagram of an ANN is shown in Fig. 6.

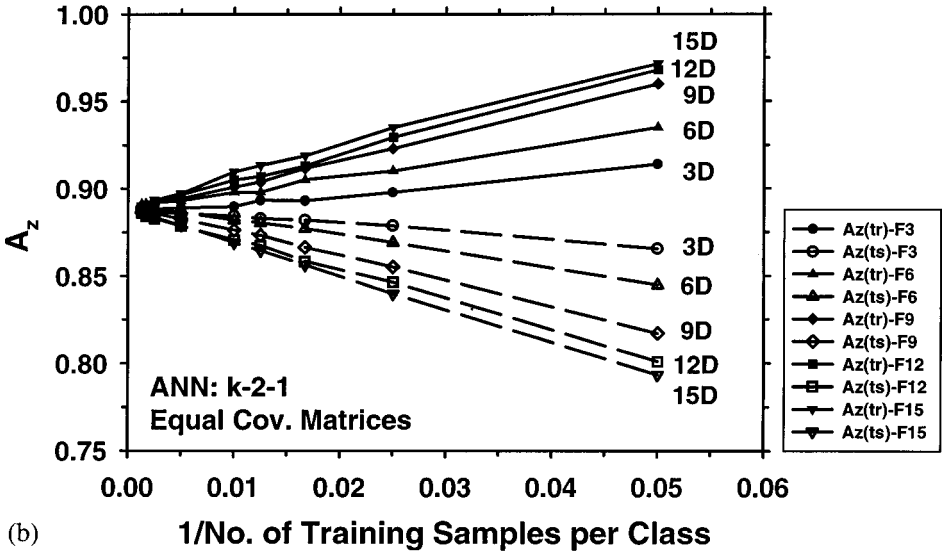
III. RESULTS

In our simulation study, we compared the performance of the linear, quadratic, and backpropagation neural network classifiers for the different class distributions in the feature spaces of dimensionality ranging from 3 to 15. The number of repeated experiments N_e was chosen to be 20 for all cases in the multivariate normal feature spaces and 100 in the checkerboard feature space. The number of data set partitionings N_p in each experiment ranged from 1 to 20. These choices are a compromise between computation time and estimation accuracy, especially for ANN classifiers with a large number of hidden nodes in high dimensional feature spaces. As shown in the graphs discussed below, some of the performance curves may exhibit fluctuations that could be reduced by a larger number of experiments. However, the general trend of the performance curves should not be changed by the statistical uncertainties.

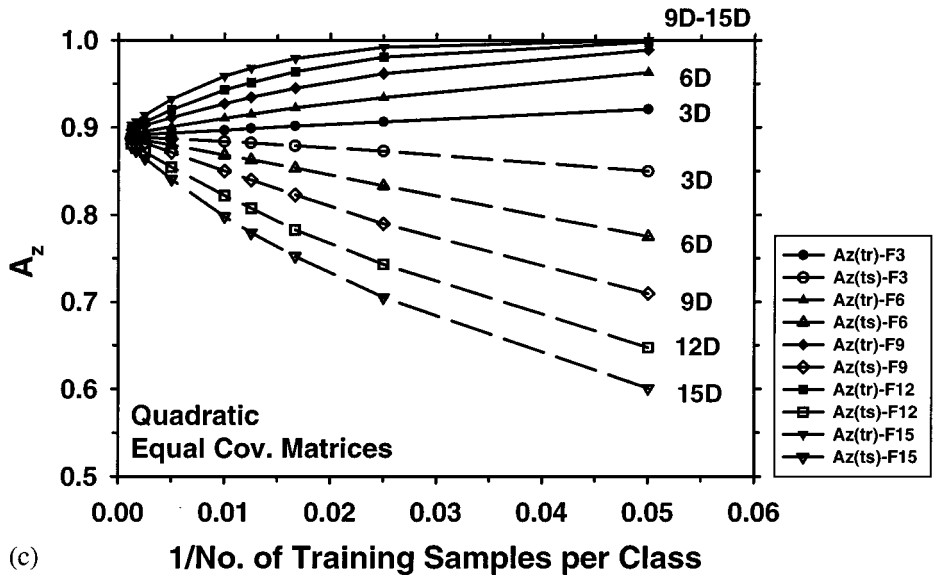
(1) Multivariate normal distributions—Equal covariance matrices and unequal means: For class distributions with equal covariance matrices, the linear discriminant is theoretically the optimal classifier when the design sample size is large. However, when the design sample size is small, the performances of all classifiers are biased. Figures 7(a)–7(c) show the dependence of the A_z obtained from resubstitution (training), $A_z(\mathrm{tr})$, and the A_z obtained from the hold-out method (testing), $A_z(\mathrm{ts})$, on $1/N$ for the linear, ANN, and



(a)



(b)



(c)

FIG. 7. The dependence of the A_z obtained from resubstitution (training-solid lines), A_z(tr), and the A_z obtained from the hold-out method (testing—dashed lines), A_z(ts), on $1/N$ for the class distributions with equal covariance matrices and unequal means. (a) Linear, (b) ANN, and (c) quadratic classifier. Legend: F3=3D feature space, etc.

