Comparison of Machine Learning and Traditional Classifiers in Glaucoma Diagnosis

Kwokleung Chan*, Te-Won Lee, *Associate Member, IEEE*, Pamela A. Sample, Michael H. Goldbaum, Robert N. Weinreb, and Terrence J. Sejnowski, *Fellow, IEEE*

***Abstract*—Glaucoma is a progressive optic neuropathy with characteristic structural changes in the optic nerve head reflected in the visual field. The visual-field sensitivity test is commonly used in a clinical setting to evaluate glaucoma. Standard automated perimetry (SAP) is a common computerized visual-field test whose output is amenable to machine learning. We compared the performance of a number of machine learning algorithms with STATPAC indexes mean deviation, pattern standard deviation, and corrected pattern standard deviation. The machine learning algorithms studied included multilayer perceptron (MLP), support vector machine (SVM), and linear (LDA) and quadratic discriminant analysis (QDA), Parzen window, mixture of Gaussian (MOG), and mixture of generalized Gaussian (MGG). MLP and SVM are classifiers that work directly on the decision boundary and fall under the discriminative paradigm. Generative classifiers, which first model the data probability density and then perform classification via Bayes' rule, usually give deeper insight into the structure of the data space. We have applied MOG, MGG, LDA, QDA, and Parzen window to the classification of glaucoma from SAP. Performance of the various classifiers was compared by the areas under their receiver operating characteristic curves and by sensitivities (true-positive rates) at chosen specificities (true-negative rates). The machine-learning-type classifiers showed improved performance over the best indexes from STATPAC. Forward-selection and backward-elimination methodology further improved the classification rate and also has the potential to reduce testing time by diminishing the number of visual-field location measurements.**

***Index Terms*—Bayes rule, neural network, standard automated perimetry, STATPAC, support vector machine.**

I. Introduction

GLAUCOMA is a progressive optic neuropathy with characteristic structural changes in the optic nerve head reflected in the visual field [1]. Three million people in the United States and as many as 100 million worldwide are affected by glaucoma. It is the second leading cause of blindness in all North Americans.

In the clinical setting, glaucoma is commonly evaluated using visual-field testing or funduscopic examination of the optic disk [2]. Standard automated perimetry (SAP) is currently the visual function test most relied upon to measure visual function in glaucoma. Automated threshold perimetry gives detailed quantitative data. However, even with all the experience that has accumulated for evaluating standard perimetry, sometimes interpreting the results of SAP can be problematic. Early detection often requires interpretation of borderline visual-field results [3]. Separating true vision loss due to glaucoma from fluctuations in the field is extremely difficult and challenging. In this paper, a number of machine learning classifiers will be applied to glaucoma diagnosis from SAP and compared with STATPAC, a specialized statistical analyses package currently employed by clinicians to interpret SAP.

The motivation behind this paper is to develop a better understanding of the machine classification process, to evaluate the classification in terms of receiver operator characteristics (ROCs) curves, and to analyze the weaknesses and strengths of known classifiers to this problem. The detailed analysis allows us to compare the results not only in terms of their accuracy but also in terms of other properties such as training and testing speed, feature selection method, ease of use, and possible interpretation. These issues are important to the application of machine classifiers in glaucoma research and to clinicians and researchers who would like to get an understanding of the classification process and analysis. Similar approaches may also be helpful in diagnosing other diseases.

This paper is outlined as follows: Section II summarizes several discriminative and generative machine classifiers that are used in this study. Section III describes the data-acquisition method and STATPAC that is currently a state-of-the-art method for glaucoma analysis. In Section IV, we describe the training and testing data and the application process to the machine classifiers. In Section V, we evaluate the results in terms of ROC and classification accuracy. Section VI depicts the feature selection methodology for one classifier evaluated with ROC curves for different numbers of features. In Section VII, we discuss the results in comparison to STATPAC, within the machine classifiers and also within the generative and discriminative class of classifiers. We conclude in Section VIII and express our near future research goals within this framework.

II. Machine Classifiers

A. Discriminative and Generative Classification

In a two-class classification problem, we are given a training dataset $\{\mathbf{x}_i, y_i\}$, $i = 1, \ldots, N$ where $\mathbf{x}_i \in \mathcal{R}^D$ (could contain both continuous and discrete entries) is the input and $y_i = \pm 1$ is the output label. When performing classification, one approach

Manuscript received October 3, 2001; revised April 30, 2002. The work of P. A. Sample was supported by the National Insitutes of Health (NIH) under Grant EY08208. The work of M. H. Goldbaum was supported by the NIH under Grant EY13235. *Asterisk indicates corresponding author.*

*K. Chan is with the University of California at San Diego, La Jolla, CA 92093 USA, and also with the The Salk Institute for Biological Studies, La Jolla, CA 92097 USA.(e-mail: kwchansalk.edu).

T.-W. Lee, P. A. Sample, M. H. Goldbaum and R. N. Weinreb are with the University of California at San Diego, La Jolla, CA 92093 USA.

T. J. Sejnowski is with the University of California at San Diego, La Jolla, CA 92093 USA, and also with the The Salk Institute for Biological Studies, La Jolla, CA 92097 USA.

Publisher Item Identifier 10.1109/TBME.2002.802012.

is to first model the class-conditional probability $p(\mathbf{x}|C_\pm)$ for each class $C_\pm$, and then employ the Bayes' rule

$$P(C_\pm|\mathbf{x}) = \frac{p(\mathbf{x}|C_\pm)P(C_\pm)}{p(\mathbf{x})}. \tag{1}$$

Under the Cox Axioms [4], the Bayes' rule is the only consistent way to manipulate beliefs and plausibility, if they are represented by real numbers. Classification using (1) is also known as the generative paradigm, since the probability of generating the data point $\mathbf{x}$ is first modeled. This effectively reduces the problem of classification to that of modeling the class-conditional probability distribution $p(\mathbf{x}|C_\pm)$ for the two classes.

However, it has always been difficult to model $p(\mathbf{x}|C_\pm)$ accurately. Naive Bayes' classifier [5] assumes independency between components of input $\mathbf{x}$. Modeling $P(C_\pm|\mathbf{x})$ through $p(\mathbf{x}|C_\pm)$ is known to be inefficient [6] as it generally requires the estimation of more parameters. Take the example of performing classification by classical linear discriminant analysis (LDA): modeling the two classes of data with Gaussian densities of same variance but different means. It takes $D(D+1)/2 + D + D$ parameters in this approach. The resulting classifier is well known to be a linear discriminant function $u(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} + b$ which only needs $D+1$ parameters. For a dataset of finite size, this means that we have fewer data points for each parameter in the generative approach. Unless the equivariance assumption fits well to the data, the classical LDA will be less efficient, for the sole purpose of classification. On the other hand, the logistic regression [7] makes fewer assumptions about the classes and is generally more robust against outliers and noise in the data. Another weakness of the generative approach is that the model parameters are usually optimized by maximum-likelihood (ML) estimation [8]. It is widely believed that discriminative classifiers are to be preferred since the discriminative criterion is more closely related to classification error.

The above suggests we may be better off using the discriminative approach in which the posterior probabilities $P(C_\pm|\mathbf{x})$ are directly estimated. Logistic regression is a well-known example of the discriminative approach and is widely used in medical research. Decision trees, such as CART [9] or C4.5 [10], are another kind of discriminative classifier. Recently, attention has shifted to neural-network-type classifiers [11], [12] and the support vector machine (SVM) [13]. In some of these classifiers, there estimation of the posterior probabilities is unnecessary. The classifier simply returns the label y by applying discrimination functions on the input $\mathbf{x}$.

The advantage of discriminative classifiers is that they concentrate on the decision boundary and, hence, are usually robust against irrelevant outliers in the training data. However, they provide less insight into the structure of the data space and it is difficult to handle data containing missing entries. The multilayer perceptron (MLP) and SVM often serve as black boxes in classification and it is very difficult for humans to comprehend how the decision is made.

B. MLP

The MLP [11], [14], [15], also termed feedforward network, is a generalization of the single-layer perceptron studied in [16]. The MLP is a universal approximator to any real valued functions. In fact, a feedforward network of just two layers (not including the input layer) can in principle approximate any continuous function [17]. The MLP has been successfully applied to a wide class of problems such as face recognition [18] and optical character recognition [19].

In a two-class classification problem, for a given input $\mathbf{x} = (x_1, \ldots, x_D)^T$

$$z_j = g\left(\sum_{d=1}^{D} w_{jd}x_d + w_{j0}\right) \tag{2}$$

$$f = h\left(\sum_{j=1}^{J} v_j z_j + v_0\right) \tag{3}$$

$z_j,\ j = 1, \ldots, J$ are the activations of the hidden-layer units. w_{jd} are the weights between the input and the hidden layer. Similarly, v_j are weights connecting the hidden layer to the output unit f. The terms w_{j0} and v_0 are the biases for the hidden and output units. $h(t)$ and $g(t)$ are continuous sigmoid function, usually of the form $\tanh(t)$ or the logistic function $1/(1+e^{-t})$.

