Indexes for Three-Class Classification Performance Assessment—An Empirical Comparison

Mehul P. Sampat, *Member, IEEE*, Amit C. Patel, Yuhling Wang, Shalini Gupta, Chih-Wen Kan, Alan C. Bovik, *Fellow, IEEE*, and Mia K. Markey, *Member, IEEE*

***Abstract*—Assessment of classifier performance is critical for fair comparison of methods, including considering alternative models or parameters during system design. The assessment must not only provide meaningful data on the classifier efficacy, but it must do so in a concise and clear manner. For two-class classification problems, receiver operating characteristic analysis provides a clear and concise assessment methodology for reporting performance and comparing competing systems. However, many other important biomedical questions cannot be posed as "two-class" classification tasks and more than two classes are often necessary. While several methods have been proposed for assessing the performance of classifiers for such multiclass problems, none has been widely accepted. The purpose of this paper is to critically review methods that have been proposed for assessing multiclass classifiers. A number of these methods provide a classifier performance index called the volume under surface (VUS). Empirical comparisons are carried out using 4 three-class case studies, in which three popular classification techniques are evaluated with these methods. Since the same classifier was assessed using multiple performance indexes, it is possible to gain insight into the relative strengths and weakness of the measures. We conclude that: 1) the method proposed by Scurfield provides the most detailed description of classifier performance and insight about the sources of error in a given classification task and 2) the methods proposed by He and Nakas also have great practical utility as they provide both the VUS and an estimate of the variance of the VUS. These estimates can be used to statistically compare two classification algorithms.**

***Index Terms*—Classification evaluation, ideal observer analysis, three-class receiver operating characteristic (ROC), volume under surface (VUS).**

I. Introduction

ASSESSMENT of classifier performance is critical for fair comparison of methods, including considering alternative models or parameters during system design. The assessment must not only provide meaningful data on the classifier efficacy, but it must do so in a concise and clear manner.

Many interesting problems in biomedicine can be represented as "two-class" classification tasks. For example, one may develop a computer-aided diagnosis (CAD_x) system to predict whether a lesion is due to a benign or malignant process based on its appearance on medical imaging. For such problems, receiver operating characteristic (ROC) analysis is the accepted standard for performance assessment [1]. An ROC curve is a plot of the sensitivity versus 1-specificity that shows the tradeoffs in these quantities that can be achieved with the classifier under study. The area under the ROC curve (AUC) is widely used as a summary index for the classifier performance. Thus, for two-class classification problems, ROC analysis provides a clear and concise assessment methodology for reporting performance and comparing competing systems. The ROC methodology has a number of advantages: 1) it has a single summary measure (AUC) and a simple graphical view; 2) the values of the AUC for chance and perfect performance are clearly specified; 3) it does not depend on disease prevalence; and 4) it is invariant to monotonic transformations of the decision variable.

However, many other important biomedical questions cannot be posed as "two-class" classification tasks. Instead, more than two classes are often necessary. For example, it is more informative to predict disease stage than merely to classify whether the disease is present or not. As an example, in computer-aided detection of cancer, one must distinguish between false positive detections, benign lesions, and malignant lesions. While several methods have been proposed for assessing the performance of classifiers for such multiclass problems, none has been widely accepted. Note that while some measures have been defined only for the extension to three classes, others can be used to assess classifiers for tasks involving three or more classes.

The purpose of this paper is to critically review methods that have been proposed for assessing multiclass classifiers. Empirical comparisons are carried out using 4 three-class case studies. Since the same classifier was assessed using multiple performance indexes, it is possible to gain insight into the relative strengths and weakness of the measures.

II. Materials and Methods

A. Datasets

Our aim was to compare methods that have been proposed for evaluating classifier performance when three or more classes are to be distinguished. Toward this goal, we applied several

Manuscript received December 1, 2007; revised August 3, 2008 and October 13, 2008. First published January 20, 2009; current version published May 6, 2009. This work was supported in part by a Predoctoral Traineeship Award from the U.S. Army Medical Research and Materiel Command under Grant W81XWH-04-1-0406 and by an Early Career Award from Wallace H. Coulter Foundation (Markey).

M. P. Sampat is with the Center for Neurological Imaging, Department of Radiology, Brigham and Women's Hospital, Boston, MA 02115 USA (e-mail: mehul.sampat@ieee.org).

A. C. Patel is with the University of Texas Southwestern, Dallas, TX 75390 USA (e-mail: a.c.patel11@gmail.com)

Y. Wang is with the Department of Biomedical Engineering, Charlottesville, VA 22908 USA (e-mail: yw9f@virginia.edu)

S. Gupta and A. C. Bovik are with the Department of Electrical and Computer Engineering, University of Texas, Austin, TX 78712 USA (e-mail: shalinig@ece.utexas.edu; bovik@ece.utexas.edu).

C.-W. Kan and M. K. Markey are with the Department of Biomedical Engineering, University of Texas, Austin, TX 78712 USA (e-mail: wendykan@mail.utexas.edu; mia.markey@mail.utexas.edu).

Digital Object Identifier 10.1109/TITB.2008.2009440

performance measures to the outputs of three different classifiers applied on four different datasets.

1) Simulated Datasets: As we wanted to study the behavior of the various indexes for different classification settings, we first investigate their response under very specific conditions using two simulated datasets. For both simulated datasets, we use a single feature x for each class c_i and assume that the conditional distributions $p(x|c_i)$ for each class are normal distribution functions with mean and standard deviations (μ_1, σ_1), (μ_2, σ_2), and (μ_3, σ_3), respectively. For all classes, the standard deviation was set to one $(\sigma_1 = \sigma_2 = \sigma_3 = 1)$. For the first dataset, the separation between the classes was empirically defined as $\text{dist} = |\mu_1 - \mu_2| + |\mu_2 - \mu_3|$. This separation distance was systematically varied from 0 to 15 to create different scenarios with varying degrees of overlap between the three classes. We set $\mu_1 = 1$, $\mu_2 = 1$, and $\mu_3 = 1$, and then increased μ_2 in increments of 0.5 and μ_3 in increments of 1. For all values of μ_2 and μ_3, $\sigma_1 = \sigma_2 = \sigma_3 = 1$.

For the second simulated dataset, two situations were considered: 1) the conditional distribution functions of classes 1 and 2 are completely overlapped, while that of class 3 is completely separated from these classes ($\mu_1 = 1$, $\mu_2 = 1$, and $\mu_3 = 7$) and 2) the conditional distribution functions for classes 2 and 3 are completely overlapped, while that of class 1 is completely separated from these classes ($\mu_1 = 1$, $\mu_2 = 7$, and $\mu_3 = 7$). For both cases, $\sigma_1 = \sigma_2 = \sigma_3 = 1$.

2) Fisher Iris Dataset: Fisher's Iris dataset (Iris flower dataset) [2] is a popular three-class dataset commonly used in classification experiments and is routinely found in the pattern recognition literature. This dataset is a multivariate dataset that consists of 50 samples from three species of Iris flowers (*I. setosa*, *I. virginica*, and *I. versicolor*). For each sample, four features were measured: the length and the width of sepal and petal. It is known that one class is linearly separable from the other two, which are not linearly separable from each other. This dataset allows us to compare the behavior and effectiveness of three-class classifier performance indexes when a varying number of input features (one to four in this case) are employed to represent each object.

3) BI-RADS Mammography Dataset: The dataset consists of 326 nonpalpable, mammographically suspicious breast mass lesions that underwent biopsy (core or excisional) at Duke University Medical Center. Each case was interpreted by one of seven radiologists. The mass margin, mass shape, and mass density were reported according to the Breast Imaging Reporting and Data System (BI-RADS) lexicon [3] and encoded as described in a previous publication [4]. The mass size and patient age were also collected. The radiologist's "gut assessment" of the likelihood of malignancy on a scale of 1–5 was provided for each lesion. Few biopsied lesions were rated as having a very low likelihood of malignancy (1 or 2); so the dataset for this study includes only those rated from 3 to 5, where 3 represents "indeterminate," 4 represents "likely malignant," and 5 represents "malignant."

4) Spectroscopy Dataset: This dataset was obtained from a pilot clinical study conducted to measure the spectroscopic signature of oral mucosa lesions suspicious for dysplasia or carcinoma [5], using oblique polarization reflectance spectroscopy (OPRS) [6], [7]. Twenty-seven patients over the age of 18 years who were referred to the Head and Neck Clinic at the University of Texas M. D. Anderson Cancer Center were inducted into the study. For each patient, spectroscopic measurements were typically performed on one to two visually abnormal sites and one visually normal site. Biopsies were conducted for all sites. A total of 57 sites were measured, of which 22 were visually and histopathologically normal (normal), 13 sites were visually abnormal but histopathologically normal (benign), 12 were visually abnormal sites that proved to be mild dysplasia (MD) on histopathology, and 10 were visually abnormal sites that proved to be severe high grade dysplasia or carcinoma (SD) after histopathology. In each measurement, parallel and perpendicular spectra were collected. Five spectral signals were measured in this study. These are parallel signals, perpendicular signals, diffuse reflectance spectrum, the ratio of parallel to perpendicular, and the depolarization ratio [5]. Six features are extracted from each spectrum: average nuclear size, average of parallel signals, average of perpendicular signals, average of diffuse signals, average of the ratio of parallel to perpendicular signals, and average of the depolarization ratio signals [5]. Additional details on preprocessing and feature extraction are described in [5].

B. Classifiers

Classifiers can provide either a single continuous decision variable or a set of posterior probabilities for each class. Different performance indexes were designed with different classifier output formats in mind. For example, for a three-class classification task, Mossman [8] proposed a performance index that uses the three probabilities for belonging to each class as input. In comparison, Nakas and Yiannoutsos [9] developed an index that requires a single decision variable as input. Note that when a single decision variable is used, one must know the order in which the classes occur. One such example of the class order is that objects from class c_3 tend to have larger measurements than objects from class c_2 and that objects from c_2 tend to have larger measurements than objects from c_1.