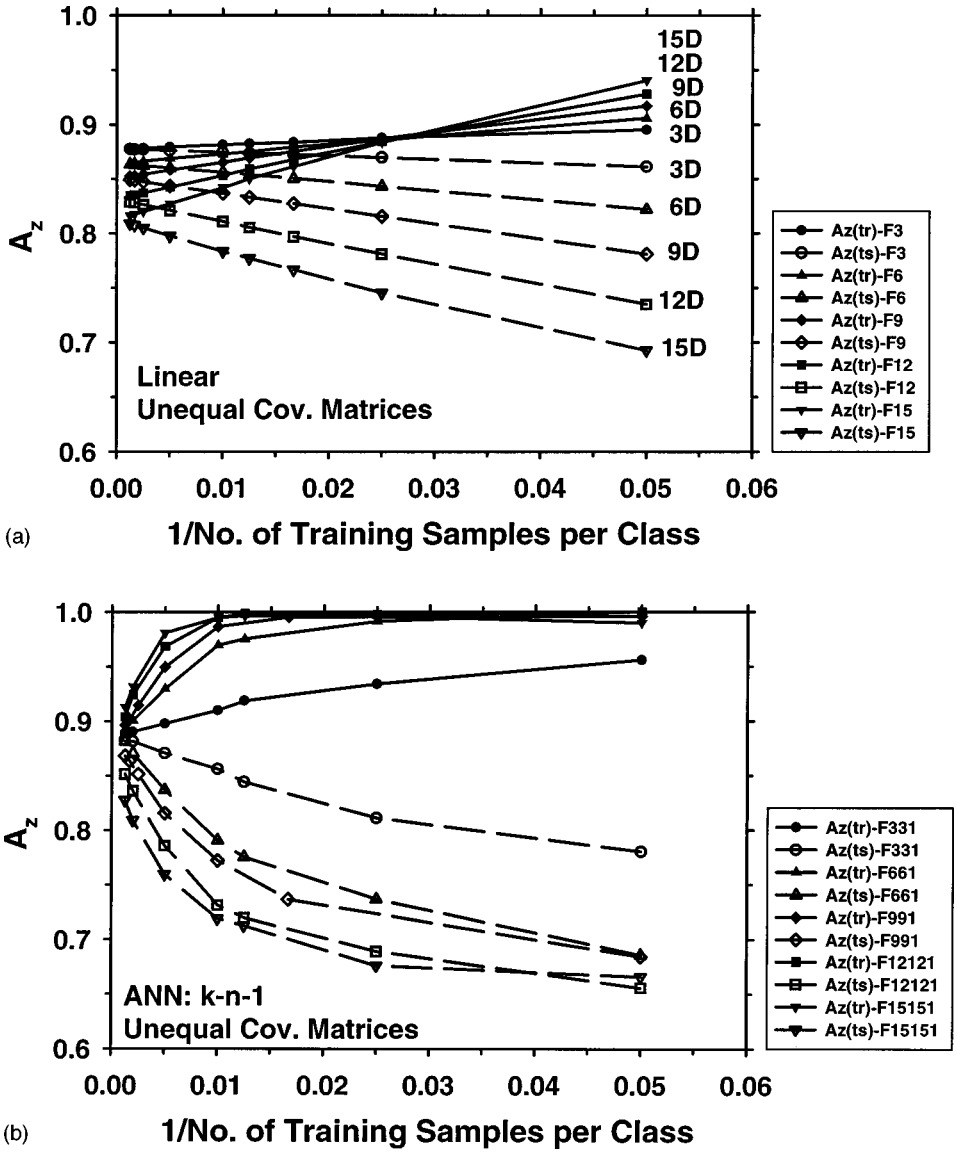


FIG. 8. The performances of the classifiers for class distributions with unequal covariance matrices and unequal means. (a) Linear, (b) ANN classifier. Legend: F3=3D feature space, etc., solid lines $=A_z(\mathrm{tr})$, dashed lines$=A_z(\mathrm{ts})$.

quadratic classifier, respectively. Two hidden nodes were used for the ANN $(k-2-1)$ because it is the smallest number of hidden nodes in a nonlinear ANN. An ANN with only one hidden node will be a linear classifier and behave in a similar manner as the linear discriminant. On the other hand, ANNs with a large number of hidden nodes (not shown) will overfit the design samples and have poor generalizability to the unknown cases, similar to the ANN curves to be discussed below. All three classifiers can reach the optimal classification accuracy of $A_z=0.89$ in the limit of large N. The curves for the linear classifier and the ANN $(k-2-1)$ at 400 training epochs (iterations) are approximately linear over the entire range. The quadratic classifier does not reach the approximately linear region until N is greater than about 100 $(1/N<0.01)$ in the higher-dimensional feature space. The biases on both the resubstitution and hold-out curves for the quadratic classifier are greater than those for the linear classifier and the ANN $(k-2-1)$. The large biases again indicate overfitting and poor generalization by the quadratic classifier in the equal-covariance-matrices situation.

(2) Multivariate normal distributions—Unequal covariance matrices and unequal means: The performances of the classifiers for class distributions with unequal covariance matrices are shown in Figs. 8(a)–8(b). The linear discriminant and the ANN $(k-2-1)$ classifier (not shown) are again approximately linear over the entire range of N studied. However, the A_z at $1/N=0$ decreases as the dimensionality of the feature space increases. This is because both the linear discriminant and the near-linear ANN $(k-2-1)$ cannot make use of the class separability due to the differences in the covariance matrices which is the second term in the Bhattacharyya distance. The second term increases relative to the first term, the squared Mahalanobis distance, when the Bhattacharyya distance is fixed and the dimensionality of the feature space increases.

The performance curves of the ANN at large N improve when a greater number of hidden nodes and a sufficient number of training epochs are used. The number of hidden nodes required to reach the optimal classification of $A_z=0.89$ at $1/N=0$ increases with the dimensionality of the feature