The MLP is the most popular architectures among other neural networks, such as the radial basis function [20], because it can be efficiently trained by error backpropagation [21]. The proper error function in classification is, however, not the mean-squared error (MSE), but the negative log likelihood function [22]

$$-\log\mathcal{L} = -\sum_{i=1}^{N} y_i \log f_i + (1-y_i)\log(1-f_i). \tag{4}$$

Here, it is assumed that the logistic function is used for $h(t)$ and the output label y takes the values of $\{1, 0\}$ instead of $\{+1, -1\}$. Despite having a different error function, the equations in the error backpropagation remain unchanged. The error function in (4) has multiple local minima. This requires repeated training from different random initial conditions and convergence to the global solution is not guaranteed.

C. SVM

The SVM is a recently developed technique for solving a variety of classification and regression problems [23]–[25]. The basic idea of SVM is to find the decision plane that has maximum distance (margin) from the nearest training patterns. The general form of the decision function $u(\mathbf{x})$ for SVM is

$$u(\mathbf{x}) = \sum_{i}^{N} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b \tag{5}$$

where $k(\mathbf{x}_i, \mathbf{x}_j)$ is known as the kernel function; the α_is are chosen by the SVM through training, subjected to constraints $\sum \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq A$. A is a user-defined penalty term regulating the generalization performance of the SVM. Upon training, only a fraction of the α_is will be nonzero. The architecture of the SVM in classification is shown in Fig. 1. SVMs have demonstrated good generalization performance in face recognition [26], text categorization [27], and optical character recognition [28], [29]. It has also been applied to data from gene expression [30], DNA and protein analysis [31], [32].

D. MOGs

As mentioned in Section I, the generative approach is to model the class-conditional density $p(\mathbf{x}|C_\pm)$. Since the input of the glaucoma data contains only continuous valuables (see

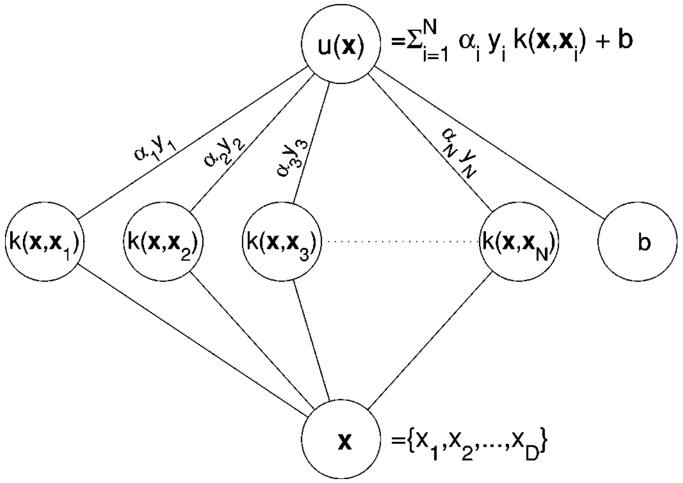

u(x)
$=\Sigma_{i=1}^{N} \alpha_i y_i k(\mathbf{x},\mathbf{x}_i) + b$
$\alpha_1 y_1$
$\alpha_2 y_2$
$\alpha_3 y_3$
$\alpha_N y_N$
$k(\mathbf{x},\mathbf{x}_1)$
$k(\mathbf{x},\mathbf{x}_2)$
$k(\mathbf{x},\mathbf{x}_3)$
$k(\mathbf{x},\mathbf{x}_N)$
b
x
$=\{x_1,x_2,\ldots,x_D\}$

Fig. 1. A visualization of the architecture of SVM in classification. $\mathbf{x}$ is the D-dimensional input vector; $k(\mathbf{x},\mathbf{x}_i)$ is the kernel function between $\mathbf{x}$ and support vectors $\mathbf{x}_i$; $u(\mathbf{x})$ is the output where α_i and y_i are the weight and training labels, respectively, associated with $\mathbf{x}_i$.

Section III) we may want to model each $p(\mathbf{x}|C_{\pm})$ by a normal multivariant density. This would result in a classical LDA or a quadratic discriminant analysis (QDA), depending on whether or not the two normal densities are constrained to have the same covariance. However, in many careful studies of real data the distributions usually do not follow a normal distribution but have slightly heavier tails, skewed or even bi-modal structure. In these problems a single Gaussian is not flexible enough to model adequately the distribution of data.

In simple nonparametric methods such as the histogram method, the input space is divided into many small hypercubes and then $p(\mathbf{x})$ is estimated for each of them. Besides not providing much useful insight in the statistical structure of the data, binning the data space subjects the classifiers to the curse of dimensionality. To model properly the probability distribution of the data, semi-parametric models with "in-between" flexibility are useful. The mixture of Gaussians (MOG) [33], [34] has been popular for its simplicity.

Adopted to our classification problem, the probability densities for the positive and negative classes are each modeled first as a mixture of multivariant normal densities [35], [36],

$$p(\mathbf{x}) = \sum_{m}^{M} p(\mathbf{x}|m)P(m) \tag{6}$$

where for each cluster m,

$$p(\mathbf{x}|m) = \frac{1}{\sqrt{(2\pi)^D|\Sigma_m|}} \times \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_m)^{\top}\Sigma_m^{-1}(\mathbf{x}-\boldsymbol{\mu}_m)\right]. \tag{7}$$

The expectation-maximation (EM) algorithm [37] is used to find the parameters $P(m)$, $\boldsymbol{\mu}_m$ and Σ_m. $P(C_{\pm})$ needed in (1) can also be obtained by ML. Similar to the MLP, multiple trials are required to avoid local minima. However, a learning rate is not required as EM automatically chooses the optimal one. The architecture of the MOG classifier is shown in Fig. 2.

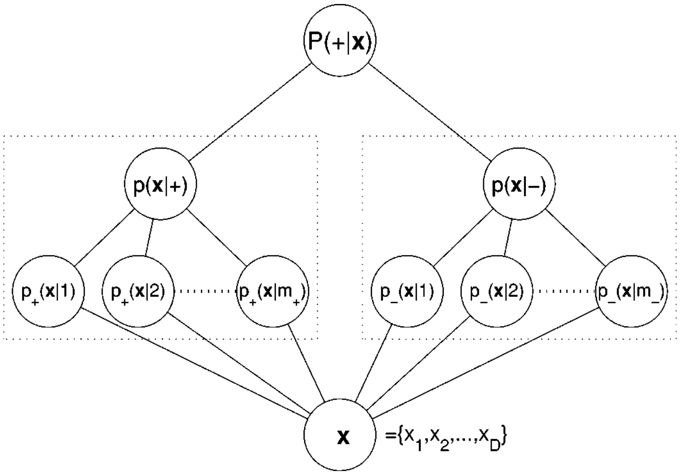


Fig. 2. Architecture of the MOG used in a binary classification setting. $p(\mathbf{x}|\pm)$ are the generative models for the two classes. Each is composed of a MOGs $[p_+(\mathbf{x}|m_+)$ or $p_-(\mathbf{x}|m_-)]$. Output $P(+|\mathbf{x})$ is obtained by applying Bayes' rule (1) on $p(\mathbf{x}|\pm)$.

E. Mixture of Generalized Gaussians (MGGs)

Although the MOGs provides a more flexible model to fit the density of the data, it would be undesirable to fit a density of long tails with two Gaussians. In addition to adequately fitting the data density, a user may also want to understand the structure of the data in terms of number of real clusters and their deviation from normality. With the development of the generalized Gaussian mixture model [38], we are able to model the class-conditional densities $p(\mathbf{x}|C_{\pm})$ with higher flexibility, while preserving the possibility to comprehend the statistical properties of the data in terms of means, variances, and kurtosis, etc. The MGG uses the same mixture model (6) as the MOG. However, each cluster is now described by a linear combination of non-Gaussian random variables $\mathbf{s}_m$

$$p(\mathbf{x}|m) = \int \delta[\mathbf{x} - (\mathbf{A}_m\mathbf{s}_m + \mathbf{b}_m)]p(\mathbf{s}_m)d\mathbf{s}_m \tag{8}$$

i.e., $\mathbf{s}_m$s are the independent hidden sources in cluster m responsible for generating the observation $\mathbf{x}$s given $\mathbf{A}_m$ and $\mathbf{b}_m$. $\mathbf{s}_m$ will assume a generalized Gaussian density [39] of zero mean, unit variance, and shape parameter $\boldsymbol{\beta}_m$

$$p(\mathbf{s}_m|\boldsymbol{\beta}_m) = \prod_{d}^{D} p(s_{md}|\beta_{md}) \tag{9}$$

$$p(s_{md}|\beta_{md}) = \omega(\beta_{md})\exp\left[-c(\beta_{md})|s_{md}|^{2/(1+\beta_{md})}\right] \tag{10}$$

where $\omega(\beta)$ is the normalization constant [39]. β is a measure of kurtosis of the source

$$\text{kurtosis} = \frac{\Gamma\left[\frac{5}{2}(1+\beta)\right]\Gamma\left[\frac{1}{2}(1+\beta)\right]}{\Gamma\left[\frac{3}{2}(1+\beta)\right]^2} - 3 \tag{11}$$

and will be adapted together with $\mathbf{A}_m$, $\mathbf{b}_m$ and $P(m)$ during training. This is done by gradient ascent on the data likelihood [38], [40].

