A General Model for Finite-Sample Effects in Training and Testing of Competing Classifiers

Sergey V. Beiden, Marcus A. Maloof, *Member, IEEE,* and Robert F. Wagner, *Fellow, IEEE*

***Abstract*— The conventional wisdom in the field of statistical pattern recognition (SPR) is that the size of the finite test sample dominates the variance in the assessment of the performance of a classical or neural classifier. The present work shows that this result has only narrow applicability. In particular, when competing algorithms are being compared, the finite training sample more commonly dominates this uncertainty. This general problem in SPR is analyzed using a formal structure recently developed for multivariate random-effects receiver operating characteristic (ROC) analysis. Monte Carlo trials within the general model are used to explore the detailed statistical structure of several representative problems in the sub-field of computer-aided diagnosis in medicine. The scaling laws between variance of accuracy measures and number of training samples and number of test samples are investigated and found to be comparable to those discussed in the classic text of Fukunaga, but important interaction terms have been neglected by previous authors. Finally, the importance of the contribution of finite trainers to the uncertainties argues for some form of bootstrap analysis to sample that uncertainty. The leading contemporary candidate is an extension of the 0.632 bootstrap and associated error analysis, as opposed to the more commonly used cross-validation.**

***Index Terms*— pattern recognition, classifier design and evaluation, discriminant analysis, ROC analysis, components-of-variance models, bootstrap methods.**

I. Introduction

The conventional wisdom in the field of statistical pattern recognition (SPR) regarding the uncertainties in estimates of the performance of classifiers trained and tested with a finite number of samples may be summarized as follows: The *bias* of measures of performance comes only from the finite size of the training sample; the sampling *variance* (and thus the error bars) of these measures comes mainly from the finite number of test samples [1, p. 218]. The validity of this wisdom regarding the bias of accuracy measures is well established [1]–[7]. The purpose of the present paper is to investigate the issue of the variance using a general multivariate statistical model.

Analysis of the variability in estimates of the performance of classifiers trained and tested with a finite number of samples has been provided by Fukunaga and Hayes [2], [3], and extensive reviews of the problem have been given by Fukunaga [1], Raudys and Jain [4], and Jain et al. [5], with further elaboration by Raudys [6]. The focus of this analysis was on the case where only a single classifier is under evaluation. In the present paper we provide a more general treatment for the case of either a single or of two competing classifiers. We shall see in the light of the present work that some previous summary rules of thumb regarding the variability of estimates have only a narrow range of applicability.

Results of any statistical assessment must be accompanied by a statement of the level of "generalizability" of the results. We define the levels of generalizability in the context of SPR as follows. If performance is estimated in such a way that the effect of the finite number of test samples is explicitly accounted for in the analysis, but not the effect of the finite number of training samples, one says that performance estimates "generalize only to a population of testers." The mean and error bars obtained from such analysis are estimates of the range of performance expected if the experiment is repeated many times, drawing independently from a population of testers on each replication but without varying the training. If performance is estimated in such a way that the finite number of training samples as well as the finite number of test samples is accounted for in the analysis, one says that performance estimates "generalize to a population of trainers and a population of testers." The uncertainties are then estimates of the range of performance expected if the experiment is repeated many times, each time drawing independently from a population of trainers as well as testers.

An analogous problem and set of issues have been the subject of contemporary research in the field of medical imaging [8]–[11]. This problem gave rise to the development of the field of random-effects (or multivariate) receiver operating characteristic analysis. In the next section we will provide the general random-effects model for the context of SPR. In following sections we analyze several problems that display the general structure of this problem and compare the results with previous expectations. The results have implications for the methods of resampling used in uncertainty analysis in SPR that we discuss in the concluding section.

II. Random-effects ROC Analysis

Receiver operating characteristic (ROC) analysis (or simply "operating characteristic" analysis [1]) is a general approach to the assessment of systems for binary classification, i.e., where the task is to assign a sample to one of two classes which,

S.V. Beiden and R.F. Wagner are with the Center for Devices & Radiological Health, Food & Drug Administration, Rockville, MD 20857. E-mail: svb@cdrh.fda.gov, rfw@cdrh.fda.gov.

M.A. Maloof is with the Department of Computer Science, Georgetown University, Washington, DC 20057-1232. E-mail: maloof@cs.georgetown.edu.

The authors are listed alphabetically.

Manuscript received February 20, 2002; revised July 2, 2003.

for definiteness, we shall refer to as the abnormal vs. the normal class [12], [13].[1] The ROC curve describes the trade-off between the true-positive fraction (TPF), i.e., the percent correct on the actually abnormal cases, and the false-positive fraction (FPF), i.e., the percent incorrect on the actually normal cases. The normalization of these fractions makes them independent of the prevalence of actually abnormal or actually normal cases; thus, the ROC curve itself—in contrast to measures such as probability of misclassification (PMC [1]–[3])—is also independent of these prevalences. The commonly used summary measures of ROC performance include the area under the entire ROC curve (i.e., the TPF averaged over all FPFs) [12] and also the partial areas under the portions of the curve above a specified TPF or below a specified FPF [14]. Although we shall use the total area to exemplify the approach in this paper, the statistical structure of the problem considered here does not change if partial areas are considered, or if the TPF at a given FPF (or vice versa) or even PMC is considered. There is a rich literature on estimation of ROC measures of accuracy and their uncertainties (reviewed in [8], [12], [13], [15]) and validated software for these tasks is available on the Web [16].

A. The General Random-Effects Model

Random-effects (or multivariate) ROC analysis is a solution to the assessment problem when multiple random effects contribute to the uncertainty in performance analysis. It has become the contemporary standard in medical imaging assessment where two obvious random effects are those due to the variability in difficulty and finite sampling of patient cases or images, and the variability due to the range of skill of the readers of images. Note for our present purposes that (aside from subjective factors) the variability of reader skill in imaging follows from the limitation and variability of their finite training. The problem in SPR is thus in one-to-one correspondence with that in imaging. In SPR the random effects are those due to the variability in difficulty and finite sample of the test cases, and the variability due to the range of difficulty and finite sample of the training cases. It might seem that the presence of two random effects would require a model of uncertainty with two terms. However, even for the task of assessing a single classifier, the most general model requires three terms. This follows since it is necessary to allow for the possibility that the range of difficulty of the test cases may depend on the range of difficulty of the training cases—a so-called "interaction" effect or cross-term. When comparing two classifiers it is also necessary to include the interactions of the previous three effects across classifiers, leading to the requirement for a model with six terms that we now describe.