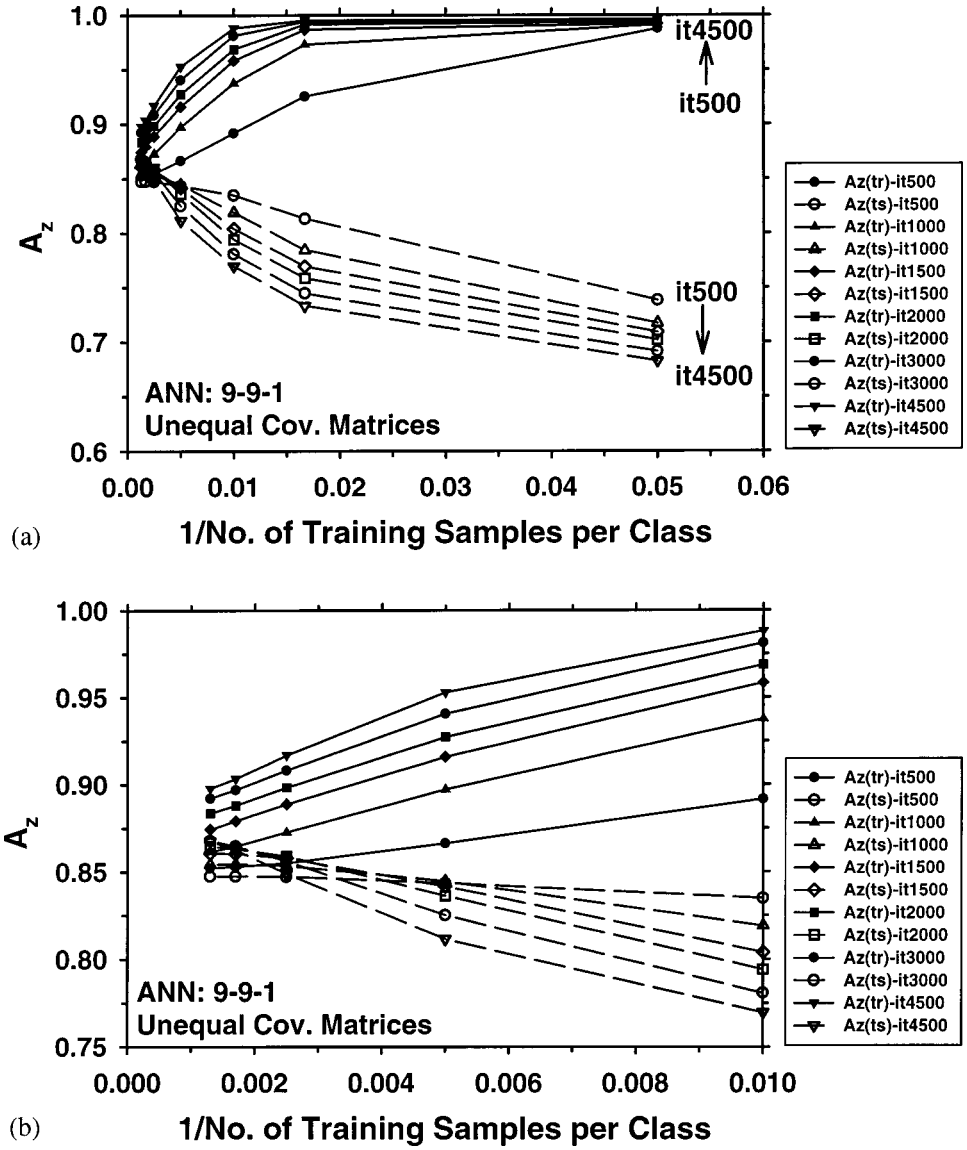


FIG. 9. The dependence of the performance curves on the number of training epochs for an ANN with nine hidden nodes in a 9D feature space: ANN(9−9−1). Legend: it500=500 training epochs, etc., solid lines=A_z(tr), dashed lines=A_z(ts). The expanded view in (b) shows the trend of the curves at large sample sizes.

space. Figure 8(b) shows the performance of the ANNs when the number of hidden nodes is equal to the dimensionality in each feature space. Since the number of weights to be trained increases rapidly with increasing number of nodes in an ANN, the number of epochs required for training the ANN to achieve a reasonable classification accuracy increases accordingly. The resubstitution and hold-out performance curves of each ANN shown in Fig. 8(b) were chosen at the smallest number of training epoch that resulted in approximately the highest A_z value when the hold-out curve was extrapolated to $1/N=0$. The number of training epochs required to reach the highest A_z increased as the dimensionality and the number of hidden nodes in the ANN increased. It ranged from about 4000 to 10 000 for the conditions shown in Fig. 8(b). We did not attempt to perform an exhaustive search for the ''optimal'' number of hidden nodes in each feature space because of the extensive computation time required for the search. Instead, we evaluated ANNs with a few different numbers of hidden nodes in each feature space and chose the ''best'' ANN within those studied. With this approximation we observed that, in a k-dimensional feature space and with these class distributions, an ANN with approximately k hidden nodes can approach the optimal performance when the design sample size and the number of training epochs are sufficiently large, as shown in Fig. 8(b).

To illustrate the training of an ANN with a large number of hidden nodes, we show the dependence of the resubstitution and the hold-out curves on the number of training epochs for ANN (9−9−1) in Fig. 9. A number of commonly discussed problems of an ANN can be observed. In the small N region below about 60 samples per class, overparametrization and over-training are obvious, i.e., near perfect classification during training [A_z(tr) greater than 0.95] and poor generalization [A_z(ts) below about 0.8]. The problem becomes more pronounced with an increasing number of training epochs. In the middle range of 200 to 400 samples per class where A_z(ts) increases to a maximum then decreases with further training, an ''optimal'' number of training epoch exists. Only in the region with a sufficiently large N (greater than about 500 per class), A_z(ts) increases with

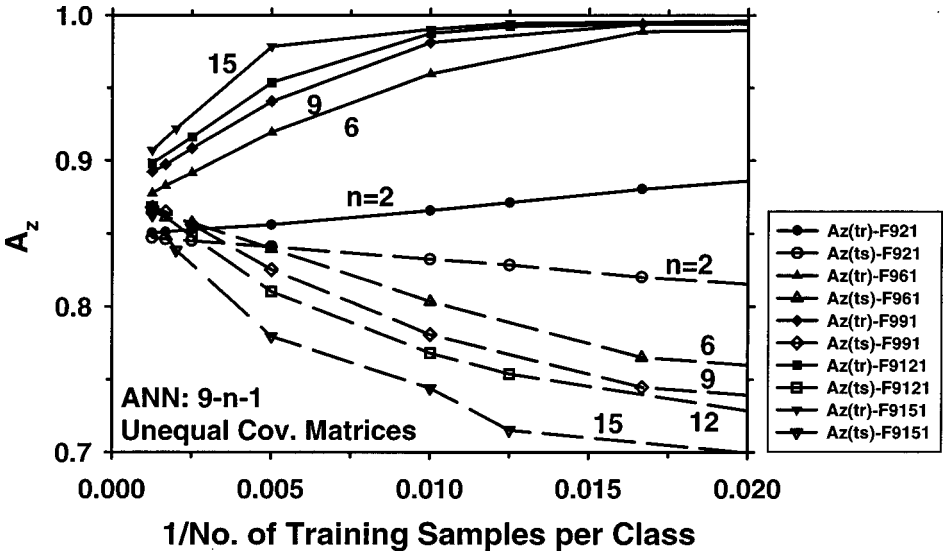


FIG. 10. The dependence of the performance curves of an ANN on the number of hidden nodes in the 9D feature space for class distributions with unequal covariance matrices and unequal means. Legend: F921=ANN with two hidden nodes, etc., solid lines=A_z(tr), dashed lines =A_z(ts).

increasing number of training epochs within the range studied. The A_z(ts)-vs-$1/N$ curve becomes linear for N greater than about 200. This dependence of ANN on training epoch is generally observed for ANNs with a large number of hidden nodes and in high-dimensional feature spaces, although the design sample size required in order to avoid overtraining and over-parametrization varies. It reinforces our general experience that the ANNs with a large number of weights can overfit the design samples easily and provide poor generalization when the sample size is small.