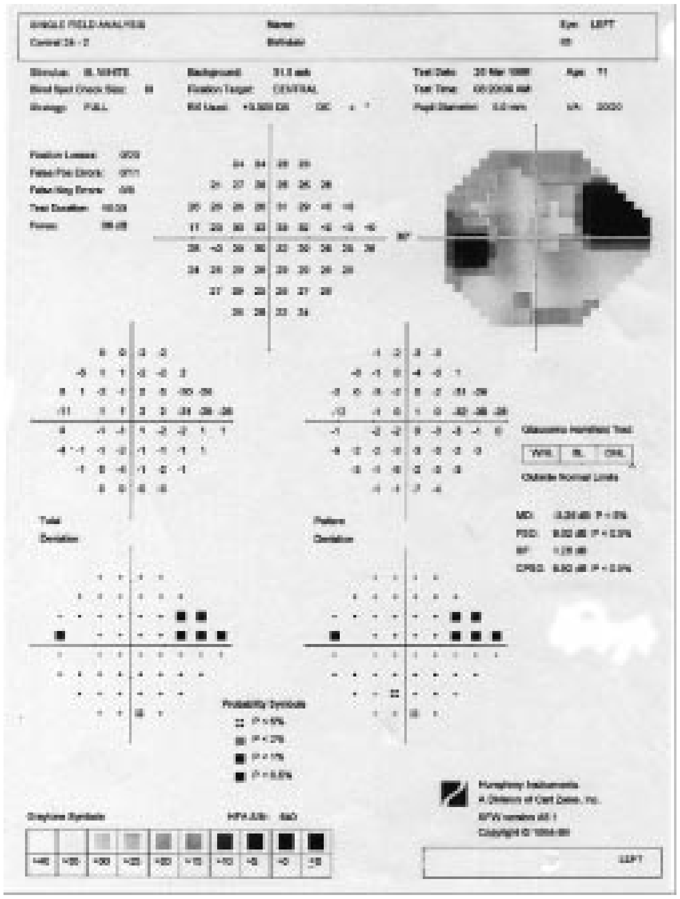


Fig. 3. A sample STATPAC printout from the HFA. Top row: absolute sensitivities and gray scale plot over the 54 locations on the retina. Middle: age corrected total deviation and pattern deviation (total deviation compensated by global depression). Bottom: probability plots of total deviation and pattern deviation.

F. Parzen Windows

The Parzen window is a kernel-based nonparametric approach to density estimation [41], [42]

$$p(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h^D}H\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) \qquad (12)$$

where $H(\mathbf{u})$ is known as the *Parzen window* and has to satisfy $H(\mathbf{u}) \geq 0$ and $\int H(\mathbf{u})d\mathbf{u} = 1$. If we use the isotropic Gaussian Parzen window $H(\mathbf{u}) \propto \exp(-|\mathbf{u}|^2/2)$, it becomes a special instance of the MOG density estimation (6) and (7). Goodness of fit to data density and performance of the resulting Bayes classifier (1) largely depend on the choice of the width parameter h. Drawbacks of the Parzen windows method are that it provides very little information on the structure of the data and requires storage of the entire training set for classification.

III. SAP

A. Humphrey Visual-Field Analyzer

In SAP, a target 0.47° in diameter of variable intensity is flashed for 200 ms against a background of 31.5 apostilbs [10 candelas/meter squared (cd/m^2)]. The most commonly used procedure worldwide is the full threshold SAP test, program 24-2 or 30-2 of the Humphrey Visual Field Analyzer (HFA, Humphrey-Zeiss, Dublin, CA). With the 24-2 program of HFA, the target is randomly presented to 54 locations over 24° at 2-dB resolution. The displayed outputs (Fig. 3) are the absolute sensitivity, the gray scale, the age-corrected total deviation (numerical and probability), the pattern deviation (numerical and probability), the glaucoma Hemifield test (GHT) result, and several global indexes (see below). The age-corrected total deviation is the absolute sensitivity subtracted from an age-matched normal surface. The pattern deviation is the total deviation compensated by global depression to account for cataracts or other nonglaucoma conditions that may globally depress the visual field. The initial output of the Humphrey field analyzer (HFA) is the absolute sensitivity at each of the 54 visual-field locations. This output is represented in decibels relative to the maximum intensity of the machine (set at 0 dB) with a minimum of 40 dB. The values for 52 locations (two locations corresponding to the blind spot are excluded) and the age of the patient will constitute the raw input of our classifiers.

B. STATPAC

The HFA comes with a statistical analysis package (STATPAC) that provides both the raw data and several specialized statistical analyses related to diagnosing glaucoma. The purpose of these analyses is to aid the clinician in interpretation of the visual field. The global indexes included in STATPAC are mean deviation (MD), pattern standard deviation (PSD), short-term fluctuation (SF), corrected pattern standard deviation (CPSD) and the GHT. MD is the depression of the patient's overall field (all test locations averaged) as compared with the age-corrected normative database within the HFA. PSD is a measurement of the degree to which the shape of the field departs from the age-corrected reference fields. Glaucoma typically begins as a localized loss of visual sensitivities. SF is an index of the consistency of the patient's answers during the field test and is obtained by testing twice at ten predetermined points. CPSD is the PSD corrected for SF in attempt to remove the effects of patient variability during the test and to reveal only irregularities caused by actual field loss.

The GHT divides the superior hemifield into five zones and compares locations within each zone to those within a mirror image zone in the inferior hemifield. The five pairs of mirroring sectors are compared and a difference score for each is determined. Glaucoma rarely affects both hemifields in the early stages of the disease. So, the GHT has a high sensitivity for early glaucoma relative to other clinically used measures. If the difference score is outside the 99.5% limits in any one pair compared with the difference score found in age-corrected normal eyes, the field is flagged "outside normal limits (ONL)." If it is outside the 97% limit the field is flagged "borderline (BL)". Fewer differences are considered "within normal limits (WNL)." These analyses are universally used to help the clinician in interpretation of the visual field.

We will evaluate the efficiency of GHT, and the indexes PSD and CPSD in glaucoma diagnosis. Their results will be used as the baseline against which our classifiers' performance are measured.

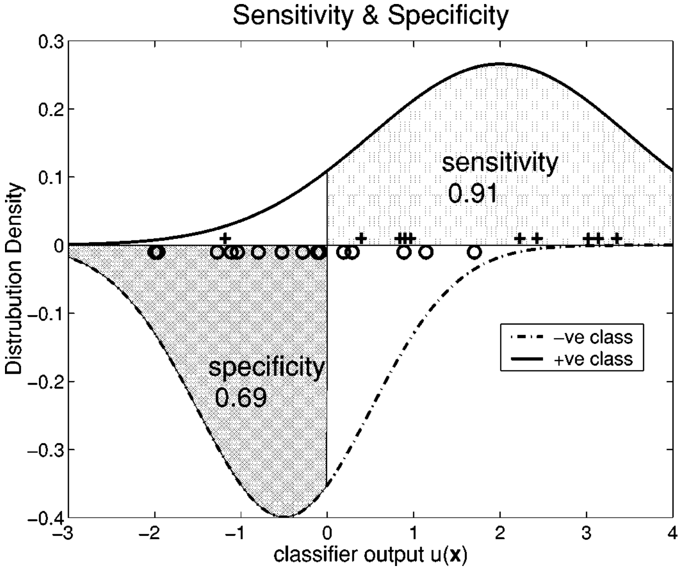


Fig. 4. Sample distribution of classifier output $u(\mathbf{x})$ on a two classes dataset.

IV. Experimental Settings

A. Data

Currently there is no gold standard to determine whether or not a patient has glaucoma. Instead, we use glaucomatous optic neuropathy (GON) as our teaching label [43]. Patients and normal controls were labeled based on masked analysis of simultaneous stereophotographs of the optic disk and ocular history without reference to visual fields (the SAP data). Our glaucoma dataset contains a collection of 156 eyes with GON and 189 eyes without GON. The STATPAC indexes MD, SF, PSD, and CPSD are scalar values and can be directly used as classifiers output $u(\mathbf{x})$ (Fig. 4) to create the ROC curves (see Section V-A). Since the amount of data available is limited, we used a 25-fold cross-validation scheme to evaluate the classifiers. The dataset was divided uniformly into 25 subsets. Each subset was in turn held aside as the test set when the other 24 were used to train the classifiers. The results on the 25 subsets were combined into one single ROC plot for each classification method. To facilitate training, we first normalized each of the 52 locations raw sensitivity threshold values and age to have zero mean and unit variance.

B. Machine Classifiers

The user-chosen parameters in all machine classifiers were set by optimizing their cross-validation performance in five out of the 25 partitions. The MLP was setup and trained using the MATLAB Neural Network Toolbox 4.0 (The MathWorks, Natick, MA). The network contained a hidden layer of 10 tanh units and a logistic output unit. The network was trained using the Levenberg–Marquardt method [44]. Early stopping was used to prevent overfitting. This was done by reserving one of the 24 subsets constituting the training set as the “stopping set.” In each fold of the cross validation, 20 networks were trained and their output were averaged to return a single $P(C_+|\mathbf{x})$ value for each testing data point $\mathbf{x}$. In the SVM, we tried both the linear $(A = 0.5)$ and Gaussian ($\sigma = 3.6$ and $A = 1.5$) kernels. We implemented the sequential minimal optimization [45], [46] in MATLAB code to train the SVM.