In this study, three types of classifiers are used to provide examples of both output formats. A linear regression model is used to obtain a single continuous decision variable while a *k*-nearest-neighbor (kNN) classifier and a Bayes classifier [10] provide posterior probabilities for each class. Each of the three classifiers is briefly described in the following sections. Since the purpose of the study is to evaluate performance indexes, using three types of classifiers is also valuable as it enables us to assess the classifier comparisons that would result from using a given performance index. This is an important consideration since a common use of performance measures is to empirically select among different models since the no free lunch theorem [10] tells us that we generally cannot know *a priori* which classifier will serve us best. Leave-one-out cross-validation was used as a sampling method for all classifiers. In other words, for each dataset, each object was excluded once from training and reserved as the test case, while the remaining cases were

used to train the classifier. Using this sampling method, more predictions can be obtained with the given amount of data by interchanging the training and test sets.

1) k-Nearest Neighbor: The *k*-NN classifier is simple, yet robust in its application. Let each observation in the dataset be described by a d-dimensional feature vector $\vec{x}$. The same features are then obtained for an unknown observation. Next, the distance in feature space between the unknown observation and the known observations in the dataset can be found. Based on these distances, the *k* NNs, or the *k* observations from the dataset, that have the smallest distance measure from the unknown observation are identified. The *a posteriori* probability that the unknown observation belongs to each output class is obtained by calculating the proportion of cases of each class that are included among the *k* NNs of the unknown observation in feature space. For example, if $k = 13$, and the number of NNs for classes 1, 2, and 3 for a case are 3, 4, and 6, respectively, the *a posteriori* probabilities for this case of belonging to each of the classes are 3/13, 4/13, and 6/13, respectively. In order to obtain a discrete classification, the unknown observation is assigned to the class to which the highest proportion of the known observations belongs.

2) Bayes Classifier: Let $P(c_i)$ denote the *a priori* probability that a sample belongs to class c_i, where $i = 1, 2, \ldots, N$. Let each sample be represented by a d-dimensional feature vector $\vec{x}$. Let $p(\vec{x}|c_i)$ denote the class-conditional probability density function. It represents the probability distribution function for feature vector $\vec{x}$ given that $\vec{x}$ belongs to class c_i. Let $P(c_i|\vec{x})$ be the *a posteriori* probability, which is the probability that the sample belongs to class c_i given the feature vector $\vec{x}$. Given $P(c_i)$ and $p(\vec{x}|c_i)$, the *a posteriori* probability for a sample represented by the feature vector $\vec{x}$ is given by the Bayes formula [10] $P(c_i|\vec{x}) = p(\vec{x}|c_i) \times P(c_i)/p(\vec{x})$, where $p(\vec{x}) = \sum_{i=1}^{N} p(\vec{x}|c_i) \times P(c_i)$. The formula is applicable for all probability density functions; however, the normal density function is often used to model the distribution of feature values of a particular class. For simplicity, this assumption was also made in our study. The general multivariate normal density function in d dimensions is given by

$$p(x) = \frac{1}{(2\pi)^{d/2}\ |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^t \sum^{-1} (x-\mu)\right]$$

where μ is the d-component mean vector, $\sum$ is the $d \times d$ covariance matrix, and $|\sum|$ and $\sum^{-1}$ are its determinant and inverse, respectively. In this paper, a normal multivariate density function was used to model the features for each class for all datasets. The parameters μ and $\sum$ of the probability density function for each class are calculated from the training observations belonging to that class. These parameters are estimated by computing the mean and covariance of the features of the training observations. As noted before, some performance methods operate on the continuous classifier output whereas others require discrete classifications. For any given test observation described by the feature vector $\vec{x}$, a discrete classification can be obtained using the Bayes decision rule, which is: decide c_i if $P(c_i|\vec{x}) > P(c_j|\vec{x})\ \forall j \neq i$.

3) Linear Regression: Linear regression is a method that models the linear relationship between a dependent variable y and d independent variables (features), represented by the vector $\vec{x}$ as $y = \vec{x}\lambda + \varepsilon$, where ε is a random error term. The parameters λ are selected such that the sum of residuals (the difference between the predicted and observed values) squared is minimized. A linear regression model produces a single continuous output value for each observation, the value of which depends on the labels selected to represent the various classes. One example of such an encoding scheme is: normal $= 1$, benign $=$ 2, and malignant $= 3$. The encoding of the class labels does not affect the computation of the measures of classifier performance. Moreover, if the target classes have a natural ordering biologically, the results are easier to interpret if that ordering is retained. For example, [1 = normal, 2 = mild dysplasia, 3 = severe dysplasia] is a better choice than [1 = mild dysplasia, 2 = normal, 3 = severe dysplasia] since mild dysplasia is the intermediate state biologically.

C. Conversion of Classifier Outputs

From Section II-B, we observe that for each object, the classifiers considered in this study can have two types of outputs.

Type A: Percent likelihood of belonging to each of the three classes, e.g., for the Bayes classifier, one obtains three posterior probabilities for each object.

Type B: One continuous output for each object, e.g., the output from a linear regression classifier.

As mentioned earlier, for some three-class ROC analysis methods, the input (which is the classifier's output) must be of type A (e.g., Mossman method), whereas for other methods, the input must be of type B (e.g., Nakas method). Thus, to be able to evaluate the performance of all classifiers with all methods, we need to define methods to transform the classifier's output from type A to type B, and vice versa. In Section II-C1 and II-C2, we define the rules for these transformations.

1) Conversion From Percent Likelihood to a Continuous Output (Conversion From Type A to Type B): Let $P(c_1|\vec{x})$, $P(c_2|\vec{x})$, and $P(c_3|\vec{x})$ represent the *a posteriori* probability for a sample represented by the feature vector $\vec{x}$ for classes 1, 2, and 3, respectively. Let ctsOp denote the continuous output. Then the rule for conversion from type A to type B is as follows:

$$\text{If } P(c_1|\vec{x}) \geq P(c_2|\vec{x}) \text{ and } P(c_1|\vec{x}) \geq P(c_3|\vec{x}),$$
$$\text{then ctsOp} = 0.5 + P(c_1|\vec{x})$$
$$\text{If } P(c_2|\vec{x}) \geq P(c_1|\vec{x}) \text{ and } P(c_2|\vec{x}) \geq P(c_3|\vec{x}),$$
$$\text{then ctsOp} = 1.5 + P(c_2|\vec{x})$$
$$\text{If } P(c_3|\vec{x}) \geq P(c_1|\vec{x}) \text{ and } P(c_3|\vec{x}) \geq P(c_2|\vec{x}),$$
$$\text{then ctsOp} = 2.5 + P(c_3|\vec{x}).$$

Note that 0.5, 1.5, and 2.5 are ordered but arbitrary offsets. Any other set of ordered offsets can also be used.

TABLE I
GENERAL CONFUSION MATRIX FOR ASSESSING THE PERFORMANCE IN A THREE-CLASS CLASSIFICATION TASK

		Actual			
		Class c_1	Class c_2	Class c_3	TOTAL
Predicted	Class c_1	$V(c_1,c_1)$	$V(c_2,c_1)$	$V(c_3,c_1)$	$V(\bullet,c_1)$
	Class c_2	$V(c_1,c_2)$	$V(c_2,c_2)$	$V(c_3,c_2)$	$V(\bullet,c_2)$
	Class c_3	$V(c_1,c_3)$	$V(c_2,c_3)$	$V(c_3,c_3)$	$V(\bullet,c_3)$
	TOTAL	$V(c_1,\bullet)$	$V(c_2,\bullet)$	$V(c_3,\bullet)$	$V(\bullet,\bullet)$

$V(c_1, c_1)$ is the number of items that truly belong to class c_1 and were classified as class c_1 $V(c_1, c_2)$ is the number of items that truly belong to class c_1 but were classified as class c_2, $V(c_1,\bullet)$ is the number of items that truly belong to class c_1; etc.

2) Conversion From a Continuous Output to Percent Likelihood (Conversion From Type B to Type A): A continuous output can be converted to percent likelihood by fitting Gaussian distributions to the continuous output for each class to obtain three distributions: $P(\mathrm{ctsOp}|c_1)$, $P(\mathrm{ctsOp}|c_2)$, and $P(\mathrm{ctsOp}|c_3)$, where ctsOp denotes the output of the linear regression classifier. Then, for each object i, three likelihood values are obtained by computing $P(\mathrm{ctsOp}_i|c_1)$, $P(\mathrm{ctsOp}_i|c_2)$, and $P(\mathrm{ctsOp}_i|c_3)$, where ctsOp_i represents the continuous output of the classifier for the *i*th object.

D. Classifier Performance Indexes

1) Overview: Several methods have been proposed for assessing the performance of multiclass classifiers. Some of these include pairwise AUC comparisons [11], Hand and Till's *M* function (HTM) [11], one-versus-all (OVA) AUC comparisons [11], the modified HTM (HT3, [12]), Scurfield's method [13], Mossman's three-way ROC [8], He's method [14], and Nakas's method [9]. In this study, performance measures were empirically compared on the outputs of classifiers on the 4 three-class datasets described before. Each method was implemented in MATLAB (The MathWorks, Inc., Natick, MA) and is available for download as part of our Classifier Performance Evaluation Toolbox from our Web site (http://bmil.bme.utexas.edu/files/bmil/publications/CPET-0.1b.zip).

The confusion matrix (shown in Table I) is an excellent way of organizing discrete classification data, based on the actual state of an object versus its predicted state as determined by a classifier. In the confusion matrix, $V(c_i, c_j)$ represents how many subjects that were actually of type c_i were classified as type c_j. For a three-class decision problem, using classes c_1, c_2, and c_3, $V(c_1, c_1)$ represents the number of class c_1 subjects correctly classified as class c_1, $V(c_1, c_2)$ represents the number of class c_1 subjects incorrectly classified as class c_2, etc., for a total of nine different combinations. These nine values can be formatted into a 3×3 confusion matrix, as seen in Table I. This can be easily extended to more than three classes. An *N*-class problem would have N^2 combinations for $V(c_i, c_j)$, i.e., an $N \times N$ confusion matrix would be needed to represent all the values of $V(c_i, c_j)$.