The performance curves of ANNs with different numbers of hidden nodes in the 9D feature space are shown in Fig. 10. The curves for a given ANN were again chosen at a training epoch in which the hold-out curve approached approximately the highest performance at $1/N=0$. The chosen training epoch ranged from 600 to 12 000 for the 2- to 15-hidden-node ANNs shown. When the number of hidden nodes is small, the highest A_z obtained by extrapolation to $1/N=0$ appears to be below the theoretical optimum of 0.89. For example, the A_z extrapolated to $1/N=0$ is about 0.85 for ANN (9−2−1), and is about 0.87 for ANN (9−6−1). The ANN with nine hidden nodes appears to approach the optimal A_z of 0.89 in the limit of $1/N=0$. However, the ANN (9−9−1) does not reach the approximately linear region until N is greater than about 200 (easier to see in Fig. 9). As can be seen from the hold-out curves, increasing the number of hidden nodes further will increase overfitting, reduce generalizability, and increase train time without gaining true improvement in performance for classification of unknown case samples.

The quadratic classifier is the theoretically optimal classifier for the class distributions with unequal covariance matrices. It can optimally utilize the class separability contributed by both the differences in the means and the covariance matrices. The performance curves for the quadratic classifier (not shown) in feature spaces of different dimensionalities are very similar to those obtained for the equal covariance matrices situation [Fig. 7(c)]. The A_z of the quadratic classi-

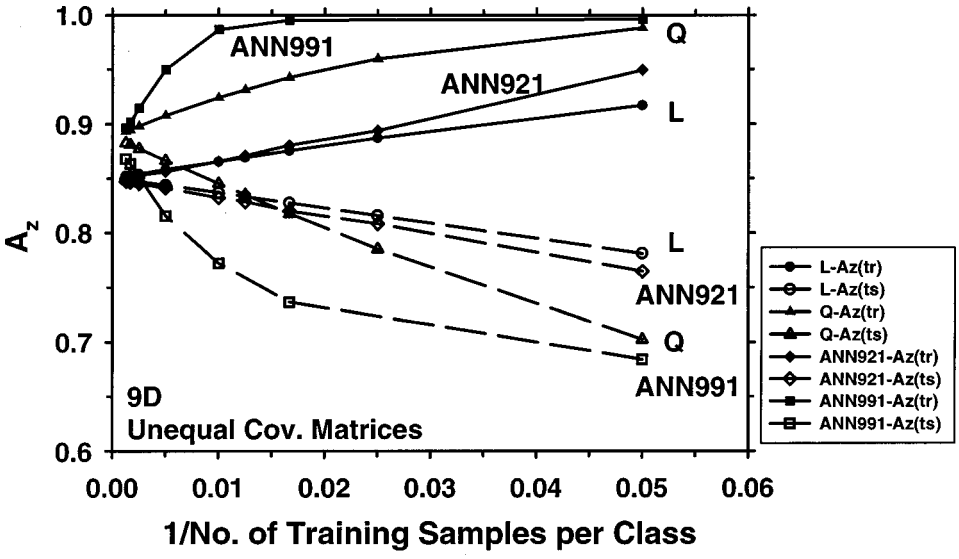


FIG. 11. Comparison of the performance curves of the linear, quadratic, ANN(9−2−1), and ANN(9−9−1) classifiers in the 9D feature space for class distributions with unequal covariance matrices and unequal means. Legends: L=linear; Q=quadratic, ANN=neural network, solid lines=A_z(tr), dashed lines=A_z(ts).

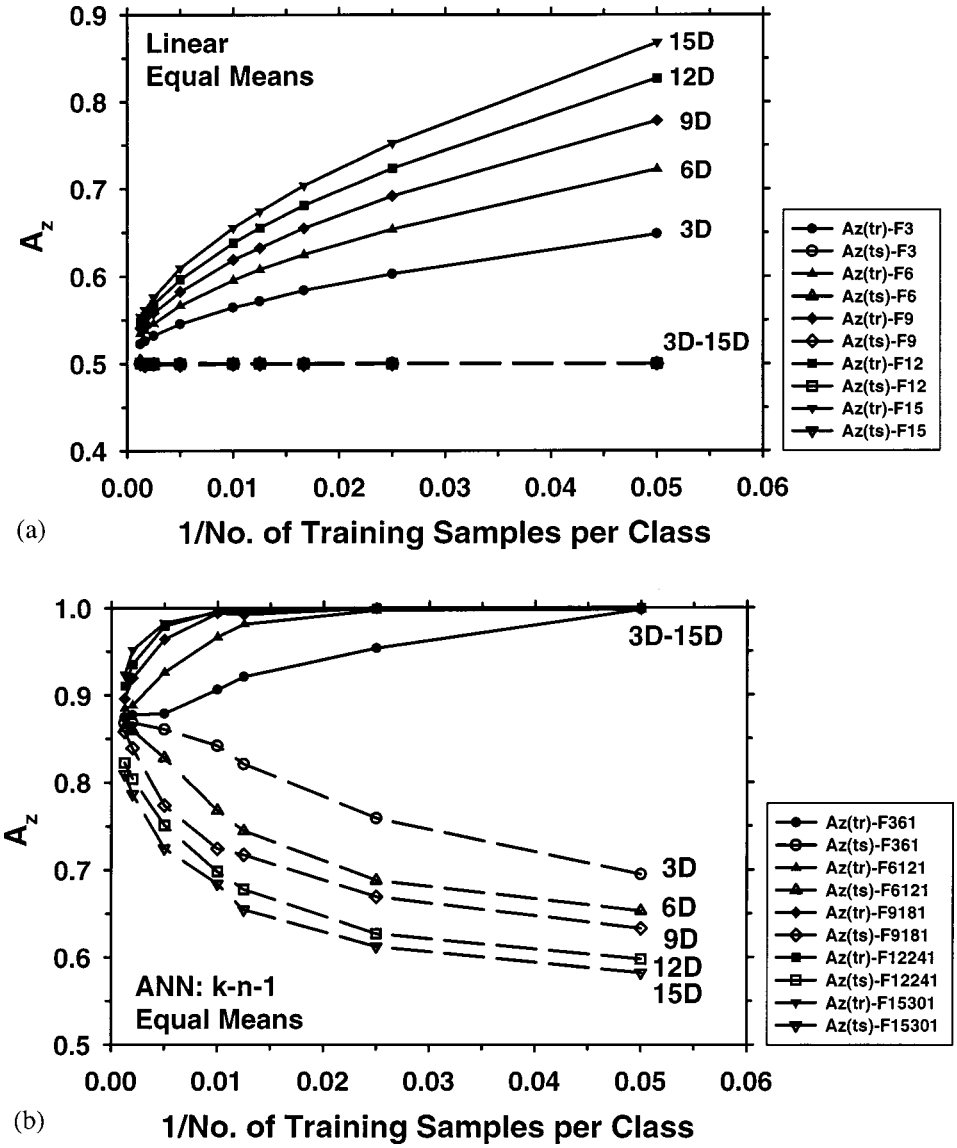


FIG. 12. The dependence of the performance curves on dimensionality of feature space for the class distributions with unequal covariance matrices and equal means. (a) Linear, (b) ANN classifier. Legend: F3=3D feature space, etc. F921 =ANN with two hidden nodes, etc. solid lines=A_z(tr), dashed lines=A_z(ts).

fier reaches the optimal value of 0.89 in the limit of large N for all dimensionalities studied.