For the generative classifiers, $h = 3$ was used for the Parzen window classifier. For the MOG and MGG classifiers, due to the limited availability of data, we first performed principle component analysis (PCA) on the normalized data to reduce the dimension. The data were projected onto the subspace of the first eight components, which accounted for more than 80% of total variance. Class-conditional densities $p(\mathbf{x}|C_\pm)$ were modeled separately on the two classes. In the MOG, there were two clusters for the glaucoma class and one cluster for the normal class. In the MGG, only 1 cluster was found for each class.

For a fair comparison, all other classifiers were also trained in the eight dimensional subspace, to single out the gain or loss from dimension reduction. Some user-chosen parameters were re-estimated in this reduced-dimensional space: five hidden units for MLP; $\sigma = 3.4$ for the Gaussian SVM; and $h = 2$ for the Parzen window classifier.

V. Results

A. ROC Curve

A simple way to assess the performance of classifiers is to compare their average misclassification rate. In biomedical data, the dataset usually comes with a higher proportion of normal ($-$ve class). Classifiers will, hence, tend to achieve an overall low misclassification rate by sacrificing the $+$ve class data, resulting in a higher misclassification rate on the $+$ve class than the $-$ve class. However, the misclassification cost associated with the $+$ve class is usually higher than that of $-$ve class. For a classifier which outputs a scalar value for a given data point $\mathbf{x}$, showing its likelihood of belonging to the $+$ve class [such as $P(C_+|\mathbf{x})$ or $u(\mathbf{x})$ in (5)], we would like to pick a decision threshold other than 0.5 for $P(C_+|\mathbf{x})$ (or 0 for $u(\mathbf{x})$), in favor of the $+$ve class. In Fig. 4, the data points of two classes are placed next to the x axis according to the value of$u(\mathbf{x})$ given by a classifier. The fractional densities (smoothed histograms) show the distribution of $u(\mathbf{x})$ for the two classes. Different classifiers would have different distributions for $u(\mathbf{x})$. The true-positive rate, also known as sensitivity, is the fraction (or %) of positively labeled test data classified as $+$ve. This is also the area under the positive class density curve, to the right of the decision threshold (zero in Fig. 4). Specificity is the true-negative rate, the fraction of negatively labeled examples classified as $-$ve.

Varying the threshold level θ such that a given example $\mathbf{x}$ will be classified as positive if $u(\mathbf{x}) > \theta$ leads to a tradeoff between sensitivity and specificity. The receiver operating characteristic (ROC) [47] curve is a plot of sensitivity versus 1–specificity (or true-positive rate versus false-positive rate). The area under the ROC curve summarizes the quality of classification over a wide range of misclassification costs [48]. We estimate the variance of the ROC area by a nonparametric method as described in [49]. There is an interesting interpretation of the ROC area: it is equal to the probability of a random sample from positive class $\mathbf{x}_+$ being assigned by the classifier a $u(\mathbf{x}_+)$ [or $P(C_+|\mathbf{x}_+)$] value greater than that of a random sample from negative class. i.e., ROC area $=$ Prob $(u(\mathbf{x}_+) > u(\mathbf{x}_-))$, for any random sample pair $(\mathbf{x}_+, \mathbf{x}_-)$ drawn from the two classes.

TABLE I
A COMPARISON OF PERFORMANCE OF CLASSIFIERS BY THEIR ROC AREAS AND SENSITIVITIES AT SELECTED SPECIFICITIES

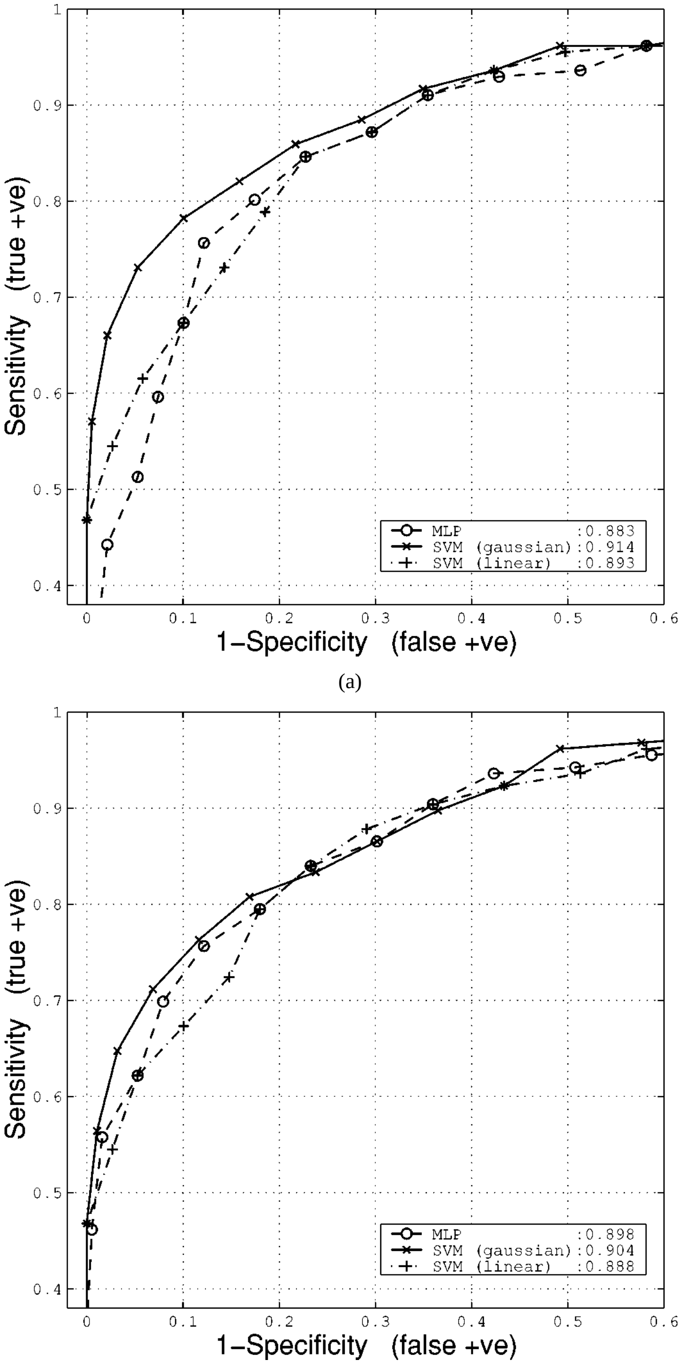
	ROC area	Sensitivities at	
	(std. err.)	specificities of	
		0.90	0.75
STATPAC:			
GHT†		0.667	
MD	0.838 (0.022)	0.654	0.731
SF	0.694 (0.029)	0.365	0.532
PSD	0.883 (0.020)	0.756	0.846
CPSD	0.844 (0.025)	0.737	0.782
Discriminative:			
(full dim.)			
MLP	0.883 (0.019)	0.660	0.859
gaussian SVM	0.914 (0.016)	0.776	0.878
linear SVM	0.893 (0.017)	0.660	0.853
(PCA reduced dim.)			
MLP	0.898 (0.017)	0.713	0.846
gaussian SVM	0.904 (0.016)	0.744	0.833
linear SVM	0.888 (0.018)	0.667	0.853
Generative:			
(full dim.)			
LDA	0.824 (0.023)	0.583	0.756
QDA	0.916 (0.016)	0.788	0.865
Parzen Window	0.892 (0.017)	0.673	0.840
(PCA reduced dim.)			
LDA	0.880 (0.018)	0.647	0.833
QDA	0.921 (0.015)	0.782	0.872
Parzen Window	0.903 (0.016)	0.724	0.808
MOG	0.923 (0.014)	0.769	0.846
MGG	0.902 (0.016)	0.750	0.821

† The specificity of GHT cannot be varied. Reported sensitivity corresponds to the specificity of 1.0 on the test data.

B. Classification Results

The ROC areas for the classifiers are summarized in Table I, grouped into the following categories: STATPAC, Discriminative (in full- and PCA-reduced dimension) and Generative (full- and PCA-reduced dimension). We also listed the sensitivities of the classifiers at specificities of 0.90 and 0.75. The ROC curves for the four categories of classifiers are plotted in Figs. 5 and 6. The index PSD is very competitive to our classifiers. The CPSD, which is derived from PSD by correcting for SF in visual-field sensitivity, does not exhibit improvement over PSD. The GHT identified 66.7% of the glaucoma eyes as "outside normal limits" and all of the normals as "within normal limits" or "borderline," Probably this is because GHT was designed to have a specificity value of 99.5%.