The model we discuss here is referred to as a components-of-variance model [17]. The indexing of variables and their interactions in the model is the most general one for the problem of random effects of training and testing and the fixed effects of competing classifier architectures. The model contains all terms with one, two, or three indices relevant to this problem. For any specified accuracy measure from those listed in Section I, denoted here generically as A, the model can be written for the SPR problem as (compare [8], [9], [17]):

$$\begin{aligned} A_{ijk} \quad = \quad & \mu_i + (tr)_j + (ts)_k + (tr{\cdot}ts)_{jk} + \\ & (m{\cdot}tr)_{ij} + (m{\cdot}ts)_{ik} + (m{\cdot}tr{\cdot}ts)_{ijk} \quad . \end{aligned} \tag{1}$$

Here, i indicates a particular classification algorithm, j denotes a particular sample training set, and k is a particular sample test set. The term μ_i represents the contribution of classifier i to the expected value of the accuracy index, while the remaining terms are independent zero-mean random variables—referred to as components of variance. (NB: No assumptions of normality are required for the approach of [9] that we follow here.) The terms with a single index are the pure training- and pure testing-sample contributions to the variability, with variances σ^2_{tr} and σ^2_{ts} respectively. The terms with two subscripts represent two-way interactions between training and test set, classifier and training set, and classifier and test set, with variances $\sigma^2_{tr{\cdot}ts}$, $\sigma^2_{m{\cdot}tr}$, and $\sigma^2_{m{\cdot}ts}$ respectively. (The letter m comes from the imaging literature where it stands for *modality*, modality here being the classifier.) The term with three subscripts represents the three-way interaction among classifier, training set, and test set, with variance $\sigma^2_{m{\cdot}tr{\cdot}ts}$. There are then six components of variance in this general model.

The magnitudes of the components of variance can be understood as follows. The variance strength σ^2_{tr} reflects the range of training case difficulty and the finite size of the training case set. The variance strength σ^2_{ts} reflects the range of test case difficulty and the finite size of the test case sample. The variance strength for the training-test interaction, $\sigma^2_{tr{\cdot}ts}$, is high (or low) depending on whether the sampled range of test case difficulty depends strongly (or weakly) on the training case difficulty (or vice versa, by the symmetry of this term). The components of variance measured by these three variance strengths in the model are perfectly correlated across the classification algorithms or modalities under comparison, since these components have no index for modality (i.e., they are unchanged across modalities). The strengths of the remaining three components of variance—$\sigma^2_{m{\cdot}tr}$, $\sigma^2_{m{\cdot}ts}$, and $\sigma^2_{m{\cdot}tr{\cdot}ts}$—correspond to analogous components that are completely uncorrelated across classifiers in the model, since these components have an index for modality. The fact that there are contributions that are correlated and uncorrelated across classifiers allows for a flexible correlation structure depending on their relative strengths.

Beiden, Wagner, and Campbell (BWC) pointed out that for the analogous random-effects problem in imaging [9] it is possible in principle to perform a family of six different experiments on the population of readers and patients that will allow one to solve for the strengths of the six model variance components. Their approach translates to an analogous set of experiments that are possible in principle for the SPR problem. In SPR, one may intend to draw both testers and trainers randomly on replication of the experiment; in this case testers and

[1] Rigorous generalization of the ROC paradigm to the problem of three (or more) classes is greatly complicated by the fact that six (or more) independent measures and their mutual trade-offs must be analyzed. A general approach to this problem has eluded investigators for many decades; the default condition has been to reduce the more general problem to a series of two-class problems (i.e., one class at a time versus all of the others).

trainers are said to be random effects. Or one may intend to hold trainers fixed and draw testers randomly on replication of the experiment; in that case trainers are said to be a fixed effect and testers are said to be a random effect. One may look at one classification algorithm at a time, or one may consider the difference in performance between two competing classifiers—with trainers fixed or random. Different (unobservable) model variance components contribute to the (observable) experimental variances depending on the experiment. A well-conditioned system of six linear equations describes these experiments (see [9], translated into the SPR paradigm for completeness in Appendix I) and can be directly inverted to yield the model variance components from the six observed variances. In the practical world of a finite set of available samples, BWC replace the population experiments with bootstrap resampling [18] of the finite data set, obtain bootstrap estimates of the six observable variances, and then obtain finite-sample estimates of the six model variance components by solving the system of equations. The approach yields distribution-free maximum-likelihood estimates in the sense of Efron and Tibshirani [18]. The finite-sample uncertainties in these estimates of the model variances may then be obtained using the technique of the jackknife-after-bootstrap [11], [18]. (Appendix II contains an application of the technique as a cross-check on the work presented below.)

B. Monte Carlo Simulations

To obtain insight into the structure of the random-effects problem in the context of statistical pattern recognition, we analyze the following construct. We suppose that we have many institutions, each with its own independent set of *training* samples. (Thus, the institutions are analogous to readers in imaging.) For tractability we limit our simulations to ten institutions. We suppose that each institution designs two competing classifiers using its own training patients. For our simulations and to simplify the interpretation of results, we consider the special case where the two classifiers are constrained to have the same architecture across institutions. Finally, we suppose that some outside institution has provided a single independent set of *test* samples to be classified by each of the training institutions. (Variations on this experimental paradigm are possible; the present one has been found to be the most statistically powerful for probing the variables of interest here.) We perform 300 Monte Carlo simulation trials of this exercise for a number of simple but very instructive problems in which it is desired to compare competing classifiers. We solve for the strengths of the six components of variance using methods developed for the imaging problem and adapted to the SPR problem in Appendix I.

C. The Tasks and the Classifiers

The problems selected for the present analysis were chosen because they bracket a range of practical problems in our sub-field of statistical pattern recognition, namely, computer-aided diagnosis (CADx) in medicine. All of the problems used a feature space of nine-dimensions, with the features taken to be independent and normally distributed. Although the parameters chosen for the present analysis are inspired by contemporary problems in the CADx field, we emphasize that the approach and model are completely general.