Figure 11 shows a comparison of the performance of the linear, quadratic, and the ANN classifiers with two and nine hidden nodes. The biases on the resubstitution and the hold-out curves of the quadratic classifier are not as large as those of the ANN (9−9−1) classifier. However, in the regime of small design sample sizes, the hold-out curve of the optimal quadratic classifier can be much lower than the corresponding curves of the linear classifier or ANN with one or two hidden nodes. This result indicates that the theoretically optimal classifier may not be the optimal choice when the available design sample size is small and over-parametrization becomes an important consideration.

(3) Multivariate normal distributions—Unequal covariance matrices and equal means: Figure 12(a) shows the dependence of A_z on $1/N$ for the linear classifiers for the class distributions with equal means. Since the Mahalanobis distance is zero when the means of the two class distributions are equal, the linear classifier performs no better than random guessing in the hold-out situation (A_z(ts)=0.5). However, it is somewhat surprising that the resubstitution curve can be biased to very high A_z values, when the design sample is small. The bias increases with increasing dimensionality of the feature space because the severity of overfitting to the design samples worsens with increased parameterization in the linear discriminant function. This indicates that the predicted performance of a classifier can be unrealistically optimistic if the test samples are not independent of the design samples.

For the class distributions with equal means, it is much more difficult to train the ANN classifier. The number of hidden nodes and the number of training epochs required for the ANN to approximate the decision surfaces, which are spherical hypersurfaces in the k-dimensional feature space, increase as k increases. Figure 12(b) shows the A_z-vs-$1/N$ curves for the ANNs in which the number of hidden nodes is 2 times the dimensionality of the feature space. The number of training epochs required to approach the highest perfor-

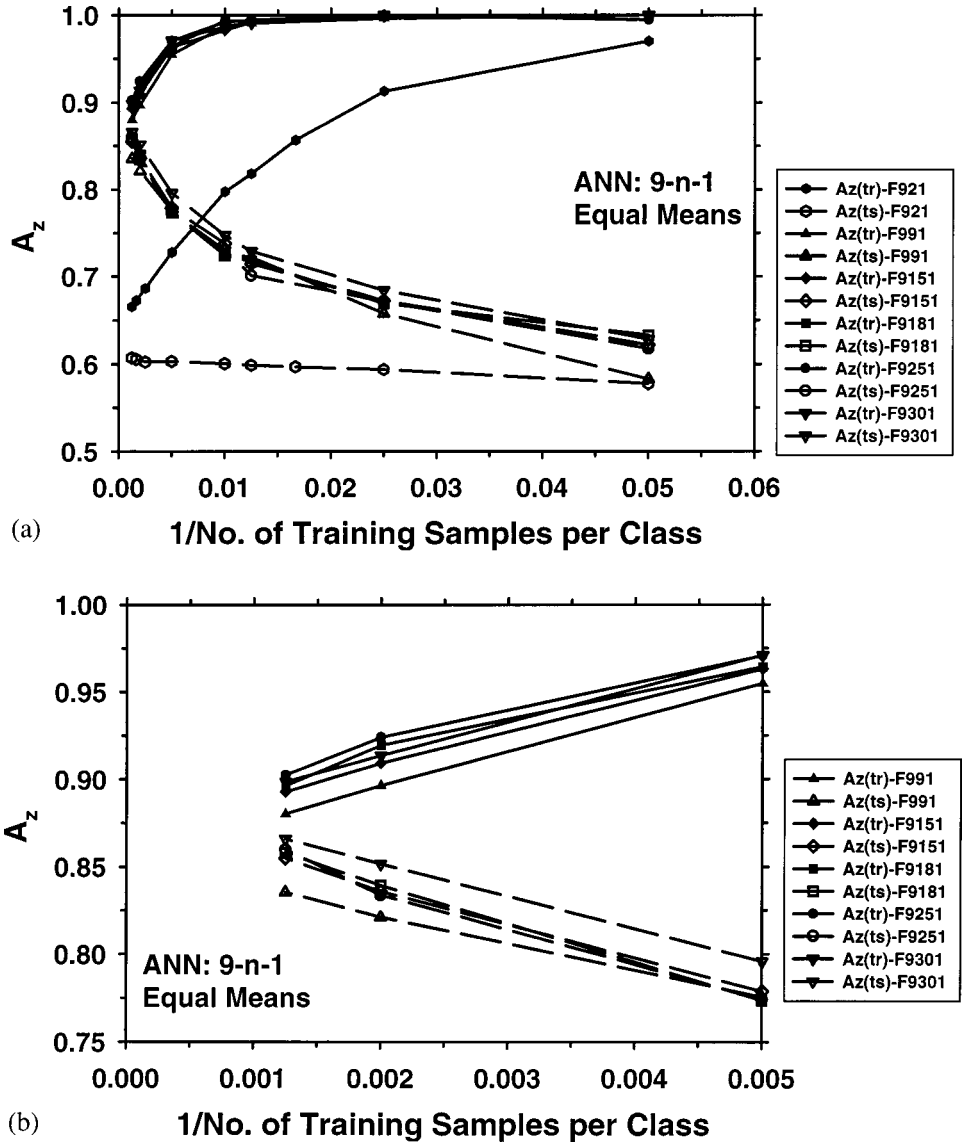


FIG. 13. (a) The dependence of the performance curves of an ANN on the number of hidden nodes in the 9D feature space for class distributions with unequal covariance matrices and equal means. In the expanded scale (b), the approximately linear regions of the curves can be observed. Solid lines $= A_z$(tr), dashed lines $= A_z$(ts).

mance for a given ANN architecture ranges from about 1800 to 20 000 in these cases. Again we did not attempt an exhaustive search for the ''optimal'' number of hidden nodes in each case. These ANNs were chosen because they appear to approach the maximum performance of $A_z = 0.89$ in the limit of large N and their number of hidden nodes is a simple multiple of the dimensionality. Compared to the class distributions with unequal means, for a given dimensionality, the number of hidden nodes and the number of training epochs required for achieving the near maximum performance at large N are greater in this equal-mean situation. Figure 13(a) shows an example of the dependence of the performance curves on the number of hidden nodes in the 9D feature space. Figure 13(b) is an enlarged view of the curves in Fig. 13(a) in the range where the sample size is greater than 200 per class. The hold-out performance of ANN(9−9−1) at $1/N = 0$ reaches about 0.85. When the number of hidden nodes is greater than nine, the performances of the ANNs at $1/N = 0$ are similar and approach the optimal A_z.