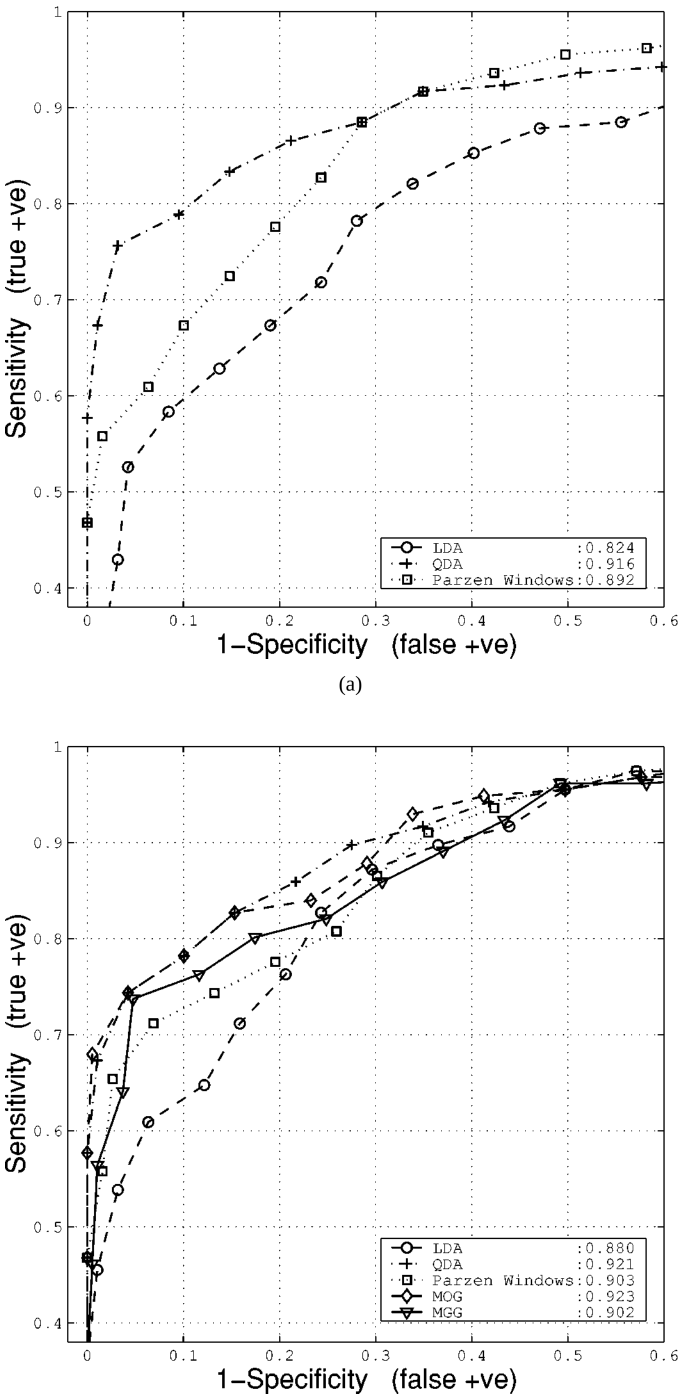
To compare the classifiers quantitatively, we compute the p-values for a two-tails test among the ROC areas of the classifiers using the nonparametric method outlined in [49]. Since the ROC curves are generated from classifiers applying on the same data set, correlation between the ROC areas must be taken into account. For instance, the difference in ROC areas between the SVMs with Gaussian and linear kernels on the full-dimensional data is 0.021. This is small compared with their std. err. of 0.016. However, a p-value of 7% is obtained on a two-tails test when correlation between ROC areas are take into account. In Tables II–IV, we report the p-values between the ROC areas of the classifiers. For the full-dimensional data,



(a)

(b)

Fig. 5. ROC curves for the discriminative classifiers on (a) full-dimension and (b) PCA-reduced dimension data. The area under the curves for each classifier is given in the insert.



(a)

(b)

Fig. 6. ROC curves for the generative classifiers on (a) full-dimension and (b) PCA-reduced dimension data.

the SVM with Gaussian kernel is significantly better than PSD at 5% level, while the QDA is significantly better than PSD at 1% level. In the PCA-reduced dimension, the QDA and MOG are significantly better than PSD at 1% level. From Tables III and IV, only the LDA showed substantial improvement when working on data in PCA-reduced dimension. Both linear and Gaussian SVMs performed worse in the PCA-reduced dimension, though the differences were not statistically significant. The ROC curves of the best classifier from each category are plotted together in Fig. 7 for comparison.

VI. Feature Selection

It is always useful and interesting to identify the subset of input variables that contribute most in the classification. Eliminating irrelevant input variables that introduce noise often improves classification. Since exhaustive search over all possible combinations of input variables to identify the best subset is prohibitively expensive, here we use forward selection and backward elimination [12], to rank the variables and identify the subset that would give best classification. Forward selection is sequentially adding variables one at a time, choosing the next variable that most increases or least decreases classification. Backward elimination starts with all input variables and sequentially deletes the next variable that most decreases or least increases classification. These two greedy feature selection methods may not find the optimal feature set, but, nonetheless, their time complexity is only quadratic in the number of features as compared with the exponential growth for exhaustive search.

From Table I, QDA is the best classifier that can work directly on the full-dimensional input. In addition, its fast training and global convergence allows repeated training in a short period of time. We performed forward selection and backward elimination using the QDA to rank the input variables. The ROC area from the 25-fold cross-validation was used as the criteria for selecting the next variable to add/delete. In Fig. 8, the ROC areas are plotted as a function of number of variables included from the list ranked by forward selection and backward elimination. Performance of the full 53 dimensions input was achieved by using less than ten most important input variables. Also, both forward and backward selection methods peaked around 20 variables. This can reasonably be interpreted as the intrinsic dimensionality of the data, since visual fields next to each other should have correlated sensitivities. Besides giving better classification accuracy, using only 20 input variables reduces the time spent by the patients in SAP from 15 to 6 min/eye and, hence, more screening can be done. Moreover, real-time classification is made possible when the visual locations are tested in the order ranked by feature selection.

In Fig. 9 we plot the ranks given by forward selection and backward elimination to the 53 input variables on a two dimensional space. Visual-field locations are labeled 1–54 as displayed in the lower right insert (cf. Fig. 3). Locations 18 and 31 corresponding to the blind spot are omitted. Variables near the origin (e.g., location 5 and 47, etc.) are considered by both forward selection and backward elimination most important toward glaucoma diagnosis. The two rankings from forward and backward selection agreed with each other well, as the variables lie approximately along the diagonal.

We could have used the Gaussian SVM in our forward and backward feature selection, since it performed as well as the QDA and does not required multiple runs. This would also verify how the set of optimal variables changes with different classifiers. However, the parameters A and σ depend on the number of variables used and have to be selected carefully in order to give reliable ranking result. This is a major concern for the wrapper-type methods [50], [51]. There are feature selection methods tailored for individual classifiers. We are

TABLE II
p-VALUES FOR THE TWO–TAILS TEST BETWEEN ROC AREAS OF THE STATPAC PSD INDEX AND THE MACHINE-LEARNING CLASSIFIERS

		STATPAC	full dim.			PCA reduced dim.		
		PSD	MLP	SVM-g	SVM-l	MLP	SVM-g	SVM-l
ROC area		0.883	0.883	0.914	0.893	0.898	0.904	0.888
STATPAC								
PSD	0.883		0.956	0.023	0.606	0.336	0.180	0.801
full dim.								
LDA	0.824	0.018	0.001	<.0005	<.0005	<.0005	<.0005	<.0005
QDA	0.916	0.002	0.041	0.922	0.130	0.183	0.379	0.071
Parzen	0.893	0.633	0.433	0.067	0.919	0.545	0.282	0.622
PCA reduced dim.								
LDA	0.880	0.877	0.880	0.016	0.065	0.104	0.053	0.222
QDA	0.921	0.008	0.008	0.611	0.062	0.055	0.108	0.041
Parzen	0.903	0.221	0.100	0.319	0.192	0.547	0.943	0.095
MOG	0.923	0.005	0.004	0.485	0.029	0.023	0.054	0.019
MGG	0.902	0.208	0.173	0.365	0.512	0.720	0.901	0.359

TABLE III
p-VALUES FOR TWO-TAILS TEST BETWEEN ROC AREAS OF THE DISCRIMINATIVE CLASSIFIERS

	full dim.			PCA reduced dim.		
	MLP	SVM-g	SVM-l	MLP	SVM-g	SVM-l
ROC area	0.883	0.914	0.893	0.898	0.904	0.888
full dim.						
MLP		0.002	0.361	0.196	0.052	0.662
SVM-g			0.069	0.140	0.234	0.041
SVM-l				0.555	0.287	0.275
PCA reduced dim.						
MLP					0.474	0.277
SVM-g						0.139

TABLE IV
p-VALUE FOR THE TWO-TAILS TEST BETWEEN ROC AREAS OF THE GENERATIVE CLASSIFIERS

	full dim.		PCA reduced dim.				
	QDA	Parzen	LDA	QDA	Parzen	MOG	MGG
ROC area	0.916	0.892	0.880	0.921	0.903	0.923	0.902
full dim.							
LDA	<.0005	<.0005	0.001	<.0005	<.0005	<.0005	<.0005
QDA		0.132	0.036	0.718	0.375	0.599	0.363
Parzen			0.193	0.489	0.066	0.020	0.472
PCA reduced dim.							
LDA				0.015	0.024	0.006	0.170
QDA					0.164	0.605	0.004
Parzen						0.078	0.955
MOG							0.001

currently studying several methods for our SAP data, including the LEAPS [52] for LDA, saliency metric [53] for MLP and some others for SVM [54], [55].