We studied two levels of mean class separability: Mean ROC areas equal to 0.88 and 0.76, corresponding to signal-to-noise ratios or so-called detectability indexes, d', of 1.66 and 1.0, respectively. (The quantity d' (squared) is equivalent to the Mahalanobis distance (squared) [6].) Roughly speaking, the former value might be considered typical of mature CADx modalities, the latter more typical of research work in that field. Competing classifiers in this work had comparable performance in the mean. (The "donut" problem below was an exception.) Our focus here is on the components of variance.

We considered the cases where a total of 125, 250, 375, or 500 samples were available for each of the two classes. The lower half of this range corresponds to typical values available in the CADx community. The larger numbers were included to discover the dependence of the results on sample numbers. We made the ratio of training samples to test samples equal to 4:1 (but scaling laws deduced below provide broader coverage). For example, the 125 were partitioned into 100 and 25; i.e., each institution provided its own independent training set of 100 normal and 100 abnormal cases, and an independent outside group provided the common test set comprised of 25 normal and 25 abnormal cases (and similarly for the other total values called out above). Since the test samples were always independent from the training samples, we have what one could call a multi-institutional version of the *holdout* method [1].

We simulated three different tasks which bracket a range of practical problems in CADx (see [19] for formal definitions of classifiers):

- Task (a): The naive Bayes classifier vs. the nine- (or nineteen-) nearest-neighbor classifier for linearly separable data;
- Task (b): The naive Bayes classifier vs. the quadratic discriminant for linearly separable data;
- Task (c): The linear classifier vs. the quadratic discriminant for the so-called "donut" problem where the means of the two classes are identical but the variance of one class is greater than that of the other class; thus for this task there is no linear separability of the data but there is separability with the quadratic classifier.

In all experiments (except the "donut" problem) the distribution of the "normal" class is centered at the multivariate zero and that of the "abnormal" class at the multivariate unity. The value of $d' = 1.0$ (alternatively 1.66) is obtained by setting the population covariance matrix for each class to be diagonal with all elements equal to 9 (alternatively, 3.266). For the "donut" task, both distributions are centered at the origin; the covariance matrices are taken to be proportional to the identity matrix, with the ratio between proportionalities of 2.7. For this problem the mean area under the ROC curve for the linear classifier is 0.5 (guessing) and for the quadratic classifier is 0.9.

For each task, both classifiers were trained at the "ten institutions" with their own independent training sets, and then tested on the common test set as above. On each Monte Carlo

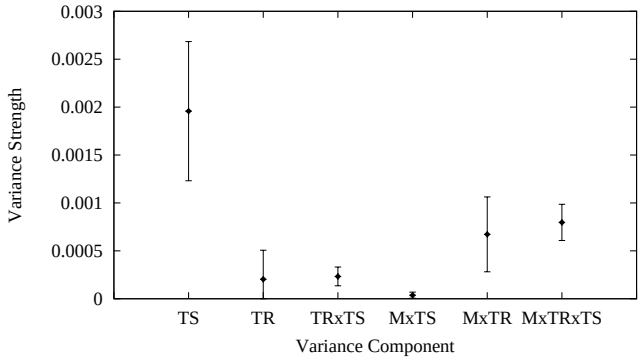

Fig. 1. Mean ($\pm$ std. dev. over 300 Monte Carlo trials) variance strength of the six variance components for Task (a), naive Bayes versus 9-NN, $d' = 1.66$, $N_{train} = 100$ per class, $N_{test} = 25$ per class.

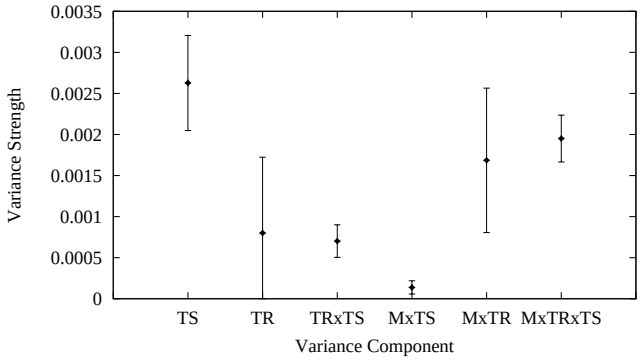

Fig. 2. Mean ($\pm$ std. dev. over 300 Monte Carlo trials) variance strength of the six variance components for Task (a), naive Bayes versus 9-NN, $d' = 1.0$, $N_{train} = 100$ per class, $N_{test} = 25$ per class.

trial, the BWC analysis was run to obtain estimates of the six variance components. In the next section we present the mean estimates, $\pm$ one standard deviation over the 300 trials.

III. Results

In Fig. 1 we present the results for Task (a) with nine nearest neighbors (9-NN), $d' = 1.66$, and 125 patients per class. For this example, the variance strengths are seen to be dominated by the finite-test-set component (TS), i.e., the component due to the range of difficulty of the test samples and the finite-test-sample size. (Capital letters are used in the figures to label the variance strength of the corresponding lower-case component.) Recall that if we are interested in the assessment of the classifiers one at a time, all six of the components of variance in the figure contribute according to the general linear model [8], [9]. Thus, we see here an example of the conventional wisdom that the sampling variance is dominated by the finite test sample [1]. Note, however, that the test-sample size is very small in this example (25 per class). We will return to this issue below.

When one is interested in comparing the performance of two competing classifiers, however, it is only the so-called modality-interaction components—the last three in the figure—that contribute uncertainty in the general model because these are the only components that are independent across classifiers [8], [9]; this is formalized in the subscript notation with the presence of the classifier index i in Equation 1. (The first three components are identical across classifiers and thus do not add randomness to the task of seeing a difference between them.)

We see in the figure that the modality-by-test-set component (M$\times$TS)—i.e., the part of the test-case-sample variance that is uncorrelated across algorithms—is negligible for this task. It is only the modality-by-training-set (M$\times$TR) and modality-by-training-by-test-set components (M$\times$TR$\times$TS) that contribute. We thus see that the limitation of the finite size of the training sample clearly dominates that of the finite size of the testing sample for the task of comparing classifiers. Although this seems intuitively reasonable or even obvious, we have not found previous observations concerning this critical point in the literature.