The quadratic discriminant is again the theoretically optimal classifier for the class distributions with unequal covariance matrices. Its performance curves (not shown) are very similar to those plotted in Fig. 7(c), except that the extrapolated A_z values at $1/N = 0$ do not reach as high as those in the equal covariance matrices situation. By using the approximately linear region of the A_z-vs-$1/N$ curve at N greater than 100, the extrapolated A_z ranges from about 0.873 to 0.885 for the 3D to 15D feature spaces. In this case, it is much more efficient to train a quadratic discriminant than the ANN. Since the linear discriminant and ANNs with few hidden nodes cannot provide effective classification regardless of the design sample size, the quadratic discriminant is obviously the optimal classifier both in terms of performance and training efficiency.

(4) Checkerboard distributions: In a feature space with checkerboard class distributions, classification is difficult for many classifiers because of the disjoint clusters of samples belonging to the same class. We compared the three classifiers in such a situation by two examples. Figure 14 shows the performance curves of the three classifiers in a 2D feature

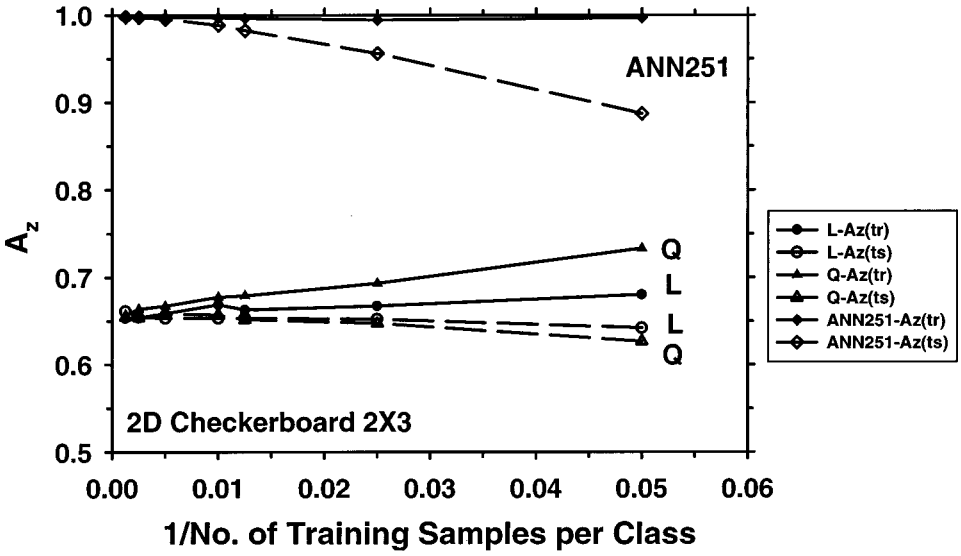


FIG. 14. Performance curves of the three classifiers for a 2×3 unit checkerboard in a 2D feature space. L=linear, Q=quadratic, ANN251=backpropagation neural network with five hidden nodes. Solid lines=A_z(tr), dashed lines=A_z(ts).

space with a 2×3 unit checkerboard distribution. Both the linear and the quadratic discriminants perform poorly even for the resubstitution method where A_z values are in the range of 0.6 to 0.7. However, the ANN(2−3−1) can achieve an A_z of 0.96 (not shown) and the ANN(2−5−1) a near-perfect classification at a training epoch of about 1200.

In a 3D feature space with a 2×2×2 unit checkerboard distribution, the difficulty in classification experienced by the linear and quadratic discriminants is even more apparent. Figure 15 shows that the hold-out curve of the linear classifier is basically the same as random guessing. The hold-out curve of the quadratic classifier is slightly higher than 0.5 at small design sample sizes but approaches 0.5 as the design sample increases. On the other hand, the ANN(3−3−1) can attain a test A_z of 0.9 (not shown) and the ANN(3−5−1) can reach near-perfect classification at large design sample sizes after about 1500 training epochs. These two examples demonstrate that an ANN classifier can be superior to the linear or quadratic classifiers for class distributions that are very different from the idealized multivariate normal distributions.

IV. DISCUSSION

Classifier design is an important field of research in computer-aided diagnosis. Yet many of the issues related to classifier design have not been explored systematically. This simulation study is a part of our on-going investigation of the sample size effects on classifier design.[7–11,15] In this study, we evaluated classifier performance for three multivariate normal class distributions with specific properties: equal covariance matrices, unequal covariance matrices, and equal means. These distributions are idealized but they do approximate a range of situations that may occur in real classification problems. Since the optimal classifier and the upper bound of classification accuracy in the limit of $1/N=0$ are

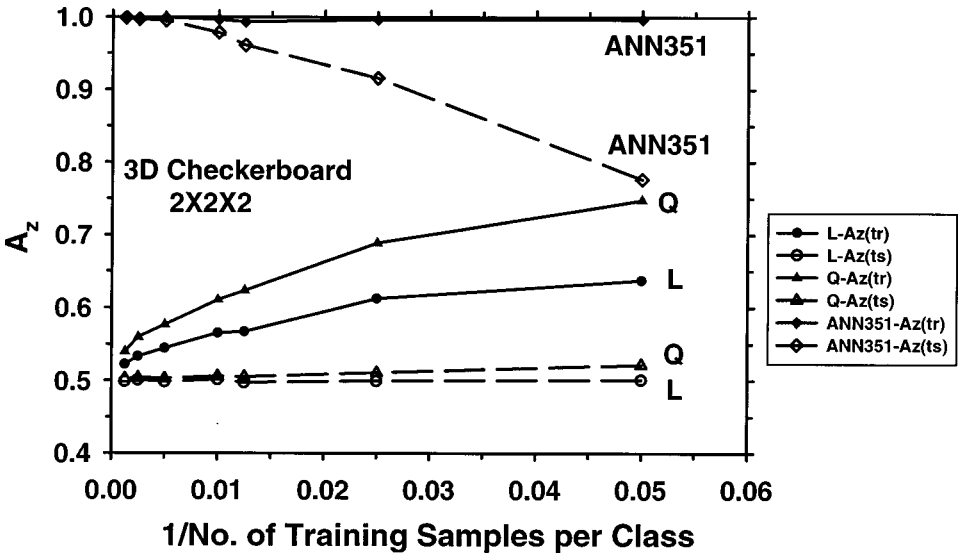


FIG. 15. Performance curves of the three classifiers for a 2×2×2 unit checkerboard distribution in a 3D feature space. Legend: L=linear, Q=quadratic, ANN351=backpropagation neural network with five hidden nodes.

known for each of these cases, we can compare the performances of the classifiers under each condition with the optimum. In addition, a checkerboard class distribution was included in the study. A comparison of the performances of the different classifiers for this class distribution can illustrate their effectiveness when the distributions are very different from multivariate normal.