VII. DISCUSSION

A. Compared with STATPAC

From Tables I and II and Fig. 7, our best machine classifiers such as Gaussian SVM, QDA, and MOG perform better than the STATPAC indexes PSD and CPSD in terms of ROC areas as well as sensitivities at selected specificities. In fact, our machine classifiers are at a relative disadvantage since the STATPAC indexes are derived using age corrected reference visual fields from a much larger normative database within the HFA. Our classifiers were trained on the raw sensitivity threshold values (and age) from a relatively limited dataset. Currently, ophthalmologists at clinics rely on the STATPAC indexes when diagnosing glaucoma from SAP. The improvement of the machine classifiers over STATPAC indexes shows promise for assisting ophthalmologists in interpreting the HFA output (Fig. 3).

Glaucoma experts have extensive experience with SAP and interpreting the results from the HFA output using STATPAC. However, there are newer perimetric tests such as short-wavelength automated perimetry [43], frequency-doubling technology perimetry [56], [57], and structural tests such as the Heidelberg Retina Tomograph and GDx Nerve Fiber Analyzer [58]. The machine classifiers explored here have great potential for extending their use on these data to assist the glaucoma experts in interpreting these less familiar but improved tests.

B. Among the Machine Classifiers

When considering the full-dimensional data, the Gaussian SVM shows significantly better performance over the MLP in terms of ROC area and sensitivities at chosen specificities. However, it did not benefit from the dimension reduction of the data as much as the MLP, both in terms of training time (Table V) and classification accuracy. In fact, both the Gaussian and linear SVM degraded in the reduced-dimension space. On the other hand, the SVM does not require repeated training with random

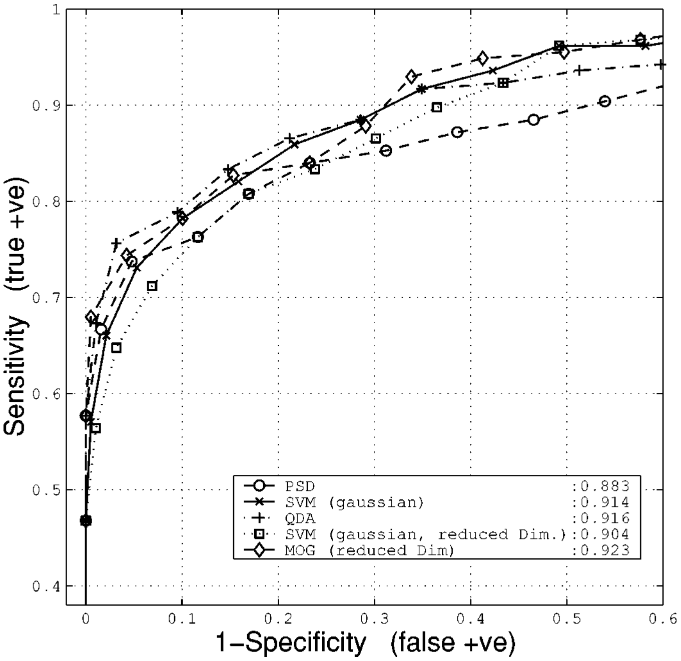


Fig. 7. ROC curves of best classifier from each category in Table I.

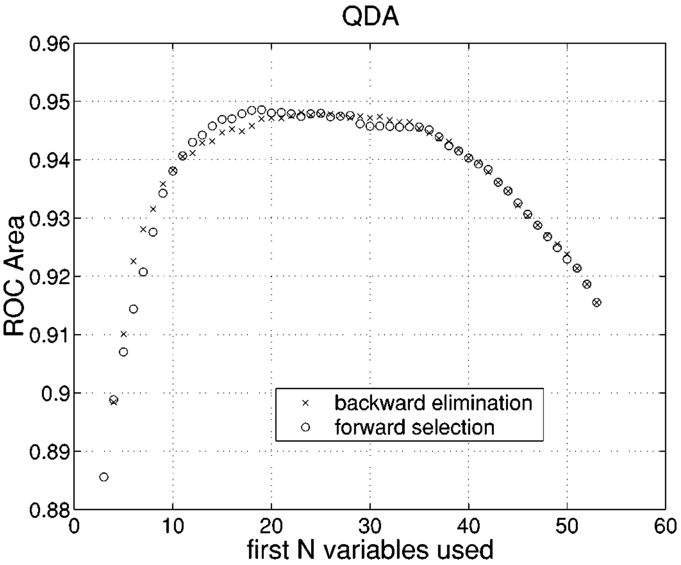


Fig. 8. ROC area as a function of number of variables used in QDA. Variables are added according to the rank given by forward selection or backward elimination.

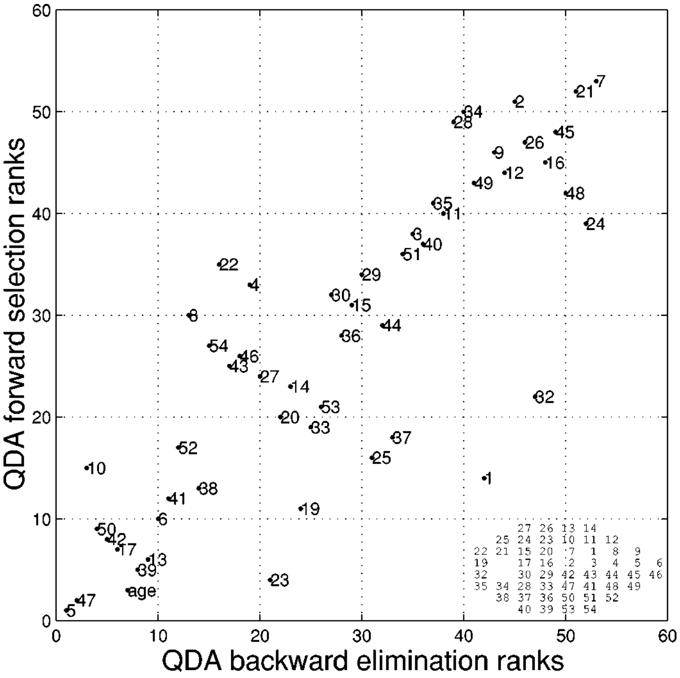


Fig. 9. Correlation of rankings given by forward selection and backward elimination on the variables (52 visual-field locations + Age). Variables close to origin (e.g., location 5 and 47, etc.) are more informative in glaucoma diagnosis. Lower right insert displays the relative positions on the retina of the variables.

TABLE V
TYPICAL TRAINING TIME (IN SECONDS) FOR THE VARIOUS CLASSIFIERS. ALL CLASSIFIERS ARE IMPLEMENTED IN MATLAB RUNNING ON A PC WITH THE PENTIUM III 500-MHZ CPU

	full dim.	PCA reduced dim.
MLP[†]	23	1.8
gaussian SVM	7.0	7.1
linear SVM	5.1	6.0
LDA	0.050	0.038
QDA	0.095	0.040
Parzen Window	1.1	0.54
MOG[†]		2.0
MGG[†]		32

[†] The MLP, MOG, and MGG classfiers require multiple runs from different random initial conditions. Reported times are from a single run only.

initial conditions. Since the SVM by itself is a linearly constrained quadratic programming problem, it will not get stuck at a local minimum during training. The performance of the SVM is insensitive to the exact choice of the parameters such as A and σ. Their values found in cross validation can be used to train the whole dataset for prediction of future unseen examples.

The LDA was the fastest to train, but it performed the poorest in our experiments. The LDA benefited most in case of the dimension reduced data set (as indicated by the p-values,Table IV) yet still did not give improved performance compared with other classifiers. In contrast, the simple Parzen window classifier achieved an ROC area similar to those of the MLP and SVMs.

QDA is an extension over the LDA as it models the normal and glaucoma class with independent Gaussian densities. This approximation fit our data distribution well and the resulted classifier was among the best in our experiments. Surprisingly, the MOG added only one Gaussian to the glaucoma class and barely improved the results over the QDA. Although the MGG is an improvement of the MOG in modeling densities of continuous variables, the MGG + Bayes' rule did not improve classification over the MOG. One explanation is that our data were close to Gaussian and already modeled well by the mixture model, and the extra flexibility in modeling non-Gaussian densities introduced additional free parameters. This resulted in overfitting the training data and poor generalization to unseen testing data. In summary, QDA is to be preferred in our SAP data for its fast training, simple implementation, global convergence, classification performance, and ease of interpretation.

C. The Linear Discriminant Function (LDF)

In this paper, the classical LDA is categorized under generative classifiers despite the word "discriminant" in its name. An LDF can be either generative or discriminative, depending on whether or not its weights are obtained from first modeling $p(\mathbf{x}|C_{\pm})$. Logistic regression, a single layer MLP or a SVM with linear dot product kernel are all effectively linear classifiers. LDFs derived from them should be considered as discriminative since they work directly on the decision boundary during training. As seen from Table I, the linear SVM performed better than the classical LDA both in the full and reduced-dimension space. Moreover, the LDA showed markable improvement when working in the reduced-dimension space. This suggests that linear SVM is generally more robust against irrelevant input variables. This is likely because the number of parameters for the classical LDA grows as $\mathcal{O}(D^2)$ (D is the dimension of the input space). As a result, our LDA is intrinsically more complex in structure. [59] theoretically and empirically compare the discriminative and generative LDFs. They concluded that generative learning has higher asymptotic error but approaches this limit faster as a function of number of training examples.