Most investigators simply study the case where the training set is a fixed effect, i.e., the same trainers are to be used on replications of the experiment. The analysis of such an experiment yields results that are only generalizable to a population of testers for a given fixed training set. In that case the TR and M$\times$TR components are not sampled and are effectively treated as zero. The example above demonstrates that uncertainties obtained in that paradigm can greatly underestimate the uncertainties expected in the more general experiment in which trainers are also a random effect.

This exercise was for the case where the number of trainers was four times the number of testers. The case where the number of testers is increased to be equal to the number of trainers—holding the latter fixed—will be examined below.

We next look at Task (a) above for the case where the intrinsic class separability is only modest ($d' = 1.0$, mean ROC area = 0.76); the results are shown in Fig. 2 and we now see a more elaborate picture. For the problem of assessing an individual classifier—where all six components contribute—the predominance of the finite-test-set component (TS) has almost disappeared. It now shares the spotlight with several other components. In this case, the conventional wisdom that the variance is due mainly to the finite test set fails again.

For the problem of comparing the two competing classifiers—where only the last three components of variance are relevant—we see the same qualitative structure as we saw above for the case of greater class separability, but now with greater variance strength as expected because of the lower signal-to-noise. Variability in the finite test set is again not the dominant contribution to the masking of the difference between competing classifiers.

Similar results were found for Task (b) and also when Task (a) was repeated with nineteen rather than nine nearest neighbors. The major differences between Tasks (a) and (b) will be recorded below in the section on scaling of results. Further details and many more examples are available in a technical report [20].

Task (c)—the "donut" problem with parameters as above—provides an interesting variation on the above theme. Here we compared the linear with the quadratic classifier. The former offers no class separability for this configuration, whereas

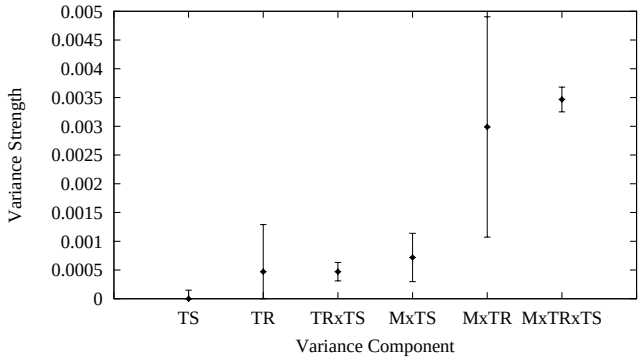

Fig. 3. Mean ($\pm$ std. dev. over 300 Monte Carlo trials) variance strength of the six variance components for Task (c), linear versus quadratic discriminant, equal class means, variance of abnormal class = 2.7 $\times$ variance of normal class, $N_{train} = 100$ per class, $N_{test} = 25$ per class.

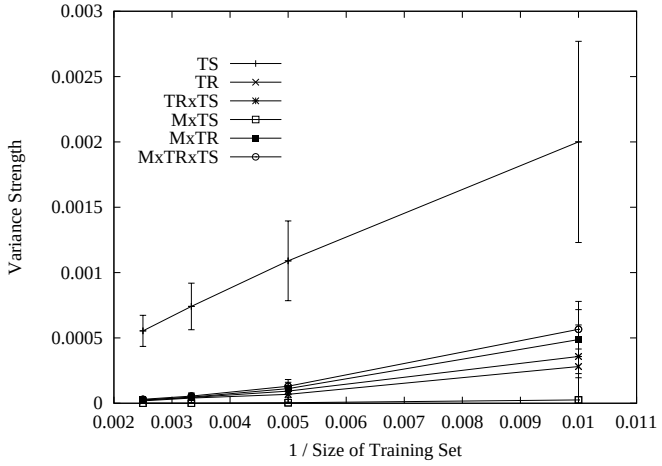

Fig. 4. Mean ($\pm$ std. dev. over 300 Monte Carlo trials) variance strength versus $1/N_{train}$ for naive Bayes versus quadratic discriminant, $d' = 1.66$. Test-set component TS (and M$\times$TS component which is zero) is consistent with this linear dependence; all others are inconsistent (cf. Fig. 5). (Note that N_{train} and N_{test} are proportional for these simulations.) Component labels as in other figures.

the latter is the appropriate Bayes classifier. We present the results in Fig. 3. The roles played by the pure test-set and the modality-by-test-set components have now reversed, compared to the examples in Figs. 1 and 2. Now there is no strength in the pure test-set component, whereas the modality-by-test-set component carries significant strength. This means the component of variance due to testers is completely uncorrelated across modalities—which might have been expected since the linear classifier is essentially guessing with this task whereas the quadratic classifier has the optimal strategy. The two rightmost of the uncorrelated components of variance across modalities have comparable strength. Again, the finite test sample does not dominate the variance analysis for this problem. Although this is an extreme example, it sheds light on the practical problem where one does not in fact know what is the optimal classifier.

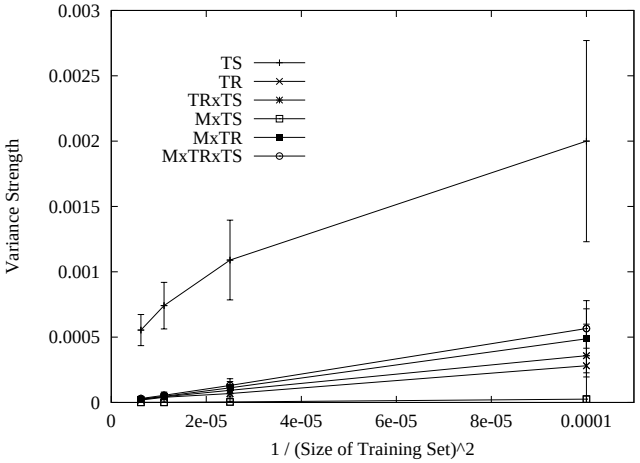

Fig. 5. Mean ($\pm$ std. dev. over 300 Monte Carlo trials) variance strength versus $1/N_{train}^2$ for naive Bayes versus quadratic discriminant, $d' = 1.66$. Test-set component TS is inconsistent with this dependence (also M$\times$TS which is zero); all others are consistent with a quadratic dependence (cf. Fig. 4). (Note that N_{train} and N_{test} are proportional for these simulations.) Component labels as in other figures.