2) Pairwise Comparisons: Pairwise comparisons break down an *N*-class classification problem into separate binary one-versus-one comparisons. For an *N*-class classification task, there exist $N(N-1)$ different binary comparisons. Thus, this method returns $N(N-1)$ pairwise AUCs for each paired comparison. The input to this method has to be of type A (Section II-C). This technique breaks down the problem into multiple binary classifications *after* a multiclass classifier has been applied to a dataset. Note that this is different from pairwise classification in which the problem is broken down into binary classification problems *before* classification, such that two-class classifiers are used. For example, when classifying breast lesions as benign, malignant but not invasive, or invasive malignancy, pairwise classification could be performed by designing separate two-class classifiers for benign versus invasive malignant, invasive malignancy versus malignant but noninvasive, and benign versus malignant but noninvasive tasks. By comparison, what is referred to as "pairwise comparison" in this study is pairwise performance *evaluation* of a multiclass classifier. This approach can be used to determine how well a classifier separates one class from another. It provides the user a detailed view of exactly which classes may cause the most trouble for the classifier. For a three-class classifier, the problem would be broken up into six different pairwise comparisons represented as six AUCs. These six AUCs could then be used to judge aspects of classifier performance. For each observation classified, three posterior probabilities are obtained. While computing the AUC for class 1 and class 2, one can use the posterior probabilities for either class 1 or class 2 to generate the ROC curve. Based on which posterior probability is used, one would obtain two different AUC values. For the three-class task, three comparisons are possible (class 1 versus 2, class 1 versus 3, and class 2 versus 3), and thus, a total of six AUCs are obtained.

3) Hand and Till M Function: In Hand and Till's *M* function [11], the classifier performance is given by HTM, calculated by $\mathrm{HTM} = [2/\{N(N-1)\}] \sum_{i<j} \bar{A}(i,j)$, where N is the number of classes in the classification problem and $\bar{A}(i,j)$ is the probability that a randomly chosen member of class i has a lower probability of belonging to class j than a randomly chosen member of class j. Note that the input to this method has to be of type A (Section II-C). When there are only two classes, the HTM function is the same as the traditional AUC method. Essentially this represents the average of all the possible pairwise comparisons. For a three-class problem, after the six pairwise AUCs have been determined, the HTM index can be found through their average. This allows for a more generalized view of classifier performance, as opposed to multiple

TABLE II
EXAMPLE CONFUSION MATRIX FOR CLASSIFYING INTO ONE CLASS VERSUS THE REST OF THE CLASSES

		Actual		
		Class a	Rest	Total
Predicted	Class a	$V(c_1,c_1)$	$V(c_2,c_1)+V(c_3,c_1)$	$V(\bullet,c_1)$
	Rest	$V(c_1,c_2)+V(c_1,c_3)$	$V(c_2,c_2)+V(c_2,c_3)+V(c_3,c_3)+V(c_3,c_2)$	$V(\bullet,\overline{c_1})$
	Total	$V(c_{1,\bullet})$	$V(\overline{c_{1,\bullet}})$	$V(\bullet,\bullet)$

AUCs. Note that a classifier with ideal performance would have an HTM value of 1.00 and a classifier with chance performance would have an HTM of 0.50.

4) OVA Comparisons: OVA comparisons cast an *N*-class classification problem as separate binary OVA comparisons. The input to this method has to be of type A (Section II-C). An *N*-class classification task requires *N* different binary OVA comparisons. Each of these *N* binary OVA comparisons has its own AUC, which can be used as a measure of how well the classifier separates one class from all the other classes. A three-class problem would result in three separate AUC measures that describe each OVA comparison. Like the pairwise comparisons, the OVA method allows for a detailed look at classifier performance; however, it is more general than pairwise in that OVA provides only one summary measure per class.

5) HT3 (Modified HTM): The HT3 [12] function is a modified version of the HTM function and is calculated by averaging the AUCs of classifying one class versus all the rest combined (OVA). In other words, HT3 is the average of OVA AUCs in analogy to the fact that HTM is the average of the pairwise AUCs. Thus, HT3= $(\mathrm{AUC}_{a,\mathrm{rest}}+\mathrm{AUC}_{b,\mathrm{rest}}+\mathrm{AUC}_{c,\mathrm{rest}})/3$, where

$$\mathrm{AUC}_{i,\mathrm{rest}} = \max\left(\frac{1}{2}, 1-\frac{2(\mathrm{rest},i)}{V(c_1,\cdot)}-\frac{2(i,\mathrm{rest})}{V(c_2,\cdot)+V(c_3,\cdot)}\right)$$

and Table II illustrates the computation of (rest,i) and (i,rest) for $i=c_1$. In essence, this is the average of three different OVA comparisons. The input to this method is the confusion matrix. As the original Hand and Till function allowed for a more generalized version of the pairwise method through averaging, the modified Hand and Till function does the same for the OVA method. For an N-class problem, each of the *N* OVA AUCs can be calculated and averaged to obtain the index. For example, for a three-class task, HT3 is the average of three OVA comparisons.

6) Macroaverage: When assessing classifier performance, the macroaverage takes into account only the correct classification rates and is a simple average of the three correct classification rates. The macroaverage [12] is given by $\mathrm{MAVG}_3 = [V(c_1,c_1)+V(c_2,c_2)+V(c_3,c_3)]/3$ (recall the confusion matrix shown in Table I; the input to this method is the confusion matrix). This is also commonly referred to as "classification accuracy" or "percent correct" and is widely used for the assessment of multiclass classification techniques.

7) Cobweb Graph: An *N*-class classification problem can also be represented by the $N(N-1)$ misclassification values from a confusion matrix. An $N(N-1)$ equilateral polygon can be created to map the point obtained from the confusion matrix. Chance classification would be represented by the equilateral polygon whose corners are at a distance of $1/N$ from the center of the polygon. Each of the separate misclassifications

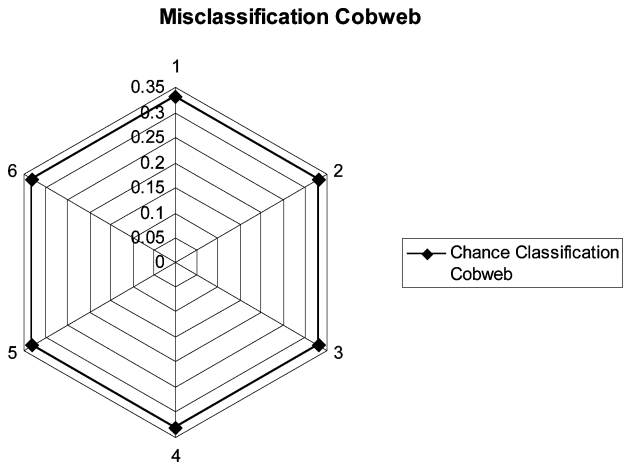

Fig. 1. Example confusion matrix for classifying into one class versus the rest of the classes.

can represent one corner on a polygon for a graphical representation. The input to this method are entries of the confusion matrix. The values chosen from a three-class classification confusion ratio matrix, which is shown in Table I, form a 6-D point. Given three classes (c_1,c_2,c_3), the following point, $(V(c_1,c_2),V(c_1,c_3),V(c_2,c_1),V(c_2,c_3),V(c_3,c_1),V(c_3,c_2))$, is obtained from a confusion ratio matrix, where $V(c_1,c_2)$ corresponds to the cell of the confusion ratio matrix that gives the fraction of class c_1 objects that were misclassified as class c_2 objects. A chance classification is shown in Fig. 1, and the point represented is (0.33, 0.33, 0.33, 0.33, 0.33, 0.33). The misclassification rates of 0.33 indicate that when confronted with an object of type c_1, the classifier would classify it as having an equal likelihood of being from any of the three classes c_1,c_2, or c_3. This method attempts to address the need for a simple visualization of classifier performance, which is one of the advantages of traditional ROC analysis that is lacking in many attempts to extend the approach to three or more classes. However, the cobweb representation is very complicated for large number of classes.

Note that methods 1–6 described before use indexes that measure the discriminability between only two classes at a time. Methods 8–12, described shortly, present measures that aim to quantify the discriminability between all three classes simultaneously. A common theme in these methods is the use of a 3-D ROC surface and the measure of the volume under this surface (VUS). Let c_j represent the true class of given sample and let o_i represent the class assigned to this sample by the classification algorithm. To describe such surfaces, we must first define the quantity $P(o_i|c_j)$, which is referred to as the conditional classification rate by Edwards and Metz [15]. $P(o_i|c_j)$ is the probability that an object is classified as belonging to class i given that it actually belongs to class j. For a three-class task, nine such possibilities exist, as shown in Table III.

Note that in Table III, $P(o_1|c_1)+P(o_2|c_1)+P(o_3|c_1)=1$, $P(o_1|c_2)+P(o_2|c_2)+P(o_3|c_2)=1$, and $P(o_1|c_3)+P(o_2|c_3)+P(o_3|c_3)=1$. Thus, for the three-class task, there are six independent variables and, in general, for the *N*-class task, there are (N^2-N) independent variables. The 3-D ROC

TABLE III
NINE CONDITIONAL CLASSIFICATION RATES FOR A THREE-CLASS CLASSIFICATION TASK

		Classification output		
		o_1	o_2	o_3
True class	c_1	$P(o_1 \mid c_1)$	$P(o_2 \mid c_1)$	$P(o_3 \mid c_1)$
	c_2	$P(o_1 \mid c_2)$	$P(o_2 \mid c_2)$	$P(o_3 \mid c_2)$
	c_3	$P(o_1 \mid c_3)$	$P(o_2 \mid c_3)$	$P(o_3 \mid c_3)$

surfaces are typically constructed by plotting the three correct conditional classification rates: $P(o_1 \mid c_1)$, $P(o_2 \mid c_2)$, and $P(o_3 \mid c_3)$. The volume under this ROC surface is the probability that the measurements of three subjects, one from each class, will be correctly classified.