D. Generative Versus Discriminative

Although the posterior probabilities $P(C_{\pm}|\mathbf{x})$ are insensitive to minor variations in class-conditional densities $P(\mathbf{x}|C_{\pm})$, since $P(\mathbf{x}|C_{\pm})$ is usually vulnerable to noise and outliers of the data [60], classifiers based on the generative model often show a lower degree of robustness. It has been proposed that robust estimators [61] be used instead of ML ones in the data generative model to guard against outliers. Discriminative classifier on the other hand are more sensitive to "outliers" near the decision boundary. In Gaussian SVM, the value of σ is adjusted to prevent overfitting. In MLP, this is done by weight decay and early stopping [11].

In addition to a binary classification, it is desirable for a classifier to output a scalar value showing its belief in classification. The generative classifiers give their output as $P(C_{+}|\mathbf{x})$, but this is not easy to obtain from discriminative classifiers such as decision trees. Methods have been developed to make SVM output probabilistic [62]. In generative classifiers, the availability of $p(\mathbf{x})$ can be used to detect outliers that belong to neither of the two classes. New examples containing missing entries can also be handled by marginalizing $p(\mathbf{x}|C_{\pm})$ over $\mathbf{x}_{\text{missing}}$.

The performance of QDA and MOG in our glaucoma problem demonstrates the power of Bayes' rule in classification, provided that we can model the underlying statistical structure of the data accurately (compare LDA, QDA, and MOG). Recently, there has been growing interest in combining the generative and discriminative approaches to train classifiers [63]–[65]. These studies aim to exploit the advantages of the two paradigms. Encouraging results were obtained in hand-written digit recognition [8]. It would be interesting to see, for example, how well a discriminative QDA or MOG perform when applied to the SAP data.

VIII. Conclusion

We have compared a variety of machine classifiers with the STATPAC indexes traditionally used for glaucoma diagnosis on SAP data. In general, the machine classifiers give statistically significant improvement over the STATPAC indexes as measured by the area under the ROC curve. They show promise for use in a clinical setting together with the STATPAC indexes for glaucoma diagnosis. Forward selection and backward elimination were used to rank the visual-field locations. Important locations for glaucoma diagnosis were identified. Properties, advantages and disadvantages of generative and discriminative machine classifiers as applied to our SAP data have been compared. The success of the machine classifiers in the SAP data suggest they may be even more promising for their applications in progression prediction and for diagnosis using other less familiar, but improved visual function or structural tests for glaucoma, such as short-wavelength automated perimetry [66] and optic nerve head topography [67].

Acknowledgment

The authors would like to thank the reviewers for their valuable comments and suggestions that have improved the quality of the paper.

References

[1] R. A. Hitchings and G. L. Spaeth, "The optic disc in glaucoma, ii: Correlation of appearance of the optic disc with the visual field," *Br. J. Ophthalmol.*, vol. 61, pp. 107–113, 1977.

[2] C. A. Johnson, "Perimetry and visual field testing," in *The Ocular Examination: Measurements and Findings*, K. Zadnik, Ed. Philadelphia, PA: Saunders, 1997.

[3] B. C. Chauhan, S. M. Drance, and G. R. Douglas, "The use of visual field indices in detecting changes in the visual field in glaucoma," *Investigat. Ophthalmol. Visual Sci.*, vol. 31, pp. 512–520, 1990.

[4] R. T. Cox, *The Algebra of Probable Inference*. Baltimore, MD: Johns Hopkins Univ. Press, 1961.

[5] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, pp. 131–163, 1997.

[6] T. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[7] D. R. Cox and E. J. Snell, *Analysis of Binary Data*, 2nd ed. London, U.K.: Chapman and Hall, 1989.

[8] L. K. Saul and D. D. Lee, "Discriminative mixture modeling," in *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press, 2002.

[9] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. New York: Chapman and Hall, 1993.

[10] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.

[11] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Clarendon, 1995.

[12] B. D. Ripley, *Pattern Recognition and Neural Netwroks*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[13] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. New York: Springer-Verlag, 2000.

[14] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[15] S. H. Haykin, *Neural networks: A Comprehensive Foundation*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall, 1999.

[16] F. Rosenblatt, "The perceptron: A probabilisitic model for information storage and organization in the brain," *Psychological Rev.*, vol. 65, no. 6, pp. 386–408, 1958.

[17] E. K. Blum and L. K. Li, "Approximation theory and feedforward networks," *Neural Networks*, vol. 4, no. 4, pp. 511–515, 1991.

[18] A. Samal and P. A. Iyengar, "Automatic recognition and analysis of human faces and facial expression: Asurvey," *Pattern Recogn.*, vol. 25, pp. 65–77, 1992.

[19] H. S. Baird, "Recognition technology frontiers," *Patern Recogn. Lett.*, vol. 14, pp. 327–334, 1993.

[20] D. S. Broomhead and D. Lowe, "Multivariable functional interpolation and adptive networks," *Complex Syst.*, vol. 2, pp. 321–355, 1988.

[21] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by back-propagating errors," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, D. E. Rumelhart and J. L. McClelland, Eds. Cambridge, MA: MIT Press, 1986, vol. 1, pp. 318–362.

[22] E. B. Baum and F. Wilczek, "Supervised learning of probability distributions by neural networks," in *Advances in Neural Information Processing Systems 0,* D. Z. Anderson, Ed. New York: Amer. Inst. Phys., 1988, pp. 52–61.

[23] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Ann. ACM Workshop Computational Learning Theory*, D. Haussler, Ed., 1992, pp. 144–152.

[24] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[25] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[26] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: An application to face detection," in *Proc. Computer Vision and Pattern Recognition '97*, Puerto Rico, pp. 130–136.

[27] S. T. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in *Proc. ACM-Conf. Information and Knowledge Management (CIKM98)*, Nov 1998, pp. 148–155.

[28] B. Schölkopf, C. Burges, and V. Vapnik, "Incorporating invariances in support vector learning machines," in *Lecture Notes in Computer Science*, C. von der Malsburg, W. von Seelen, J. C. Vorbrüggen, and B. Sendhoff, Eds. Berlin, Germany: Springer-Verlag, 1996, vol. 1112, Artificial Neural Networks—ICANN'96, pp. 47–52.

[29] C. J. C. Burges and B. Schölkopf, "Improving the accuracy and speed of support vector learning machine," in *Advances in Neural Information Processing Systems 9*. Cambridge, MA: MIT Press, 1997, pp. 375–381.

[30] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, pp. 906–914, 2000.

[31] C. H. Q. Ding and I. Dubchak, "Multi-class protein fold recognition using support vector machines and neural networks," *Bioinformatics*, vol. 17, pp. 349–358, 2001.

[32] A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K.-R. Müller, "Engineering support vector machine kernels that recognize translation initiation sites," *Bioinformatics*, vol. 16, pp. 799–807, 2000.

[33] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley, 1985.

[34] G. J. McLachlan and K. E. Basford, *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker, 1988.

[35] W.-S. Chou and Y.-C. Chen, "A new fast algorithm for the effective training of neural classifiers," *Pattern Recogn.*, vol. 25, pp. 423–429, 1992.

[36] R. L. Streit and T. E. Luginbuhl, "Maximum likelihood training of probabilistic neural networks," *IEEE Trans. Neural Networks*, vol. 5, pp. 764–783, Sept. 1994.

[37] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. Roy. Statist. Soc., Ser. B*, vol. 39, pp. 1–38, 1977.

[38] T.-W. Lee and M. S. Lewicki, "The generalized Gaussian mixture model using ICA," in *Proc. Int. Workshop Independent Component Analysis (ICA'00)*, Helsinki, pp. 239–244.

[39] G. Box and G. Tiao, *Bayesian Inference in Statistical Analysis*. New York: Wiley, 1973.

[40] S. Choi, A. Cichocki, and S. Amari, "Flexible independent component analysis," in *Neural Networks for Signal Processing VIII*, T. Constantinides, S. Y. Kung, M. Niranjan, and E. Wilson, Eds. Piscataway, NJ: IEEE Press, 1998, pp. 83–92.

[41] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, no. 3, pp. 1065–1076, 1962.

[42] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Ann. Math. Statist.*, vol. 27, pp. 832–837, 1956.

[43] P. A. Sample, C. F. Bosworth, E. Z. Blumenthal, C. Girkin, and R. N. Weinreb, "Visual function specific perimetry for indirect comparison of different ganglion cell populations in glaucoma," *Investigat. Ophthalmol. Visual Sci.*, vol. 41, pp. 1783–90, 2000.

[44] K. Levenberg, "A method for the solution of certain nonlinear problems in least square," *Quart. J. Appl. Math.*, vol. II, pp. 164–168, 1944.