A. Dependence on Sample Numbers

We have studied the scaling of all of the above results as a function of the number of trainers per class, N_{train}, and the number of testers per class, N_{test}. In Fig. 4, we show the dependence of the six variance components on the inverse of N_{train} for Task (b) with $d' = 1.66$. In Fig. 5, we show the dependence of these components on the inverse of N_{train}^2. Since N_{train} and N_{test} were made proportional to each other in the present work, the power-law dependence of curves in these figures will hold for either N_{train} or N_{test}. Fig. 4 shows that only the pure test-set component and the modality-by-test-set component (the latter trivially zero here) exhibit a linear dependence on the number of samples. Fig. 5 shows that the remaining four components exhibit an approximately quadratic dependence on the number of samples. We note that the error bars in these figures (as in Figs. 1 through 3) are standard deviations over the 300 Monte Carlo trials. Thus, the mean values shown in those figures are known approximately to within those error bars diminished by the root of 300. The mean dependences indicated by the lines therefore have very little uncertainty.

These dependences are interpreted as the dominant lowest-order Taylor series terms in an expansion of the total variance in two variables, N_{train} and N_{test}. *A priori*, one would expect the TR, TS, M$\times$TR, and M$\times$TS terms to have the inverse linear dependences on the appropriate sample numbers that are typical of variances in statistical estimation theory because they contain a single random variable; the TR$\times$TS and M$\times$TR$\times$TS terms would be expected to have inverse quadratic dependences because they contain two random variables—which scale together in the present simulations. Fukunaga and Hayes [1]–[3] have pointed out, however, that the first-order term in the numbers of trainers—realized here by the TR and M$\times$TR terms—goes to zero for the case of Bayes classifiers with normally distributed data; for those cases the leading terms are then quadratic in the number of trainers. This is

TABLE I
SCALING OF RESULTS WITH N_{train} AND N_{test} FOR TASKS (A) AND (B).

TS	$\sim$	$1/N_{test}$
TR	$\sim$	$1/N_{train}^2$ (plus*) $1/N_{train}$
TR$\times$TS	$\sim$	$1/N_{train}N_{test}$
M$\times$TS	$\sim$	$1/N_{test}$
M$\times$TR	$\sim$	$1/N_{train}^2$ (plus*) $1/N_{train}$
M$\times$TR$\times$TS	$\sim$	$1/N_{train}N_{test}$

*Admixture of the linear term accompanies a decrease in class separability.

consistent with what we observe in Fig. 5 when comparing two classifiers that are both approximations to the Bayes classifier for Task (b). We note that it is this quadratic dependence on the number of trainers that has provided the basis for the conventional wisdom that the contribution of the finite training sample to the variance is small. That wisdom ignores departures from its assumptions and the relative strengths of the Taylor terms that then result.

We exhibit a summary of the scaling behavior of the mean results observed in our simulations for Tasks (a) and (b) in Table I. (For Task (a) the number of nearest neighbors was made to scale with the overall sample size.) The scaling did not reduce to a single term for those training-sample-related terms marked with an asterisk (*) as the class separability was reduced from $d' = 1.66$ to $d' = 1.0$. The dominant behavior is quadratic at the higher signal-to-noise but admixture of a linear contribution begins to emerge as the signal-to-noise decreases; this is manifested by roughly three-fold as opposed to four-fold scaling as the sample numbers are changed by a factor two. Thus, a single Taylor term is in general insufficient to describe those training-set-related terms, as foreshadowed by the analysis of [1]–[3].

The scaling for Task (c) departs from the pattern exhibited in Table I mainly in that the M$\times$TR$\times$TS term now scales linearly (rather than quadratically) for the several cases we have studied, while the TR$\times$TS term scales closer to a quadratic dependence; the same behavior was observed, respectively, for the M$\times$TR and TR terms. Thus as the number of samples increases, the uncorrelated-across-modalities M$\times$TR$\times$TS and M$\times$TR terms will dominate their correlated-across-modalities counterparts (TR$\times$TS and TR). This is consistent with the fact that the training of a classifier that is essentially guessing decorrelates asymptotically from the training of an optimal one. Finally, we note that for all tasks and parameters we have investigated, the dependence of the TS and M$\times$TS terms is unambiguously linear in the inverse of the number of testers, as expected.

Since the scaling law for test-set contributions is always inverse-linear, we can use this law to generate good estimates of the mean results for situations other than the four-to-one ratio between trainers and testers considered above. In particular, if we are interested in the case where the number of testers is increased to be equal to the number of trainers, we can reduce by a factor four all of the test-set (TS containing) contributions in the previous analysis displayed in Fig. 1. This simple operation can be carried out by inspection of the mean results in that figure to obtain expected results when 100 trainers and 100 testers per class are used. It is then readily appreciated that almost all of the variability would come from the finite size of the training set for that case. So, the conventional wisdom would also be greatly misleading here. The paradigm of this example has special standing because the equality of the size of the training and test sets means the results are effectively normalized per unit case, independent of whether the case is a trainer or a tester.

The analysis of uncertainties in terms of components of variance may be used to design large trials based on the results of smaller trials and scaling laws, as has been previously pointed out in the imaging problem [9]. Similarly in SPR, it may be used to design and size a database for training and testing of competing classifiers [21], as suggested in [22]. Regarding the entries in Table I that show a mixture of linear and quadratic dependences (marked with the asterisk *), the conservative approach would be to assume a linear scaling.

IV. DISCUSSION AND CONCLUSIONS

We have presented here an application to the problem of statistical pattern recognition of a general components-of-variance model developed recently for random-effects ROC studies in medical imaging. Our examples and analysis show that one cannot assume that the variance of accuracy measures comes mainly from the finite test set; this is true only for a limited class of problems, namely: where only a single classifier is being studied at a time; where the finite test set is almost obviously the limiting factor due to the small test sample size; and only at the high-performance end of the signal-to-noise ratio scale. In that limited case, the finite-test sample may indeed make the dominant contribution to the variance or overall uncertainty in the assessment. Outside that context, however, the finite-training sample contributes comparable or greater strength to that variance. In particular, if we are interested in comparing competing classifiers, the contribution to the uncertainty in the *difference* in performance from the finite test sample tends to be dominated by the contribution from the finite training sample. In retrospect, this result seems intuitively obvious because trainers essentially become part of the classification algorithm, but this point does not seem to have been remarked upon in previous literature.