8) Nakas's Method: Nakas and Yiannoutsos [9] proposed a nonparametric measure for the volume under the ROC surface that requires a single decision variable as input (type B, Section II-C). Note that when a single decision variable is used, one must know the order in which the classes occur. One such example of the class order is: objects from class c_3 tend to have larger measurements than objects from classes c_2 and objects from c_2 tend to have larger measurements than objects from c_1. Let X_{1i}, X_{2j}, and X_{3k} represent the value of the decision variable for the ith, jth, and kth objects from classes c_1, c_2, and c_3, respectively, and let n_1, n_2, and n_3 be the sample sizes of classes c_1, c_2, and c_3, respectively. The measure proposed by Nakas and Yiannoutsos [9] for a three-class task is defined as

$$\widehat{\theta}_V = (1/n_1 n_2 n_3) \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} I(X_{1i}, X_{2j}, X_{3k}).$$

As mentioned before, there is an order associated with the three variables (X_{1i}, X_{2j}, X_{3k}), which is known beforehand. $I(X_{1i}, X_{2j}, X_{3k})$ is a function that equals one if X_{1i}, X_{2j}, and X_{3k} are in the correct order $(X_{1i} \le X_{2j} \le X_{3k})$ and zero otherwise. For the N-class classification task ($N > 3$), this measure was generalized as

$$\widehat{\theta}_V = (1/n_1 n_2 \cdots n_N) \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_N=1}^{n_N} I(X_{1i_1}, X_{2i_2}, \ldots, X_{Ni_N}),$$

where n_i, $i = 1, 2, \ldots, N$, are the sample sizes of the N classes and the function $I(X_{1i_1}, X_{2i_2}, \ldots, X_{N-1i_{N-1}}, X_{Ni_N})$ is defined in the same manner as that for the three-class task. Nakas and Yiannoutsos also present methods to measure the variance of the term $\widehat{\theta}_V$ by using *U*-statistics methodology. Note that the *U*-statistics approach is the same as that developed by DeLong *et al.* [16] for a two-class case and by Dreiseitl *et al.* [17] for the three-class case. A method to compare the VUS measurements from two decision variables was also developed.

9) Mossman Three-Way ROC: Mossman [8] defined a method for three-class ROC analysis in which three decision variables y_1, y_2, and y_3 are considered. These variables are subject to two constraints: $y_1 + y_2 + y_3 = 1$ and $0 \le y_i \le 1\ \forall i$. One example of such a set of decision variables is the *a posteriori* probabilities (for each class) in a three-class classification task. Thus, the input for this method is of type A (Section II-C). Following the description by Edwards and Metz [18], the decision rule considered by Mossman depends on two parameters α and β, and it can be stated as

$$\begin{aligned}
&\text{decide class 1,} && \text{iff } y_2 - y_1 \le \beta \text{ and } y_3 \le \alpha \\
&\text{decide class 2,} && \text{iff } y_2 - y_1 \ge \beta \text{ and } y_3 \le \alpha \\
&\text{decide class 3,} && \text{iff } y_3 \ge \alpha \quad (\text{where } 0 \le \alpha \le 1, -1 \le \beta \le 1).
\end{aligned}$$

In Mossman's method, the correct identification rates for each class are plotted on the same plot to form a 3-D ROC surface analogous to the ROC curve. The VUS can be used to assess classifier performance in a way similar to the AUC. Information gain, likelihood ratios, and the estimate bias and standard errors of the classifier are all easily calculated.

10) Scurfield's Method: Scurfield's method [13] can be used for three or more classes. In this method, a single decision variable is required as input (type B, Section II-C) and six ROC surfaces are generated (Table IV) for a three-class classification task. These six surfaces (see [13], Fig. 2]) describe all the ways in which the three classification outputs o_1, o_2, and o_3 can be paired with the true states of the objects c_1, c_2, and c_3. A description of each of these surfaces and the conditions each of these surfaces represents is provided in Table IV. Note that the volume of these six surfaces must sum to one. (In general, with this method, *N!* ROC surfaces are created for an *N*-class classification task.) Given a set of ROC surfaces, Scurfield also defined a new measure of discriminability based on the entropy of the volumes under the set of ROC surfaces. For the three-class case, let this measure be denoted by $D_{1:2:3}$. Let $\mathrm{vol}_1, \mathrm{vol}_2, \ldots, \mathrm{vol}_6$ represent the volumes under the six ROC surfaces. $D_{1:2:3}$ is defined as $D_{1:2:3} = \log(6) - H_{1:2:3}$, where $H_{1:2:3} = -\sum_i \mathrm{vol}_i \times \log(\mathrm{vol}_i)$. Highly separable events are indicated by higher values of $D_{1:2:3}$. The benefits of using $D_{1:2:3}$ as a measurement of classifier performance are that it is nonparametric, independent of observer criteria, and finite. Compared to methods that only look at one ROC surface, the six surfaces of Scurfield's method can illustrate the effect of the different misclassifications and their interactions.

11) Sahiner's Method: Chan and coworkers [19]–[21] are interested in the problem of classifying a given observation as malignant, benign, or normal. For this purpose, Sahiner *et al.* [20], [21] have proposed a decision rule based on an ideal observer model as this model has been extensively analyzed. Edwards *et al.* [22] showed that an *N*-class ideal observer makes classifications (decisions) by partitioning a likelihood ratio decision space. The borders of the partitions are given by hyperplanes, which are defined as follows:

$$\begin{aligned}
&\text{decide class} = c_i \\
&\text{iff} \sum_{k=1}^{N-1} \left(U_{i|k} - U_{j|k}\right) P(c_k)\, \mathrm{LR}_k \ge \left(U_{j|N} - U_{i|N}\right) P(c_N), \\
&\qquad \text{if } j < i \\
&\text{and} \sum_{k=1}^{N-1} \left(U_{i|k} - U_{j|k}\right) P(c_k)\, \mathrm{LR}_k > \left(U_{j|N} - U_{i|N}\right) P(c_N), \\
&\qquad \text{if } j > i
\end{aligned}$$

TABLE IV
SIX 3-D ROC SURFACES PROPOSED BY SCURFIELD (FOR THE THREE-CLASS CLASSIFICATION TASK) AND AN INTERPRETATION FOR EACH OF THESE SURFACES

ROC surface	Description	Interpretation
123-surface	Created by plotting $P(o_1 \mid c_1)$, $P(o_2 \mid c_2)$ and $P(o_3 \mid c_3)$.	Measures likelihood that all three classes are correctly classified.
132-surface	Created by plotting $P(o_1 \mid c_1)$, $P(o_2 \mid c_3)$ and $P(o_3 \mid c_2)$	Measures likelihood that c_1 is correctly classified and observations from c_2 are classified as c_3 and vice versa.
213-surface	Created by plotting $P(o_1 \mid c_2)$, $P(o_2 \mid c_1)$ and $P(o_3 \mid c_3)$	Measures likelihood that c_3 is correctly classified and observations from c_1 are classified as c_2 and vice versa.
231-surface	Created by plotting $P(o_1 \mid c_2)$, $P(o_2 \mid c_3)$ and $P(o_3 \mid c_1)$	Measures likelihood that all classes are incorrectly classified. Observations from class 2 are classified as c_1; those from c_3 as c_2; those from c_1 as c_3.
312-surface	Created by plotting $P(o_2 \mid c_1)$, $P(o_1 \mid c_3)$ and $P(o_3 \mid c_2)$	Measures likelihood that all classes are incorrectly classified. Observations from c_1 are classified as c_2; those from c_3 as c_1; those from c_3 as c_2.
321-surface	Created by plotting $P(o_2 \mid c_2)$, $P(o_1 \mid c_3)$ and $P(o_3 \mid c_1)$	Measures likelihood that c_2 is correctly classified and observations from c_1 are classified as c_3 and vice versa