[45] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods: Support Vector Machines*, B. Schölkopf, C. Burgees, and A. Smola, Eds. Cambridge, MA: MIT Press, 1998, pp. 185–208.

[46] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to Platt's SMO algorithm for SVM classifier design," Control Dvision, Dept. Mech Production Eng, Nat. Univ. Singapore, Singapore, Tech Rep. [CD-99-14], 1999.

[47] D. J. Goodenough, K. Rossmann, and L. B. Lusted, "Radiographic applications of receiver operating characteristic ROC curve," *Radiology*, vol. 110, pp. 89–95, 1974.

[48] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a reciever operating characteristic ROC curve," *Radiology*, vol. 143, pp. 29–36, 1982.

[49] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operati ng characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, pp. 837–845, 1988.

[50] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Proc. Int. Conf. Machine Learning*, 1994, pp. 121–129.

[51] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artif. Intell.*, vol. 97, no. 1–2, pp. 245–271, 1997.

[52] G. Furnival and R. Wilson, "Regression by leaps and bounds," *Technometrics*, vol. 16, pp. 499–511, 1974.

[53] D. W. Ruck, S. K. Rogers, and M. Kabrisky, "Feature selection using a multilayer perceptron," *J. Neural Network Comput.*, vol. 2, no. 2, pp. 40–48, 1990.

[54] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Proc. 15th Int. Conf. Machine Learning*, San Francisco, CA, 1998, pp. 82–90.

[55] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs," in *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. Cambridge, MA: MIT Press, 2001, pp. 668–674.

[56] C. Johnson and S. J. Samuels, "Screening for glaucomatous visual field loos with frequency-doubling perimetry," *Investigat. Ophthalmol. Visual Sci.*, vol. 38, pp. 413–425, 1997.

[57] A. Turpin, A. M. McKendrick, C. A. Johnson, and A. J. Vingrys, "Development of efficient threshold strategies for frequency doubling technology perimetry using computer simulation," *Investigat. Ophthalmol. Visual Sci.*, vol. 43, pp. 322–331, 2002.

[58] L. M. Zangwill, C. Bowd, C. C. Berry, J. Williams, E. Z. Blumenthal, C. A. Sanchez-Galeana, C. Vasile, and R. N. Weinreb, "Discriminating between normal and glaucomatous eyes using the heidelberg retina tomograph, GDx nerve fiber analyzer, and optical coherence tomograph," *Arch. Ophthalmol.*, vol. 119, pp. 985–993, 2001.

[59] A. N. Ng and M. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press, 2002.

[60] P. Rousseeuw and B. van Zomeren, "Unmasking multivariate outliers and leverage points," *J. Amer. Statist. Assoc.*, vol. 85, 1990.

[61] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley, 1992.

[62] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 1999.

[63] T. Jebara and A. Pentland, "Maximum conditional likelihood via bound maximization and the cem algorithm," in *Advances in Neural Information Processing Systems 11*: MIT Press, 1999, pp. 494–500.

[64] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," in *Advances in Neural Information Processing Systems 11*: MIT Press, 1999, pp. 487–493.

[65] T. Jaakkola, M. Meila, and T. Jebara, "Maximum entropy discrimination," in *Advances in Neural Information Processing Systems 12*. Cambridge, MA: MIT Press, 2000, pp. 470–477.

[66] P. A. Sample and R. N. Weinreb, "Color permetry for assessment of primary open-angle glaucoma," *Investigat. Ophthalmol. Visual Sci.*, vol. 31, pp. 1869–75, 1990.

[67] F. S. Mikelberg, C. M. Parfitt, N. V. Swindale, S. L. Graham, S. M. Drance, and R. Gosine, "Ability of the heidelberg retina tomograph to detect early glaucomatous visula field loss," *J. Glaucoma*, vol. 4, pp. 242–247, 1995.

[68] J. W. Shavlik and T. G. Dietterich, *Readings in Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1990.

Kwokleung Chan recieved the B.Sc degree from the Department of Physics, the Hong Kong University of Science and Technology, Hong Kong, in 1994 and the M.Sc degree from the Department of Physics, University of California, San Diego, in 1996. He is currently working towards the Ph.D. degree in the Computational Neurobiology Laboratory, the Salk Institute, La Jolla, CA.

His research interests are development and application of machine-learning techniques in biological and biomedial data analysis.

Te-Won Lee (S'96–A'99) received the diploma degree in March 1995 and the Ph.D. degree in October 1997 (*summa cum laude*) in electrical engineering from the University of Technology Berlin, Germany. He was a visiting graduate student at the Institute Nationale Polytechnique de Grenoble, Grenoble, France, the University of California, Berkeley, and Carnegie Mellon University, Pittsburgh, PA.

He is an Assistant Research Professor at the Institute for Neural Computation, University of California, San Diego. He is also a Research Associate at the Salk Institute for Biological Studies, La Jolla, CA, and an Adjunct Professor at Korea Advanced Institute of Science and Technology (KAIST), Seoul, Korea. From 1995–1997, he was a Max-Planck Institute Fellow and in 1999 he was a Visiting Professor at KAIST. His research interests are in the areas of unsupervised learning algorithms, artificial neural networks, and Bayesian probability theory with applications in signal and image processing. He has written one book on independent component analysis.

Dr. Lee received the Erwin-Stephan prize for excellent studies from the University of Technology, Berlin, in 1997 and the Carl-Ramhauser prize for excellent dissertations from the Daimler-Chrysler Corporation in 1999. He served as program chair and editor for the ICA-2001 conference.

Pamela A. Sample received the Ph.D. degree in Psychology with an emphasis on sensation and perception in the visual system from the University of California, San Diego, in 1988. She received the American Dissertation Fellowship from the American Association of University Women. Her thesis was the development of Short-wavelength Automated Perimetry, a test currently in use to diagnose glaucoma.

She is Professor of Ophthalmology and Director of the Visual Function Laboratory at the Shiley Eye Center, University of California, San Diego. She is the Principal Investigator of a large longitudinal study of visual function in glaucoma and ocular hypertension, currently in its 13th year of funding through the National Eye Institute. During this time she has authored or coauthored 75 peer-reviewed publications and 90 peer-reviewed abstracts. Her research interests include understanding the effects of glaucoma on the various subtypes of retinal ganglion cells, understanding normal aging and its effects on visual function, the development of new diagnostic tests for glaucoma, and the application of new techniques such as machine learning classifiers for the interpretation and classification of complex datasets from well-established, as well as newly developed diagnostic tests and other information used in the diagnosis and management of glaucoma.

Dr. Sample is a 1999 recipient of the Lew R. Wasserman Merit Award from Research to Prevent Blindness. She serves on several national and international committees. She was a panel member for development of the NEI National Vision Research Plan. She is a board member for the International Perimetric Society, an editorial board member for the *Journal of Glaucoma*, and a member of the Ethics and Regulations for Clinical Research Committee of the Association for Research in Vision and Ophthalmology.

Michael H. Goldbaum received the M.D. degree in 1965 from Tulane University School of Medicine, New Orleans, LA, an ophthalmology certificate in 1969 from the U. S. Navy, and a Retina Fellowship certificate in 1972 from Cornell University, New York Hospital. Ithaca. He received the M.S. degree in medical informatics from Stanford University, Stanford, CA, in 1988.

He is a Professor of Ophthalmology, Department of Ophthalmology, University of California, San Diego. He taught ophthalmology at the University of Illinois, Urbana, from 1973 to 1976. He has been teaching ophthalmology at the University of California at San Diego since 1977. His research interests are image understanding applied to ophthalmic images, Bayesian nets for inferencing about ophthalmic images, and machine-learning classification to improve diagnosis and prediction for glaucoma.

Dr. Goldbaum was awarded a Senior Honor Award from the American Academy of Ophthalmology in 2001. He has served on the program and planning committees of the Association for Research in Vision and Ophthalmology.

Robert N. Weinreb received the S.B. degree in electrical engineering, 1971 from the Massachusetts Institute of Technology, Cambridge, in 1971 and the M.D. degree from Harvard Medical School, Cambridge, MA, in 1975.

Both a clinician and scientist, he has been Professor of Ophthalmology at the University of California, San Diego since 1984. His notable contributions to glaucoma include developing methods for quantitatively assessing the optic disc and nerve fiber layer and elucidating cellular and molecular mechanisms of aqueous outflow.

Dr. Weinreb is the current President of the Association of Research in Vision and Ophthalmology (ARVO) and President-Elect of the Association of International Glaucoma Societies.

Terrence J. Sejnowski (M'83–SM'91–F'00) received the B.S. degree from the Case-Western Reserve University, Cleveland, OH, in 1968 and the M.A. and Ph.D. degrees from Princeton University, Princeton, NJ, in 1970 and 1978, respectively, all in physics.

He is an Investigator with Howard Hughes Medical Institute and a Professor at the Salk Institute for Biological Studies, La Jolla, CA, where he directs the Computational Neurobiology Laboratory. He is also Professor of Biology at the University of California, San Diego, where he is Director of the Institute for Neural Computation. In 1988, he founded *Neural Computation*, (MIT Press). The long-range goal of his research is to build linking principles from brain to behavior using computational models.