There is no reason to expect that the effects observed here will be diminished in importance as the dimensionality of the feature space and/or the complexity of the classifier increase beyond the examples considered here (cf. [6]). We base this statement on the fact that the Bayes classifier can only get more difficult to estimate. Thus, there is no evident case for ignoring either the first-order terms in the inverse of the number of training samples or the cross-terms in trainers and testers [1]. Also—unless the two competing classifiers are very similar—there will be no basis for expecting the interaction between training samples and classifier to be negligible. The examples analyzed in previous literature [1]–[7] are not sufficiently general to address these points. We hope to address them in further detail in future investigations.

The conclusions here regarding the role of the finite training sample argue for the necessity of using some form of resampling strategy that includes training samples as well as testing samples when assessing—and especially when comparing—classifiers under conditions where generalizability to a population of training samples as well as to a population of testing samples is desired. A counter-argument might be offered by practitioners who intend to "freeze" their classifier algorithm after the initial training. However, in practice, such an initial stance is almost always modified as more samples become available downstream and the classifier is modified accordingly. Thus, the contributions of the finite training sample will always manifest themselves: There will be more variance than expected on the basis of the finite test set alone. We therefore close with some comments on resampling strategies to address this issue.

The most commonly used resampling strategy for the training and testing of classifiers in SPR is the so-called cross-validation paradigm, where different partitionings of a given data set into trainers and testers are considered [19], [23]. In general, however, the conventional cross-validation approach cannot provide an estimate of uncertainty that includes the appropriate contribution from the finite-training set. This is clear from the limiting case of cross-validation embodied in leave-one-out training and testing. In that case, the training sets are almost identical and so the finite training set contribution to uncertainty is essentially unsampled.

A growing appreciation of the utility or even necessity of using some form of the bootstrap rather than cross-validation resampling in classifier assessment has developed over the last decade or so [19], [24]–[29] (cf. also [1]–[6]). For example, Efron compared the bootstrap with cross-validation and indicated that the former has lower variability than the latter but suffers from bias [25]. At the time of the review by Efron and Tibshirani [18], Efron's version of the bootstrap known as the 0.632 estimator [25] was found to perform "the best among all methods," but they suggested that further evidence was required before making more general recommendations. It is remarkable (and perhaps surprising) that the focus of most of this work was on estimation of *mean* performance but not of uncertainties and the associated issues of generalizability of error bars that are our focus here.

More recently Efron and Tibshirani have extended the 0.632 bootstrap to what they call the $0.632+$ bootstrap [28] and have provided methods for estimating the resulting uncertainties. We summarize the major features of the $0.632+$ bootstrap here, starting with its 0.632 predecessor. The latter is made up of two contributions: (1) Performance based on training using all of the original samples and testing on the very same samples, the so-called "apparent error"; (2) A bootstrap approach in which—for each original data sample—one keeps track of classification performance only from those bootstrap training sets that do not contain that data sample (the "leave-one-out bootstrap"). The 0.632 bootstrap is obtained by weighting the apparent error by a factor $1/e$ and the leave-one-out bootstrap by a factor $(1 - 1/e)$, or 0.632 (the large-sample limit of the expected fraction of distinct original observations in a bootstrap sample). The $0.632+$ bootstrap contains one further step that involves measuring performance when the input class labels are randomized (the no-information case) to obtain a correction for over-fitting. The overall process reduces to the 0.632 bootstrap when there is no overfitting, and to the leave-one-out bootstrap when the overfitting is maximum. These methods may be thought of as generalized cross-validation [28].

From the point of view of our present work, the most attractive feature of the more recent paper [28] is that the authors provide a formal approach not only to estimating mean performance but also to estimating the resulting uncertainties—using the original bootstrap samples. Alternatively, a conceptually straightforward (but computationally prohibitive) approach to estimating the uncertainties in bootstrap estimates is to bootstrap the bootstrap procedure. More practically, one can replace the second bootstrap with the so-called jackknife-after-bootstrap procedure [11], [18]. A lemma due to Efron and Tibshirani reduces the second step in practice to simply a sorting of the original bootstrap data [18].

For the present, we recommend the $0.632+$ bootstrap procedure, together with the methods for estimating standard errors in [28], or the alternative of the jackknife-after-bootstrap procedure [11], [18]. A subject for future investigation is to determine the properties of these estimates of uncertainty and their generalizability, i.e., the connection between such estimates of uncertainty and the general structure of the problem uncovered in the present work. The general picture of uncertainty analysis in the field of SPR is thus still incomplete.

APPENDIX I
THE MODEL EQUATIONS

The BWC approach [9] was inspired by a general analysis of components-of-variance models due to Roe and Metz [17]. The latter authors describe a large family of population experiments whose observed variances can be expressed as a linear combination of a small set of common underlying but unobserved model variance components; the particular mix of model components depends on the experiment. BWC [9] noted that one could solve for estimates of the underlying model components from a finite-sample data set of the form described in the present paper; this is done by replacing the set of population experiments with the corresponding set of bootstrap experiments, and replacing the population variances with bootstrap variance estimates. In particular, for the class of problems discussed in the present paper, there are six relevant population experiments and these will be included here for completeness. A central contribution of the Roe and Metz analysis [17] is their unambiguous and unifying subscript notation for the variance of an accuracy index A. Following their scheme, subscripts to the left of the vertical bar in the equations below refer to random effects, i.e., variables that are drawn randomly from the population when the experiment is replicated. Subscripts to the right of the vertical bar refer to fixed effects, i.e., variables that remain unchanged when the experiment is replicated.