For all three classifiers, the A_z(tr) obtained by resubstitution is biased optimistically while the A_z(ts) obtained by testing with an independent test set is biased pessimistically, relative to the A_z in the limit of $N \to \infty$, except for the situations when A_z(tr) is bounded from above by perfect classification ($A_z = 1$) or when A_z(ts) is bounded from below by random guessing ($A_z = 0.5$). The magnitude of the biases increases as the design sample size decreases and as the dimensionality of the feature space increases. In the cases where a given classifier has no discriminatory power for a given class distribution, for example, the linear discriminant for the equal-mean or checker-board class distributions, or the quadratic discriminant for the 3D checker-board class distribution, the test A_z(ts) remains almost constant at 0.5, independent of the design sample size. In many cases, the A_z-vs-$1/N$ curve cannot be approximated by a straight line that extrapolates to the A_z at $1/N = 0$ until the design sample sizes are very large, beyond the range of sample sizes that are generally available for CAD classifier design. To estimate the performance of a classifier at large N under the constraint of a small design sample, one may use the Fukunaga and Hayes resampling scheme[3] to derive several points along the A_z-vs-$1/N$ curves in the small sample size region. If the extrapolated resubstitution and hold-out curves do not converge to approximately the same A_z at $1/N = 0$, an average of the points on the two curves which correspond to the same design sample size may be a closer estimate of A_z than either A_z(tr) or A_z(ts). It may be noted that the resubstitution and the hold-out curves are not biased symmetrically from the A_z at infinite N, the average thus obtained will only be a rough estimate. It is also not valid in cases when the classifier has no discriminatory power with A_z(ts) constant at about 0.5 or when the resubstitution curve is overly optimistic with A_z(tr) constant at about 1.

In any case, caution should be taken in estimating classifier performance by extrapolation to $1/N = 0$ or by averaging the resubstitution and hold-out performance as discussed above. The estimated performance contains variances that have to be estimated using further tools. One such attempt in estimating the components of variance by a bootstrapping resampling scheme has been studied recently by Wagner *et al.*[11] These estimates reveal the amount of bias and variance in the classifier performance obtained with the finite design samples, thus allowing estimation of the sample size required to achieve a desired degree of generalizability, rather than replacing the need for a larger sample set and further studies.

With the equal-covariance-matrix class distributions, the linear discriminant is the optimal classifier as expected. The biases are low and the computation is efficient. Moreover, since the A_z-vs-$1/N$ relationship is linear over almost the entire range of design sample sizes, the classifier performance at very large N can be estimated from the small sample size performance by linear interpolation, as suggested by Fukunaga and Hayes[3] and demonstrated previously by Wagner *et al.*[9]

With the unequal-covariance-matrices and equal-mean class distributions, the linear discriminant and the back-propagation neural network with one hidden layer are inferior to the quadratic classifier when the design sample size is large. The linear discriminant cannot utilize the difference in the covariance matrices and underestimates the class separability even when an infinite number of design samples is available. The ANN needs a relatively large number of hidden nodes and a large number of training epochs in order to reach the optimal performance. Its hold-out performance and the computation efficiency are both inferior to those of the quadratic classifier. However, for the unequal-covariance-matrices and unequal-mean case and a small design sample size, the linear classifier or an ANN with very few hidden nodes, e.g., $n = 2$, provides better hold-out performance than the more complex ANNs or the optimal quadratic classifiers. These results indicate that the bias on classifier performance increases with increasing complexity (loosely related to the number of parameters to be estimated) of the classifier. The linear classifier contains $(k+1)$ independent parameters and the quadratic classifier contains $(k+1)(k+2)/2$ independent parameters in their formulations. The number of weights to be estimated for the ANN depends on the number of hidden nodes as $n(k+1)+(n+1)$. The number of weights in an ANN can therefore easily exceed that of a quadratic classifier, although the estimation of the mean and covariance matrices for the linear and quadratic discriminants may contribute additional ''complexity'' to the classifier design. Two observations can be made. First, when the available sample size is small, a simple classifier will have better generalization than a more complex classifier. Second, a complex ANN or a quadratic classifier trained with an insufficient number of design samples generalizes poorly, even if it is the optimal classifier for the class distributions. It is therefore important to select an appropriate classifier by taking into consideration the design sample size.

A further problem in classifier design is that the true population distributions of the classes in the feature space are generally unknown. It was suggested that the quantile–quantile (Q–Q) plot and the chi-square plot may be used for investigating the normality of univariate and multivariate sample distributions, respectively.[16] However, it is still unknown under what criteria the chi-square plot will indicate that it is optimal to use a classifier designed under the normality assumption. For any measure of goodness-of-fit, when the sample size is small, only the most aberrant deviations from the normal distribution can be identified as a lack of fit from these plots.[16] Therefore, there is often no *a priori* knowledge to select an ''optimal'' classifier or to predict whether the observed performance is caused by the sample size, the choice of an overly complex classifier, or by an actual poor separation of the classes in the feature space. If one observes poor generalization of a trained classifier in a

truly independent test set, it will be important to take into consideration all these factors and redesign the classifier.

In this study, we assumed that the best features have already been determined for the classification task. In a general classifier design problem, the best set of features usually has to be selected based on the available design samples. The feature selection step will introduce additional biases to the classifier performance. The number of features selected also has a strong influence on the classifier design, as can be seen from the dependence of the bias on the dimensionality of the feature space. The investigation of this more complex situation including both the feature selection and classifier training steps is underway.[17]

The term generalizability is nonspecific and needs to be qualified here. The present paper is concerned with the generalizability of the mean performance of classifiers to unknown test samples drawn from the same population of cases. We have shown in this paper that the mean performance of a classifier depends on the number of samples used to train the classifier, the architecture of the classifier, and—for multivariate-normal data—the means and covariances of the population distributions. Suppose in this context that a classifier is trained on a given finite number of design samples (patients). The mean performance of the classifier over independent replications with the same number of design samples is generalizable to studies characterized by the same number of design samples. In other words, the mean resubstitution or hold-out performance is an unbiased estimate for repeated sampling of independent design and test sample sets, respectively, when the same number of design samples is used. The classifier performance may not, however, be generalizable to studies characterized by a different number of design samples. In particular, when a very large and representative design sample size is used, the mean performance may be very different from the mean performance that characterizes the finite-training-sample condition. When the mean performance under the conditions of a finite design sample size is close to that expected with a very large design sample size, the finite-training sample performance is said to be generalizable to the population performance.