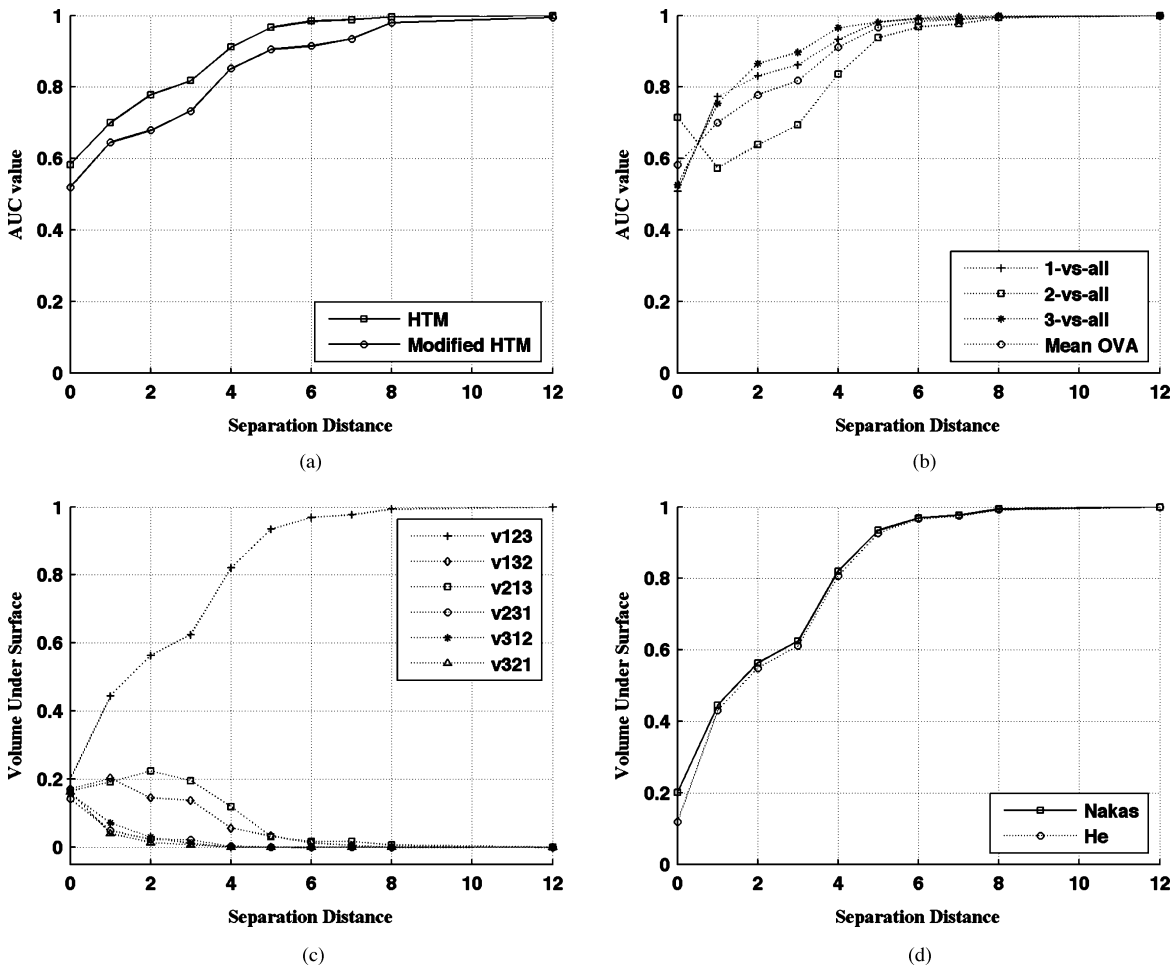

Fig. 2. Plots of VUS/AUC (for various three-class ROC analysis methods) versus separation distance between the classes in the simulated datasets. (a) HTM and modified HTM. (b) OVA. (c) Scurfield's method (volume of the six surfaces). (d) Nakas and He methods.

where $U_{i|j}$ is the utility of deciding an observation is from class c_i, given that it is actually from class c_j. $P(c_k)$ is the *a priori* probability that an observation is drawn from class c_k and LR_k is the kth likelihood ratio defined by the ratio $p(\vec{x}|c_k)/p(\vec{x}|c_N)$ of the probability density functions of the observational data. For purpose of normalization, Sahiner *et al.* [20] assume that $0 \leq U_{i|j} \leq 1$ and also the following constraints. Let M, B, and N denote the malignant, benign, and normal classes. In terms of classes $c_1, c_2,$ and c_3 used in the previous sections, $c_1 \equiv M$, $c_2 \equiv B$, and $c_3 \equiv M$.

$U_{M|M} = 1$, $U_{B|B} = 1$, $U_{N|N} = 1$: all correct decisions have maximum utility (1)

$U_{B\|M} = 0,\ U_{N\|M} = 0:$	utility of misdiagnosing a malignant case as benign/normal is minimum (0)
$U_{B\|N} = 1,\ U_{B\|N} = 1:$	classifying a benign case as normal or vice versa is not harmful
$U_{M\|B} \in (0,1),$ $U_{M\|N} \in (0,1):$	utility of misdiagnosing a benign /normal case as malignant is variable.

With these simple but realistic constraints, the ideal observer model for the three-class task is simplified considerably and provides a single decision boundary in the likelihood ratio plane separating the malignant from "nonmalignant" decisions [20], [21]. In addition, Sahiner *et al.* constructed a 3-D ROC surface by plotting the sensitivity (P_{MM}) versus the two false-positive fractions (P_{MN} and P_{MB}). Note that this is different from all other methods where the 3-D ROC surface is created by plotting the three sensitivities P_{MM}, P_{BB}, and P_{NN}. Finally, Sahiner *et al.* define a normalized volume under the surface (NVUS) for this particular 3-D ROC surface and provide two methods for computing the NVUS under the condition that the three distributions in the feature space are multivariate normal with equal covariance matrices [20], [21]. Note that the input for this case are two log likelihood ratios [20], [21]. For a given object, these can be obtained from the posterior probabilities for each class for that object.

12) He's Method: He and Frey [23] and He *et al.* [24] also use the three-class ideal observer model to develop an ROC analysis method for three-class classification tasks. The motivating problem for their work arose from myocardial perfusion single photon emission computed tomography (SPECT), where the goal was to distinguish normal, infarcted, and ischemic tissue. The input to this method are two log likelihood ratios [23], [24]. For a given object, these can be obtained from the posterior probabilities for each class for that object. For the ROC analysis task, note that as described in the previous section, the utility function $U_{i|j}$, which is the utility of deciding an observation is from class c_i given that it is actually from class c_j, must be specified for the ideal observer analysis. He and Frey [23] and He *et al.* [24] make the following simplifying assumptions about the utility functions:

$$U_{3|1} - U_{2|1} = 0 \qquad U_{3|2} - U_{1|2} = 0 \qquad U_{1|3} - U_{2|3} = 0$$
$$\text{and} \qquad U_{i|i} \geq U_{i|j}, \qquad i,j = 1,2,3, \qquad i \neq j.$$

The first assumption is described as the "equal error utility assumption" since it assumes that under all hypotheses, wrong decisions have equal utilities.

III. Results

A. Experiment 1: Simulated Datasets

In this experiment, we studied the behavior of the various indexes under specific conditions. For this task, we used the two simulated datasets described in Section II-A1. For the first dataset, for each set of values of μ_1, μ_2, and μ_3, we randomly select 50 samples from each class and then classify these samples with a Bayes classifier using a leave-one-out sampling approach. In Fig. 2, for each classification method, the value of VUS/AUC is plotted against the separation distance ("dist," as defined in Section II-A). In the second simulated dataset, two situations were considered: 1) there is complete overlap between the features of classes c_1 and c_2 while class c_3 is completely separated from these classes ($\mu_1 = 1, \mu_2 = 1, \mu_3 = 7$, and $\sigma_1 = \sigma_2 = \sigma_3 = 1$); 2) there is complete overlap between the features of classes c_2 and c_3 while class c_1 is completely separated from these classes ($\mu_1 = 1, \mu_2 = 7, \mu_3 = 7$, and $\sigma_1 = \sigma_2 = \sigma_3 = 1$). For both situations, 50 samples are selected from each class and classified with a Bayes classifier using a leave-one-out sampling approach. This experiment was repeated ten times, and the mean and standard deviation of the various indexes are reported in Table V.

Recall that for a three-class experiment, vol_4 and vol_5 represent the likelihood that all classes are incorrectly classified (Table IV). In this simulated experiment, in both situations, one class (c_3 or c_1) is separate from the other two classes. Hence, $\mathrm{vol}_4 = 0$ and $\mathrm{vol}_5 = 0$ for both situations (Table V). The parameter vol_6 measures the likelihood that samples from class c_2 are correctly classified and observations from class c_1 are classified as c_3, and vice versa. In this experiment, c_2 overlaps with c_1 (case 1) or c_2 overlaps with c_3 (case 2). This implies that samples from class c_2 cannot be correctly classified, and hence, $\mathrm{vol}_6 = 0$ for both cases. The key issue is that only with the Scurfield, pairwise, and OVA methods can one observe the differences in these situations (Table V). Note that the HTM, Nakas, and He methods provide similar measures for both situations and do not provide insight on the source of the classification error.

B. Experiment 2: Fisher Iris Dataset

It is known that each feature of the Fisher Iris dataset provides unique information about each observation. Thus, for any given classifier, we expected the VUS to increase as the number of features were increased from 1 to 4 for the three-class classification task. In this experiment, we used a Bayes classifier, and the results are shown in Table VI. From Table VI, we observe that as the number of features are increased, the VUS/AUC measures increase as expected.

Most of the three-class ROC methods provided a global view of the classification performance (i.e., a single VUS value). However, note that the pairwise comparison approach and Scurfield's methods were the only approaches that provided a detailed account of the classification performance. For example, consider the case when only first two features were used. For Scurfield's method, we see that when only the first two features are used, $\mathrm{vol}_1 = 0.68$ and $\mathrm{vol}_2 = 0.30$ (Table VI, column 3). Note that vol_1 is the volume under the 123-ROC surface that is obtained by plotting the three correct sensitivities $P(o_1|c_1)$, $P(o_2|c_2)$, and $P(o_3|c_3)$ (Table IV [13]).