The expected variance for the experiment in which both training samples and test samples are random but the classifier

is fixed contains all of the components in the present model:

$$\text{var}(A_{TR,TS|M}) = \sigma^2_{tr} + \sigma^2_{ts} + \sigma^2_{tr\cdot ts} + \sigma^2_{m\cdot tr} + \sigma^2_{m\cdot ts} + \sigma^2_{m\cdot tr\cdot ts} \ . \quad (2)$$

The expected variance for the experiment in which only test samples are random is

$$\text{var}(A_{TS|TR,M}) = \sigma^2_{ts} + \sigma^2_{tr\cdot ts} + \sigma^2_{m\cdot ts} + \sigma^2_{m\cdot tr\cdot ts} \ . \quad (3)$$

That is, σ^2_{tr} and $\sigma^2_{m\cdot tr}$ do not contribute because they correspond to fixed effects for this experiment; however, terms that include *tr* with *ts* do contribute because the presence of *ts* makes these random effects. (See [8], [9], [17].)

The remaining four experiments are those in which a *difference* in accuracy measures between competing classifiers is measured:

$$\text{var}(A_{TR,TS|M} - A_{TR,TS|M'}) = 2(\sigma^2_{m\cdot tr} + \sigma^2_{m\cdot ts} + \sigma^2_{m\cdot tr\cdot ts}) \quad (4)$$

$$\text{var}(A_{TS|TR,M} - A_{TS|TR',M}) = 2(\sigma^2_{tr\cdot ts} + \sigma^2_{m\cdot tr\cdot ts}) \quad (5)$$

$$\text{var}(A_{TS|TR,M} - A_{TS|TR,M'}) = 2(\sigma^2_{m\cdot ts} + \sigma^2_{m\cdot tr\cdot ts}) \quad (6)$$

$$\text{var}(A_{TS|TR,M} - A_{TS|TR',M'}) = 2(\sigma^2_{tr\cdot ts} + \sigma^2_{m\cdot ts} + \sigma^2_{m\cdot tr\cdot ts}) \quad (7)$$

These include comparisons between the same (M and M) or different (M and M') classifiers and between the same (*TR* and *TR*) or different (*TR* and *TR*′) training sets. Comparison of these results with those in the multiple-reader imaging problem [9] shows that the two problems are isomorphic. Bootstrap experiments are performed to obtain finite-sample estimates of the population variances on the left-hand sides. The system is then solved for estimates of the model components on the right-hand sides. Estimates of uncertainties in the estimates of the model components can be obtained using the second-order resampling strategy of the jackknife-after-bootstrap [11], [18].

APPENDIX II
JACKKNIFE-AFTER-BOOTSTRAP MONTE CARLO CHECK

We consider here a finite data set in the form of a single Monte Carlo trial in the paradigm of this paper. One can obtain estimates of the components of variance for this data set as discussed in the body of the paper using the family of bootstrap experiments of Appendix I. One may also estimate the *uncertainty* in those estimates based on that data set. This is achieved by means of a second resampling procedure called the jackknife-after-bootstrap [11], [18]. If one then progresses through a series of Monte Carlo replications of this entire exercise, one can also explore the distribution of the jackknife-after-bootstrap estimates of uncertainty and perform an overall consistency check as follows.

Figs. 1, 2, and 3 of the text displayed the *mean* estimates of the variance components and the standard deviation of these estimates over the population as represented by the Monte Carlo trials. The standard deviations from the case of Fig. 2 are re-displayed in Fig. 6 as the asterisks (*). Also shown in that figure is the root of the mean over the Monte Carlo trials of the estimate of the variance in the estimates of the variance components obtained on individual trials using the technique of the jackknife-after-bootstrap described in [11], [18]. Comparing these results we see that one may

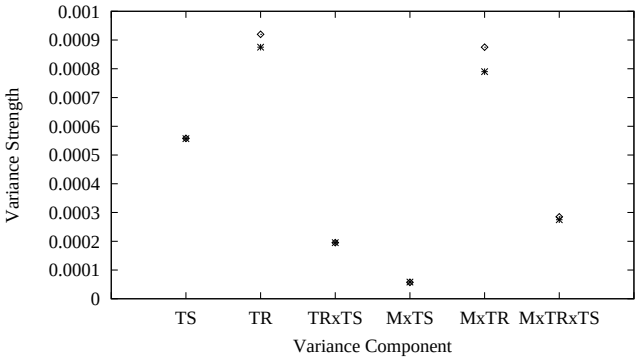

Fig. 6. Standard deviation of variance components over 300 Monte Carlo trails from Fig. 2 (asterisks); square root of mean variance estimate of variance components obtained over 300 Monte Carlo trails of jackknife-after-bootstrap procedure (diamonds).

indeed obtain approximately unbiased estimates of not only the variance components but also the uncertainties in those estimates from a data set of the kind represented by a single Monte Carlo realization of the experiment described in this paper. We note, however, that a great number of bootstraps ($\sim 25,000$) is required before the TR and M$\times$TR components achieve the level of convergence shown in the figure. With only 15,000 bootstraps—the number typically used in multiple-reader ROC studies in medical imaging as well as the present paper—the bias was roughly twice that shown in the figures. This exercise provides a check on the consistency of the present work. Further validation of the approach is provided in [9].

ACKNOWLEDGMENT

The authors gratefully acknowledge the very helpful questions and suggestions of the anonymous referees and the Associate Editor. We thank Prof. Richard Squier for supplying some of the additional computational resources needed for this study through Grant No. DAAD19-00-1-0165 from the U.S. Army Research Office. RFW gratefully acknowledges discussions of the first draft of this work with Prof. Charles E. Metz of the University of Chicago.

REFERENCES

[1] K. Fukunaga, *Introduction to statistical pattern recognition*, 2nd ed. Boston, MA: Academic Press, 1990.

[2] K. Fukunaga and R.R. Hayes, "Effects of sample size in classifier design," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 8, pp. 873–885, Aug. 1989.

[3] ——, "Estimation of classifier performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 10, pp. 1087–1101, Oct. 1989.

[4] Š.J. Raudys and A.K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 252–264, 1991.

[5] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[6] Š.J. Raudys, *Statistical and neural classifiers: An integrated approach to design.* New York, NY: Springer-Verlag, 2001.

[7] H.-P. Chan, B. Sahiner, R.F. Wagner, and N. Petrick, "Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers," *Medical Physics*, vol. 26, no. 12, pp. 2654–2668, 1999.