The term generalizability is not only used with respect to mean performance, it is also used with respect to uncertainty in performance, as reflected in estimates of error bars (standard deviations, or the corresponding variances). For example, if we think of repeating a given training and testing experiment on a classifier and if only the test samples are drawn independently on the repeated trials, then the estimated uncertainties are said to be generalizable only to a population of test samples. If, however, we think of repeating the experiment and independently drawing new training samples as well as new test samples, then the estimated uncertainties are said to be generalizable to a population of trainers and a population of testers.[17] Models for the components of variance in both paradigms are the subjects of current work in progress.[10,11] A key point of this latter work is the fact that for computer-aided diagnosis, most available software for ROC analysis only provides estimates of uncertainty that are generalizable to a population of test samples.

In this investigation, we have limited our study to only three types of classifiers: the linear discriminant, the quadratic discriminant, and the backpropagation ANNs with one hidden layer. There are, of course, many other variations of the ANN architecture and other parametric or non-parametric classifiers available for feature classification tasks. The purpose of our work is not to exhaustively evaluate all possible combinations of class distributions and classifiers. Rather, by limiting our investigation to some well-known situations, we can perform systematic analyses and gain some insights into the classifier design problems. Furthermore, we have limited our discussion here to the estimates of the mean classifier performance. Wagner *et al.*[10,11] have investigated the variances of classifier performance estimated from a finite sample set and developed models to study the relative importance of the sizes of the training and test samples. It has been demonstrated that a components-of-variance model can be estimated with a finite sample set by using a bootstrap method. More importantly, the analysis of variances can reveal the generalizability of the performance estimates to other training and test sample sets in the population. Our long term goals are to find some guidelines for designing efficient resampling schemes that can minimize the bias and variance of a trained classifier using the available samples, and to provide a quantitative design tool that can estimate the design sample size requirement for a larger ''pivotal'' study from the results of a smaller ''pilot'' study in order to achieve a desired precision in A_z and the desired generalizability.

ACKNOWLEDGMENTS

This work is supported in part by USPHS Grant No. CA 48129 and by a grant from the U.S. Army Medical Research and Materiel Command DAMD 17-96-1-6254, a Career Development Award (B.S.) DAMD 17-96-1-6012 from the U.S. Army Medical Research and Materiel Command and a Whitaker Foundation Grant (N. P.). The content of this paper does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned in this paper should be inferred. The authors are grateful to Charles E. Metz, Ph. D., for providing the LABROC1 programs.

[a]Author to whom correspondence should be addressed. Department of Radiology, University of Michigan, 1500 E. Medical Center Drive, UHB1 F510B, Ann Arbor, MI 48109-0030; Phone: 734-936-4357; Fax: 734-936-7948; Electronic mail: chanhp@umich.edu

[1]K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. (Academic, New York, 1990).

[2]S. Raudys and V. Pikelis, ''On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition,'' IEEE Trans. Pattern. Anal. Mach. Intell. **PAMI-2**, 242–252 (1980).

[3]K. Fukunaga and R. R. Hayes, ''Effects of sample size on classifier design,'' IEEE Trans. Pattern. Anal. Mach. Intell. **11**, 873–885 (1989).

[4] R. F. Wagner, D. G. Brown, J.-P. Guedon, K. J. Myers, and K. A. Wear, in *Information Processing in Medical Imaging*, edited by H. H. Barrett and A. F. Gmitro (Springer-Verlag, Berlin, 1993).
[5] R. F. Wagner, D. G. Brown, J.-P. Guedon, K. J. Myers, and K. A. Wear, ''On combining a few diagnostic tests or features,'' Proc. SPIE **2167**, 503–512 (1994).
[6] D. G. Brown, A. C. Schneider, M. P. Anderson, and R. F. Wagner, ''Effect of finite sample size and correlated/noisy input features on neural network pattern classification,'' Proc. SPIE **2167**, 180–190 (1994).
[7] H. P. Chan, B. Sahiner, R. F. Wagner, N. Petrick, and J. Mossoba, ''Effects of sample size on classifier design: quadratic and neural network classifiers,'' Proc. SPIE **3034**, 1102–1113 (1997).
[8] H. P. Chan, B. Sahner, R. F. Wagner, and N. Petrick, ''Effects of sample size on classifier design for computer-aided diagnosis,'' Proc. SPIE **3338**, 845–858 (1998).
[9] R. F. Wagner, H. P. Chan, B. Sahiner, N. Petrick, and J. T. Mossoba, ''Finite-sample effects and resampling plans: applications to linear classifiers in computer-aided diagnosis,'' Proc. SPIE **3034**, 467–477 (1997).
[10] R. F. Wagner, H. P. Chan, J. T. Mossoba, B. Sahiner, and N. Petrick, ''Components of variance in ROC analysis of CADx Classifier performance,'' Proc. SPIE **3338**, 859–875 (1998).
[11] R. F. Wagner, H. P. Chan, B. Sahiner, N. Petrick, and J. T. Mossoba, ''Components of variance in ROC analysis of CADx classifier performance. II: Applications of the bootstrap,'' Proc. SPIE **3661**, 523–532 (1999).
[12] D. J. Hand, *Discrimination and Classification* (Wiley, New York, 1981).
[13] P. A. Lachenbruch, *Discriminant Analysis* (Hafner, New York, 1975).
[14] J. A. Freeman and D. M. Skapura, *Neural Networks-Algorithms, Applications, and Programming Techniques* (Addison-Wesley, Reading, 1991).
[15] H. P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, ''Classifier design for computer-aided diagnosis in mammography: effects of finite sample size,'' Med. Phys. **24**, 1034–1035 (1997).
[16] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis* (Prentice-Hall, Englewood Cliffs, NJ, 1982).
[17] C. A. Roe and C. E. Metz, ''Variance-component modeling in the analysis of receiver operating characteristic index estimates,'' Acad. Radiol. **4**, 587–600 (1997).