Similarly, vol_2 is the volume under the 132-ROC surface that is created by plotting $P(o_1|c_1)$, $P(o_2|c_3)$, and $P(o_3|c_2)$ (Table IV [13]). Recall that vol_1 denotes the likelihood of classifying all classes correctly and vol_2 measures the likelihood of incorrectly classifying samples from class c_2 as from class c_3, and vice versa (Table IV). Fig. 3(a) shows the distribution of

TABLE V
RESULTS (VUS/AUC) FROM VARIOUS THREE-CLASS ROC ANALYSIS METHODS

Method		Volume Under Surface/ Area Under the Curve			
		Case 1: $\mu_1 = 1,\ \mu_2 = 1,\ \mu_3 = 7$ $\sigma_1 = \sigma_2 = \sigma_3 = 1$		Case 2: $\mu_1 = 1,\ \mu_2 = 7,\ \mu_3 = 7$ $\sigma_1 = \sigma_2 = \sigma_3 = 1$	
		Mean value (std. dev.)		Mean value (std. dev.)	
1.Pairwise comparisons	c_1 vs. c_2	0.55(0.04), 0.55(0.04)		1.00(0), 1.00(0)	
	c_2 vs. c_3	1.00(0), 1.00(0)		0.56(0.04), 0.56(0.04)	
	c_1 vs. c_3	1.00(0), 1.00(0)		1.00(0), 1.00(0)	
2. Hand & Till M fn		0.85(0.01)		0.85(0.01)	
3. One versus All (OVA)	c_1 vs. all	0.75(0.03)		1.00(0)	
	c_2 vs. all	0.75(0.04)		0.74(0.03)	
	c_3 vs. all	1.00(0)		0.74(0.03)	
	Mean	0.83(0.02)		0.83(0.02)	
4.Modified HTM		0.76(0.02)		0.77(0.01)	
5.Scurfield's Method		vol_1= 0.53(0.04)	vol_2 = 0(0)	vol_1 = 0.53(0.06)	vol_2 = 0.47(0.05)
		vol_3= 0.47(0.04)	vol_4 = 0(0)	vol_3 = 0(0)	vol_4 = 0(0)
		vol_5 = 0(0)	vol_6= 0(0)	vol_5= 0(0)	vol_6 = 0(0)
6. Nakas's Method		0.53(0.04)		0.53(0.06)	
7. He's Method		0.51(0.07)		0.49(0.07)	

In this experiment, a Bayes classifier was used to classify observations from the second simulated dataset. The third column shows the results for the case when there is overlap between the features of classes c_1 and c_2 while the features of class c_3 is separated from classes c_1 and c_2. The fourth column shows the results for the case of overlap between the features of classes c_2 and c_3 while class c_1 is completely separated from classes c_2 and c_3. Note that vol_1: volume of 123-surface; vol_2: volume of 132-surface; vol_3: volume of 213-surface; vol_4: volume of 231-surface; vol_5: volume of 312-surface; vol_6: volume of 321-surface [13].

TABLE VI
RESULTS (VUS/AUC) FROM VARIOUS THREE-CLASS ROC ANALYSIS METHODS

Method		Volume Under Surface/ Area Under the Curve							
		Features used							
		1		1 and 2		1, 2 and 3		1, 2, 3 and 4	
		AUCs		AUCs		AUCs		AUCs	
1.Pair-wise Compa-risons	c_1 vs. c_2	0.92, 0.83		1.00, 0.98		1.00, 1.00		1.00, 1.00	
	c_2 vs. c_3	0.67, 0.77		0.76, 0.77		0.99, 0.99		0.99, 0.99	
	c_1 vs. c_3	0.98, 0.98		1.00, 1.00		1.00, 1.00		1.00,1.00	
2. Hand & Till M fn		0.86		0.92		1.00		1.00	
3. One versus All (OVA)	c_1 vs. all	0.95		1.00		1.00		1.00	
	c_2 vs. all	0.75		0.87		0.99		1.00	
	c_3 vs. all	0.88		0.89		0.99		1.00	
	Mean	0.86		0.92		1.00		1.00	
4.Modified HTM		0.81		0.85		0.97		0.985	
5.Scurfield's Method		vol_1= 0.68	vol_2= 0.20	vol_1= 0.68	vol_2= 0.30	vol_1= 0.92	vol_2= 0.08	vol_1= 0.98	vol_2= 0.02
		vol_3 = 0.10	vol_4< 0.01	vol_3= 0.01	vol_4< 0.01	vol_3= 0.0	vol_4= 0.0	vol_3= 0.0	vol_4= 0.0
		vol_5= 0.01	vol_6< 0.01	vol_5< 0.01	vol_6< 0.01	vol_5 = 0.0	vol_6= 0.0	vol_5 = 0.0	vol_6= 0.0
		vol_T = 1.00	$D_{1:2:3}$= 1.27	vol_T = 1.00	$D_{1:2:3}$ = 1.58	vol_T = 1.00	$D_{1:2:3}$ = 2.17	vol_T = 1.00	$D_{1:2:3}$ = 2.43
6. Nakas's Method		VUS = 0.68		VUS = 0.68		VUS = 0.92		VUS = 0.98	
7. He's Method		VUS = 0.69		VUS = 0.77		VUS = 0.988		VUS = 0.994	

In this experiment, a Bayes classifier was used to classify observations from the Fisher Iris dataset. It is known that each feature in this dataset provides unique information. Thus, for any given classifier, we expected the VUS to increase as the number of features used were increased from 1 to 4 for the three-class classification task. Note that vol_1: volume of 123-surface; vol_2: volume of 132-surface; vol_3: volume of 213-surface; vol_4: volume of 231-surface; vol_5: volume of 312-surface; vol_6: volume of 321-surface [13]. vol_T is the sum of all of these six ROC surfaces.

samples in a 2-D feature space (features 1 and 2). We see that there is significant overlap between classes c_2 and c_3. This is captured by the measure vol_2. Also note that vol_3 measures the likelihood that samples from class c_1 are classified as from c_2 and vice versa. Similarly, vol_6 measures the chance of mislabeling samples from class c_1 as from c_3 and vice versa. From Fig. 3(a), it is clear that even with only the first two features, this is very unlikely, and thus, both vol_3 and vol_6 are ≤ 0.01 (Table VI, column 3).

Fig. 3(b) shows how all samples are distributed in a 2-D feature space (features 1 and 3) and Fig. 3(d) shows how all samples are distributed in a 3-D feature space (features 1–3). From these figures, we see that feature 3 helps to differentiate between classes c_2 and c_3. Thus, when we use all three features, vol_1 increases to 0.92 and vol_2 decreases to 0.08, as compared to $\mathrm{vol}_2 = 0.30$ when only two features were used (Table VI, columns 3 and 4). Thus, we can quantitatively show that feature 3 helps in discrimination of classes c_2 and c_3.

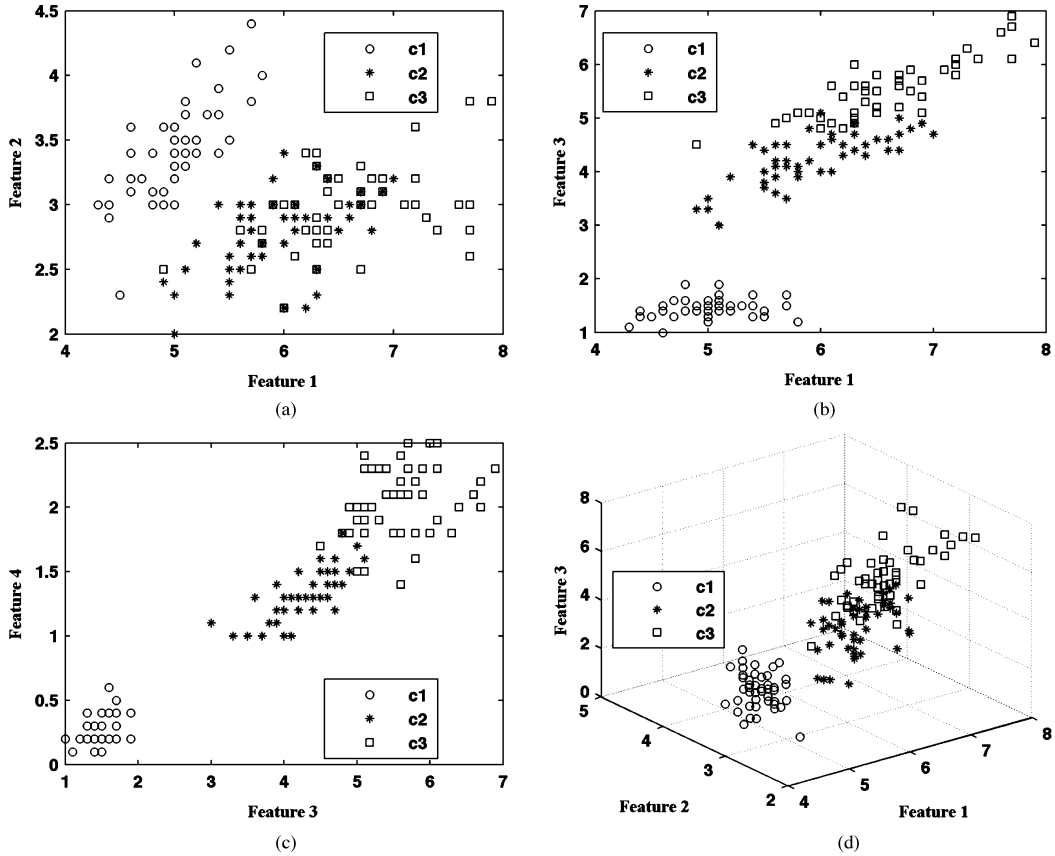

Fig. 3. Distribution of the features from the Iris dataset. (a) Distribution of samples in the 2-D feature space (features 1 and 2). (b) Distribution in 2-D feature space (features 1 and 3). (c) Distribution in 2-D feature space (features 2 and 4). (d) Distribution in 3-D feature spaces (features 1–3).

Note that the fact that feature 3 helps to distinguish between classes c_2 and c_3 was easily obtained from the results from Scurfield's method. To some degree, similar information can also be found by looking at the results from the pairwise comparisons (Table VI, columns 3 and 4). We see that the lowest AUC was obtained (0.76, 0.77) when class c_2 was compared to class c_3, and only features 1 and 2 were used. This indicates that the two features do not provide adequate discrimination between these classes. In comparison, when features 1–3 are used for the pairwise comparison of classes c_2 and c_3, an AUC of 0.99 is obtained. This shows that feature 3 helps in the discrimination of classes 2 and 3. However, the pairwise comparisons do not provide an indication of the overall classification accuracy, which is provided by the volume of the 123-ROC surface (vol_1) from Scurfield's method. Note that the volume of the 123-ROC surface and the VUS from Nakas method are identical. This experiment shows that by analyzing all six ROC surfaces, one can obtain a more detailed description of the utility of each feature to discriminate between each pair of classes.

C. Experiment 3: BI-RADS Mammography Dataset

In this experiment, observations from the BI-RADS mammography dataset were classified using all three classifiers described in Section II-B. These results are shown in Table VII. In this classification task, the goal was to predict the radiologist's gut assessment of the likelihood of malignancy for each breast lesion. Note that while the Bayes classifier did perform well for distinguishing some of the two class pairs (for example "3" versus "5"), none of the classifiers provide good discrimination between all three classes. From the results from Scurfield's method, we see that the overall classification accuracy denoted by vol_1 is 0.15, 0.12, and 0.10 for the *k*-NN, linear regression, and Bayes classifiers, respectively (Table VII).