[8] D.D. Dorfman, K.S. Berbaum, and C.E. Metz, "Receiver operating characteristic rating analysis: Generalization to the population of readers and patients with the jackknife method," *Investigative Radiology*, vol. 27, pp. 723–731, 1992.
[9] S.V. Beiden, R.F. Wagner, and G. Campbell, "Components-of-variance models and multiple-bootstrap experiments: An alternative method for random-effects receiver operating characteristic analysis," *Academic Radiology*, vol. 7, pp. 341–349, 2000.
[10] S.V. Beiden, R.F. Wagner, G. Campbell, C.E. Metz, and Y. Jiang, "Components-of-variance models for random-effects ROC analysis: The case of unequal variance structures across modalities," *Academic Radiology*, vol. 8, pp. 605–615, 2001.
[11] S.V. Beiden, R.F. Wagner, G. Campbell, and H.-P. Chan, "Analysis of uncertainties of estimates of variance components in multivariate ROC analysis," *Academic Radiology*, vol. 8, pp. 616–622, 2001.
[12] C.E. Metz, "ROC methodology in radiologic imaging," *Investigative Radiology*, vol. 21, pp. 720–733, 1986.
[13] ——, "Some practical issues of experimental design and data analysis in radiological ROC studies," *Investigative Radiology*, vol. 24, pp. 234–245, 1989.
[14] Y. Jiang, C.E. Metz, and R.M. Nishikawa, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests," *Radiology*, vol. 201, pp. 745–750, 1996.
[15] C.E. Metz, "Statistical analysis of ROC data in evaluating diagnostic performance," in *Multiple regression analysis: Applications in the health sciences*, D. Herbert and R. Meyers, Eds. New York, NY: American Institute of Physics: American Association of Physicists in Medicine, 1986, pp. 365–384.
[16] C.E. Metz, Y. Jiang, H. MacMahon, R.M. Nishikawa, and X. Pan, "ROC software," Kurt Rossmann Laboratories for Radiologic Image Research, University of Chicago, Chicago, IL, Web page, 2003. [Online]. Available: http://www-radiology.uchicago.edu/krl/roc_soft.htm
[17] C.A. Roe and C.E. Metz, "Variance-component modeling in the analysis of receiver operating characteristic index estimates," *Academic Radiology*, vol. 4, pp. 587–600, 1997.
[18] B. Efron and R.J. Tibshirani, *An introduction to the bootstrap*. New York, NY: Chapman & Hall, 1993.
[19] T. Hastie, R.J. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer-Verlag, 2001.
[20] M.A. Maloof, S.V. Beiden, and R.F. Wagner, "Analysis of competing classifiers in terms of components of variance of ROC accuracy measures," Department of Computer Science, Georgetown University, Washington, DC, Technical Report CS-02-01, January 2002, http://www.cs.georgetown.edu/~maloof/pubs/cstr-02-01.pdf.
[21] L.P. Clarke, B.Y. Croft, E. Staab, H. Baker, and D.C. Sullivan, "National Cancer Institute initiative: Lung image database resource for imaging research," *Academic Radiology*, vol. 8, pp. 447–450, 2001.
[22] L.E. Dodd, R.F. Wagner, S.G. Armato, et al. "An overview of assessment methodologies and related statistical issues for computer-assist modalities in lung imaging: A status report of the lung image database consortium (LIDC)," *Academic Radiology*, To be submitted, July 2003.
[23] A.P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.
[24] B. Efron, *The jackknife, the bootstrap, and other resampling plans*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1982.
[25] B. Efron., "Estimating the error rate of an prediction rule: Improvement on cross-validation," *Journal of the American Statistical Association*, vol. 78, no. 382, pp. 316–331, 1983.
[26] G.J. McLachlan, *Discriminant analysis and statistical pattern recognition*. New York, NY: John Wiley & Sons, 1992.
[27] B.D. Ripley, *Pattern recognition and neural networks*. Cambridge UK: Cambridge University Press, 1996.
[28] B. Efron and R.J. Tibshirani, "Improvements on cross-validation: The .632+ bootstrap method," *Journal of the American Statistical Association: Theory and Methods*, vol. 92, no. 438, pp. 548–560, 1997.
[29] 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," in *Proc. of the SPIE*, vol. 3661, 1999, pp. 523–532.

Sergey V. Beiden received the MS degree in Applied Mathematics and Physics from the Moscow Institute of Physics and Technology in 1990 and the PhD degree from the Russian Research Center Kurchatov Institute in 1993. He has held Fellowships at the University of Sheffield (UK) and The University of West Virginia (US) doing first-principles calculations of complex processes in metals, alloys, and phase transitions. From 1999–2002 he worked at the US Food and Drug Administration on random-effects ROC analysis in breast and lung cancer screening and computer-aided diagnosis. He is currently working for the CARANA Corporation in Moscow on electricity market modeling and applications of machine learning and extreme value theory to forecasting.

Marcus A. Maloof received the BS and MS degrees from the University of Georgia in 1989 and in 1992, respectively, and received the PhD degree from George Mason University in 1997. He was a post-doctoral fellow at the Institute for the Study of Learning and Expertise in Palo Alto, CA, and a visiting scholar in the Center for the Study of Language and Information at Stanford University. In 1998, he joined the faculty of Georgetown University where he is an assistant professor in Department of Computer Science. He is a member of AAAI, IEEE, and IEEE Computer Society. His research interests include on-line learning algorithms and their application to concept drift and computer security.

Robert F. Wagner received the BS in Electrical Engineering from Villanova University in 1959, the MA degree from Augustinian College in 1965, the MS degree in Physics (1965) and the PhD degree in Theoretical Physics in 1969 from Catholic University of America. From 1970–1972 he did research in photo- and electro-nuclear interactions at Ohio University. Since 1972 he has worked at the FDA's Center for Devices and Radiological Health on problems of statistical physics of the major medical imaging modalities, signal detection and statistical decision analysis, and computer-aided diagnosis, with over 150 publications (shared Best Paper, 1983, IEEE Sonics and Ultrasonics). He is a Fellow of the IEEE, SPIE, SPSE/IST (Imaging Societies), The Opt. Soc. of Amer. (OSA), and the Amer. Inst. of Medical and Biol. Engineering (AIMBE). He was elected to the NIH/FDA Senior Biomedical Research Service in 1995.