The Scurfield method allows one to look at the type of classification error. Note that the sum of the two volumes, vol_4 and vol_5, measures the likelihood of classifying all classes incorrectly (Table IV). The sum of these measures is 0.35, 0.38, and 0.41 for each of the three classifiers respectively (Table VII). Note that in this experiment, for some of the cases, the results from the pairwise comparisons are reasonably good. For example, the Bayes classifier performs well in differentiating between classes "3" and "5" (AUC = 0.93; Table VII, last column). However, it is difficult to get an assessment of the overall classification accuracy from only looking at the results of the pairwise comparisons.

D. Experiment 4: Spectroscopy Dataset

In this experiment, observations from the UT spectroscopy dataset were classified using all three classifiers described in Section II-B. These results are shown in Table VIII. In this

TABLE VII
RESULTS (VUS/AUC) FROM VARIOUS THREE-CLASS ROC ANALYSIS METHODS

Method		**Volume Under Surface/ Area Under the Curve**					
		Classifier used					
		K-Nearest neighbor		Linear regression		Bayes classifier	
		AUCs		AUCs		AUCs	
1.Pairwise Comparisons	"3" vs. "4"	0.73, 0.58		0.60, 0.66		0.83, 0.73	
	"4" vs. "5"	0.58, 0.60		0.70, 0.72		0.58, 0.73	
	"3" vs. "5"	0.83, 0.84		0.55, 0.62		0.93, 0.93	
2. Hand & Till M function		0.69		0.64		0.79	
3. One-versus-All (OVA)	"3" vs. all	0.78		0.52		0.89	
	"4" vs. all	0.51		0.68		0.67	
	"5" vs. all	0.75		0.65		0.85	
	Mean	0.68		0.62		0.80	
4. Modified HTM		0.61		N/A		0.73	
5. Scurfield's Method		vol_1= 0.15	vol_2= 0.15	vol_1 = 0.12	vol_2 = 0.19	vol_1= 0.10	vol_2 = 0.17
		vol_3 = 0.14	vol_4 = 0.21	vol_3 = 0.11	vol_4 = 0.18	vol_3 = 0.13	vol_4= 0.20
		vol_5 = 0.14	vol_6 = 0.21	vol_5 = 0.20	vol_6= 0.20	vol_5 = 0.21	vol_6= 0.20
		vol_T = 1.00	$D_{1:2:3}$ = 0.02	vol_T = 1.00	$D_{1:2:3}$ = 0.04	vol_T = 1.00	$D_{1:2:3}$ = 0.05
6. Nakas's Method		VUS = 0.13		VUS = 0.12		VUS = 0.10	
7. He's Method		VUS = 0.09		VUS = 0.20		VUS = 0.12	

In this experiment, observations from the Duke BI-RADS mammography dataset were classified using all three classifiers described in Section H-B. Note that vol_1: volume of 123-surface; vol_2: volume of 132-surface; vol_3: volume of 213-surface; vol_4: volume of 231-surface; vol_5: volume of 312-surface; vol_6: volume of 321-surface [13]. vol_T is the sum of all six ROC surfaces.

TABLE VIII
RESULTS (VUS/AUC) FROM VARIOUS THREE-CLASS ROC ANALYSIS METHODS

Method		**Volume Under Surface/ Area Under the Curve**					
		Classifier used					
		k-Nearest neighbor		Linear regression		Bayes classifier	
1.Pairwise comparisons	c_1 vs. c_2	0.59, 0.56		0.64, 0.52		0.59, 0.75	
	c_2 vs. c_3	0.67, 0.51		0.61, 0.64		0.81, 0.57	
	c_1 vs. c_3	0.68, 0.64		0.57, 0.67		0.70, 0.72	
2. Hand & Till M function		0.60		0.61		0.69	
3. One-versus-All (OVA)	c_1 vs. all	0.64		0.59		0.66	
	c_2 vs. all	0.61		0.54		0.78	
	c_3 vs. all	0.59		0.66		0.66	
	Mean	0.61		0.60		0.70	
4. Modified HTM		0.53		N/A		0.562	
5. Scurfield's Method		vol_1 = 0.26	vol_2= 0.23	vol_1 = 0.25	vol_2= 0.20	vol_1= 0.27	vol_2 = 0.23
		vol_3 = 0.21	vol_4 = 0.10	vol_3= 0.15	vol_4 = 0.14	vol_3 = 0.27	vol_4 = 0.10
		vol_5= 0.14	vol_6 = 0.07	vol_5= 0.14	vol_6= 0.12	vol_5= 0.09	vol_6= 0.05
		vol_T = 1.00	$D_{1:2:3}$ = 0.13	vol_T = 1.00	$D_{1:2:3}$ =0.05	vol_T = 1.00	$D_{1:2:3}$ = 0.23
6. Nakas's Method		VUS = 0.18		VUS = 0.25		VUS = 0.27	
7. He's Method		VUS = 0.15		VUS = 0.19		VUS = 0.19	

In this experiment, observations from the UT spectroscopy dataset were classified using all three classifiers described in Section II-B. Note that vol_1: volume of 123-surface; vol_2: volume of 132-surface; vol_3: volume of 213-surface; vol_4: volume of 231-surface; vol_5: volume of 312-surface; vol_6: volume of 321-surface [13]. vol_T is the sum of all of these six ROC surfaces.

classification task, the goal was to classify tissue sites into three categories as normal, benign, or dysplasia. (The cases of mild dysplasia (MD) and severe high grade dysplasia were combined into one category.)

The results indicate that these sets of features do not provide good classification between the various classes. From the results obtained with Scurfield's method, we see that the overall classification accuracy denoted by vol_1 is 0.26, 0.25, and 0.27 for the *k*-NN, linear regression, and Bayes classifiers, respectively (Table VIII). The sum of these volumes, vol_4 and vol_5 (which measure the likelihood of classifying all classes incorrectly), is 0.24, 0.28, and 0.19 for each of the three classifiers respectively (Table VIII).

Note that vol_3 quantifies the likelihood of incorrectly classifying samples from class c_1 as those from class c_2 and vice versa (Table IV). In this experiment, vol_3 is 0.21, 0.15, and 0.27 for each of the three classifiers respectively (Table VIII). Thus, from Scurfield's method, we can infer the type of error that is

likely to occur for a given classification technique. Often, certain kinds of errors have a "higher cost" associated with them than others. In classification experiments, this is usually quantified with a cost matrix. For an *N*-class classification task, the cost matrix is an $N \times N$ matrix in which each term cost_{ij} represents the cost of classifying a sample as belonging to class c_i when it actually belongs to class c_j [10]. As different classification techniques may have different likelihoods of making a particular type of error, by using the Scurfield method, one can find the classification technique that has the smallest likelihood for the error with the "highest cost."

Comparing the various indexes from experiments 3 and 4, we observe some interesting trends. In experiment 3, the AUCs from the pairwise comparisons are higher than those for experiment 4. Similarly, the HTM and modified HTM indexes are also larger for experiment 3. However, the Scurfield, He, and Nakas indexes for experiment 4 are greater than those for experiment 3. Note that these indexes quantify the discriminability between all three classes simultaneously. This further shows that it is difficult to get an assessment of the overall classification accuracy from only looking at the results of the pairwise comparisons.

IV. Discussion

The intention of this study was to summarize methods for multiclass performance analysis and provide insight into the different methods via empirical comparisons. All methods were used to evaluate the performance of three popular classification techniques (*k*-NN classifier, linear regression classifier, and Bayes classifier) in four experiments. In the following sections, we first present the key points from the experiments and then discuss the advantages and disadvantages of each of the three-class ROC methods.

A. Advantages and Disadvantages of the Various Three-Class ROC Methods

As demonstrated by the experiments before, we see that the various three-class methods have certain advantages and disadvantages. First we consider the methods that analyze the discrimination of two classes at a time (pairwise, HTM function, OVA, and modified HTM). The results from experiment 1 (Table V) show the major limitations of the Hand and Till function and the modified HTM method. For the two simple simulated datasets considered in this experiment, these methods generate the same results and may be misleading. In contrast, pairwise comparisons allow for an in-depth view of classifier performance. The advantage is that problem areas can be pinpointed, but the disadvantage is that as the number of classes increases, there are numerous AUC results (for *N* classes, one obtains $2N\,(N-1)$ AUCs), and this may make it cumbersome to interpret. Note that the OVA comparisons allow for a slightly less detailed view of classifier performance than pairwise comparisons. However, note that both the pairwise and OVA comparisons do not indicate how a given method would perform for classifying all classes simultaneously.

Among the methods that analyze the discrimination of all three classes, the results from the experiments show that the VUS from the Nakas, He, and Mossman methods are similar to the volume of the 123-surface in Scurfield's method. A key aspect is that the Scurfield method provides the most detailed description of classifier performance. For the three-class classification task, the six volumes from the Scurfield method provide information on all potential scenarios and can provide the investigator with valuable information on the sources of the classification error. For example, the volume under the 321-surface denotes the likelihood that observations from class 3 are misclassified as those from class 1 and vice versa. In addition, one can evaluate the effectiveness of certain features for a specific classification task. For example, from experiment 2 on the Iris Fisher dataset (Table VI), we see that when only the first two features are used, observations from class 2 are misclassified as those from class 3 (volume of 123-surface is 0.68 and volume of 132-surface is 0.30), and when three features are used, this helps to discriminate between classes 2 and 3 and improve the overall classification performance (volume of the 123-surface is 0.92 and the volume of the 132-surface reduces to 0.08). The limitation of the Scurfield method is that it does not present methods to quantify the variance of the volumes of the various ROC surfaces. Also, no guidelines exist to compare the results from two classification methods (e.g., volumes of two 123-surfaces). In comparison, the Nakas method also allows one to measure the variance of the VUS estimate and compare the VUS measurements from two decision variables. Similarly, the He method also provides two techniques to measure the variance of the VUS estimate. In situations when both the VUS and the variance of the VUS estimate are desired, we recommend the use of the He method when the classifier's output consists of a set of posterior probabilities for each class. For the case where the classifier's output is a single continuous decision variable, we recommend the use of the Nakas method. Among the various three-class ROC indexes compared in this paper, those developed by Mossman [8], Nakas and Yiannoutsos [9], and Dreiseitl *et al.* [17] have been most commonly used in various medical imaging studies [25]–[29]. One reason for their popularity is the ease of implementation, whereas, in comparison, the index proposed by Scurfield [13] is relatively more difficult to compute. Yiannoutsos *et al.* [25] used 3-D ROC surfaces and the corresponding VUS to differentiate between HIV-negative, HIV-positive neurologically asymptomatic patients, and patients with AIDS demential complex using brain metabolites quantified by proton magnetic resonance spectroscopy (MRS) [25]. Alonzo and Nakas [26] used the VUS for the determination of diagnostic markers for lung. Binder *et al.* [27] used the VUS to differentiate between melanomas from benign pigmented skin lesions. In this three-class problem, the goal was to distinguish three classes of lesions. Extending this work, Dreiseitl *et al.* [28] compared the performance of five different classification techniques for the classification of pigmented skin lesions into three classes. The VUS was used as the index for comparing classifiers. Ratnasamy *et al.* [29] studied heart failure in children. They categorized their patients into three groups based on disease severity and used the VUS to assess how accurately certain blood markers reflected clinical severity [29].

As a number of three-class ROC indexes exist, it can be difficult to select the most suitable evaluation methodology for a given classification task. This paper and the corresponding toolbox will allow researchers to compare the various three-class ROC indexes and will help them to select the most appropriate evaluation methodology for their application.

V. Conclusion

The intention of this study was to summarize methods for multiclass performance analysis and provide insight into the different methods via empirical comparisons. All methods were used to evaluate the performance of three popular classification techniques in four experiments. We recommend the use of the Scurfield method as it provides the investigator with the most detailed description of classifier performance. In addition, it can be used to test the effectiveness of features, and it also provides insight about the sources of error in a given classification task. However, the Scurfield method does not provide an estimate of the variance of the six ROC surfaces. The He and Nakas methods are particularly useful when both the VUS and an estimate of the variance of the VUS are desired. Thus, in such scenarios, we suggest the use of the He or Nakas method, depending on the type of classifier output.

Acknowledgment

The BI-RADS mammography data were compiled by Duke Advanced Imaging Laboratories and particular thanks are due to late C. E. Floyd, Jr., for this contribution. The spectroscopy data were compiled at the Head and Neck Clinic, University of Texas M. D. Anderson Cancer Center, by Dr. K. Sokolov and Dr. L. Nieman, and the authors thank them for their contributions. The authors would like to thank Dr. X. He for helpful discussions on the various three-class ROC methods. They also thank Dr. B. Sahiner for providing details about their method, and C. Kite, Z. Mahdavi, and S. Swanson for technical support.

References

[1] C. E. Metz, "Basic principles of ROC analysis," *Semin. Nucl. Med.*, vol. 8, no. 4, pp. 283–298, 1978.

[2] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annu. Eugenics*, vol. 7, pp. 179–188, 1936.

[3] American College of Radiology, *ACR BI-RADS—Mammography, Ultrasound & Magnetic Resonance Imaging*, 4th ed. Reston, VA: Amer. College of Radiol., 2003.

[4] M. K. Markey, J. Y. Lo, G. D. Tourassi, and C. E. Floyd, Jr., "Self-organizing map for cluster analysis of a breast cancer database," *Artif. Intell. Med.*, vol. 27, no. 2, pp. 113–127, 2003.

[5] L. T. Nieman, C. W. Kan, A. Gillenwater, M. K. Markey, and K. Sokolov, "Probing local tissue changes in the oral cavity for early detection of cancer using oblique polarized reflectance spectroscopy: A pilot clinical trial," *J. Biomed. Opt.*, vol. 13, no. 2, p. 024011, 2008.

[6] A. Myakov, L. Nieman, L. Wicky, U. Utzinger, R. Richards-Kortum, and K. Sokolov, "Fiber optic probe for polarized reflectance spectroscopy in vivo: Design and performance," *J. Biomed. Opt.*, vol. 7, no. 3, pp. 388–397, 2002.

[7] K. Sokolov, L. T. Nieman, A. Myakov, and A. Gillenwater, "Polarized reflectance spectroscopy for pre-cancer detection," *Technol. Cancer Res. Treat.*, vol. 3, no. 1, pp. 1–14, 2004.

[8] D. Mossman, "Three-way ROCs," *Med. Decis. Making*, vol. 19, no. 1, pp. 78–89, 1999.

[9] C. T. Nakas and C. T. Yiannoutsos, "Ordered multiple-class ROC analysis with continuous measurements," *Stat. Med.*, vol. 23, no. 22, pp. 3437–3449, 2004.

[10] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley–Interscience, 2000.

[11] D. J. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Mach. Learning*, vol. 45, no. 2, pp. 171–186, 2001.

[12] C. Ferri, J. Hernandez-Orallo, and M. A. Salido, *Volume Under the ROC Surface for Multi-Class Problems. Exact Computation and Evaluation of Approximations*. Valencia, CA: Univ. Politecnica de Valencia, 2003, pp. 1–40.

[13] B. K. Scurfield, "Multiple-event forced-choice tasks in the theory of signal detectability," *J. Math. Psychol.*, vol. 40, no. 3, pp. 253–269, 1996.

[14] X. He and E. C. Frey, "The meaning and use of the volume under a three-class ROC surface (VUS)," *IEEE Trans. Med. Imag.*, vol. 27, no. 5, pp. 577–588, May 2008.

[15] D. C. Edwards and C. E. Metz, "Optimization of restricted ROC surfaces in three-class classification tasks," *IEEE Trans. Med. Imag.*, vol. 26, no. 10, pp. 1345–1356, Oct. 2007.

[16] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

[17] S. Dreiseitl, L. Ohno-Machado, and M. Binder, "Comparing three-class diagnostic tests by three-way ROC analysis," *Med. Decis. Making*, vol. 20, no. 3, pp. 323–331, 2000.

[18] D. C. Edwards and C. E. Metz, "Analysis of proposed three-class classification decision rules in terms of the ideal observer decision rule," *J. Math. Psychol.*, vol. 50, no. 5, pp. 478–487, 2006.

[19] H.-P. Chan, B. Sahiner, L. M. Hadjiiski, N. Petrick, and C. Zhou, "Design of three-class classifiers in computer-aided diagnosis: Monte Carlo simulation study," in *Medical Imaging 2003: Image Processing,* vol. 5032. San Diego, CA, USA: SPIE, 2003, pp. 567–578.

[20] B. Sahiner, H.-P. Chan, and L. M. Hadjiiski, "Performance analysis of 3-class classifiers: Properties of the 3D ROC surface and the normalized volume under the surface," in *Medical Imaging 2006: Image Perception, Observer Performance, Technology Assessment*, vol. 6146, San Diego, CA: SPIE, 2006, pp. 87–93.

[21] B. Sahiner, H. P. Chan, and L. M. Hadjiiski, "Performance analysis of three-class classifiers: Properties of a 3-D ROC surface and the normalized volume under the surface for the ideal observer," *IEEE Trans. Med. Imag.*, vol. 27, no. 2, pp. 215–227, Feb. 2008.

[22] D. C. Edwards, C. E. Metz, and M. A. Kupinski, "Ideal observers and optimal ROC hypersurfaces in N-class classification," *IEEE Trans. Med. Imag.*, vol. 23, no. 7, pp. 891–895, Jul. 2004.

[23] X. He and E. C. Frey, "Three-class ROC analysis—The equal error utility assumption and the optimality of three-class ROC surface using the ideal observer," *IEEE Trans. Med. Imag.*, vol. 25, no. 8, pp. 979–986, Aug. 2006.

[24] X. He, C. E. Metz, B. M. Tsui, J. M. Links, and E. C. Frey, "Three-class ROC analysis—A decision theoretic approach under the ideal observer framework," *IEEE Trans. Med. Imag.*, vol. 25, no. 5, pp. 571–581, May 2006.

[25] C. T. Yiannoutsos, C. T. Nakas, and B. A. Navia, "Assessing multiple-group diagnostic problems with multi-dimensional receiver operating characteristic surfaces: Application to proton MR spectroscopy (MRS) in HIV-related neurological injury," *Neuroimage*, vol. 40, no. 1, pp. 248–255, 2008.

[26] T. A. Alonzo and C. T. Nakas, "Comparison of ROC umbrella volumes with an application to the assessment of lung cancer diagnostic markers," *Biomet. J.*, vol. 49, no. 5, pp. 654–664, 2007.

[27] M. Binder, H. Kittler, S. Dreiseitl, H. Ganster, K. Wolff, and H. Pehamberger, "Computer-aided epiluminescence microscopy of pigmented skin lesions: The value of clinical data for the classification process," *Melanoma Res.*, vol. 10, no. 6, pp. 556–561, 2000.

[28] S. Dreiseitl, L. Ohno-Machado, H. Kittler, S. Vinterbo, H. Billhardt, and M. Binder, "A comparison of machine learning methods for the diagnosis of pigmented skin lesions," *J. Biomed. Inf.*, vol. 34, no. 1, pp. 28–36, 2001.

[29] C. Ratnasamy, D. D. Kinnamon, S. E. Lipshultz, and P. Rusconi, "Associations between neurohormonal and inflammatory activation and heart failure in children," *Amer. Heart J.*, vol. 155, no. 3, pp. 527–533, 2008.

Authors' photographs and biographies not available at the time of publication.