Comparing Machine Learning Classifiers for Diagnosing Glaucoma from Standard Automated Perimetry

Michael H. Goldbaum,[1] *Pamela A. Sample,*[2] *Kwokleung Chan,*[3,4] *Julia Williams,*[2] *Te-Won Lee,*[3,4] *Eytan Blumenthal,*[2] *Christopher A. Girkin,*[5] *Linda M. Zangwill,*[2] *Christopher Bowd,*[2] *Terrence Sejnowski,*[3,4] *and Robert N. Weinreb*[2]

Purpose. To determine which machine learning classifier learns best to interpret standard automated perimetry (SAP) and to compare the best of the machine classifiers with the global indices of STATPAC 2 and with experts in glaucoma.

Methods. Multilayer perceptrons (MLP), support vector machines (SVM), mixture of Gaussian (MoG), and mixture of generalized Gaussian (MGG) classifiers were trained and tested by cross validation on the numerical plot of absolute sensitivity plus age of 189 normal eyes and 156 glaucomatous eyes, designated as such by the appearance of the optic nerve. The authors compared performance of these classifiers with the global indices of STATPAC, using the area under the ROC curve. Two human experts were judged against the machine classifiers and the global indices by plotting their sensitivity–specificity pairs.

Results. MoG had the greatest area under the ROC curve of the machine classifiers. Pattern SD (PSD) and corrected PSD (CPSD) had the largest areas under the curve of the global indices. MoG had significantly greater ROC area than PSD and CPSD. Human experts were not better at classifying visual fields than the machine classifiers or the global indices.

Conclusions. MoG, using the entire visual field and age for input, interpreted SAP better than the global indices of STATPAC. Machine classifiers may augment the global indices of STATPAC. (*Invest Ophthalmol Vis Sci.* 2002;43:162–169)

Classification permeates medical care. Much of the management of glaucoma depends on the diagnosis of glaucoma or the risk of its progression. The specific aims of this study were (1) to determine which machine learning classifiers best interpret standard automated perimetry and (2) to compare the performance of the best classifiers with the global indices in STATPAC 2 and with experts in glaucoma.

Appropriate glaucoma evaluation requires examination of the optic disc and visual field testing. Automated threshold perimetry has grown in popularity largely because it provides calibrated, detailed quantitative data that can be compared over time and among different centers. Interpretation of the visual field remains problematic to most clinicians.[1]

From the [1]Ophthalmic Informatics Laboratory, [2]Glaucoma Center and Research Laboratories, Department of Ophthalmology, [3]Institute for Neural Computation, University of California at San Diego, La Jolla, California; [4]Computational Neurobiology Laboratory, Salk Institute, La Jolla, California; and [5]University of Alabama, Birmingham.

Supported by National Institutes of Health Grants EY08208 (PAS) and EY13235 (MHG), and a grant from the Foundation for Eye Research (EB).

Submitted for publication June 1, 2001; revised August 29, 2001; accepted September 14, 2001.

Commercial relationships policy: N.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Corresponding author: Michael H. Goldbaum, Department of Ophthalmology, 9500 Gilman Drive, La Jolla, CA 92093-0946; mgoldbaum@ucsd.edu.

It is difficult to separate true visual field loss from fluctuations in visual field results that may arise from learning effects, fatigue, and the long-term fluctuation inherent in the test.[2,3] This fluctuation makes the identification of glaucoma and the detection of its progression difficult to establish.

We investigated classification techniques to improve the identification of glaucoma using SAP. Neural networks have been previously applied in ophthalmology to interpret and classify visual fields,[4–6] detect visual field progression,[7] assess structural data from the optic nerve head,[8] and identify noise from visual field information.[9] Neural networks have improved the ability of clinicians to predict the outcome of patients in intensive care, diagnose myocardial infarctions, and estimate the prognosis of surgery for colorectal cancer.[10–12] We applied a broad range of popular or novel machine classifiers that represent different methods of learning and reasoning.

Methods

Subjects

Population Source and Criteria. Visual field data came from our longitudinal study of visual function in glaucoma. Normal subjects were recruited from the community, staff, and spouses or friends of subjects. Primary open-angle glaucoma patients were recruited from the Glaucoma Center, University of California at San Diego. Informed consent was obtained from all participants, and the study was approved by the Institutional Review Board of the University of California at San Diego. This research follows the tenets of the Declaration of Helsinki.

Exclusion criteria for both groups included unreliable visual fields (defined as fixation loss, false-negative and false-positive errors $\geq$ 33%),[13] angle abnormalities on gonioscopy, any diseases other than glaucoma that could affect the visual fields, and medications known to affect visual field sensitivity. Subjects with a best-corrected visual acuity worse than 20/40, spherical equivalent outside $\pm$5.0 diopters, and cylinder correction >3.0 diopters were excluded. Poor quality stereoscopic photographs of the optic nerve head served as an exclusion for the glaucoma population. A family history of glaucoma was not an exclusion criterion.

Inclusion criteria for the glaucoma category were based on optic nerve damage and not visual field defects. The classification of an eye as glaucomatous or normal was based on the consensus of masked evaluations of two independent graders of a stereoscopic disc photograph. All photograph evaluations were accomplished using a stereoscopic viewer (Asahi Pentax Stereo Viewer II) illuminated with color-corrected fluorescent lighting. Glaucomatous optic neuropathy (GON) was defined by evidence of any of the following: excavation, neuroretinal rim thinning or notching, nerve fiber layer defects, or an asymmetry of the vertical cup/disc ratio > 0.2. Inconsistencies between grader's evaluations were resolved through adjudication by a third evaluator.

Inclusion criteria for the normal category required that subjects have normal dilated eye examinations, open angles, and no evidence of visible GON. Normal optic discs had a cup-to-disc ratio asymmetry $\leq$ 0.2, intact rims, and no hemorrhages, notches, excavation, or nerve fiber layer defects. Normal subjects had intraocular pressure (IOP) $\leq$

Investigative Ophthalmology & Visual Science, January 2002, Vol. 43, No. 1

22 mm Hg with no history of elevated IOP. Excluded from the normal population were suspects with no GON and with IOP ≥ 23 mm Hg on at least two occasions. These suspects are part of a separate study on classification of stratified patient populations.

Only one eye per patient was included in the study. If both of the eyes met the inclusion criteria, only one of the eyes was selected at random. The final selection of eyes totaled 345, including 189 normal eyes (age, 50.0 ± 6.7 years; mean ± SD) and 156 eyes with GON (62.3 ± 12.4 years).

Optic Nerve Photographs. Color simultaneous stereoscopic photographs were obtained using a Topcon camera (TRC-SS; Topcon Instrument Corp of America, Paramus, NH) after maximal pupil dilation. These photographs were taken within 6 months of the field in the data set. Stereoscopic disc photographs were recorded for all patients with the exception of a subset of normal subjects ($n = 95$) for whom photography was not available. These normal subjects had no evidence of optic disc damage with dilated slit-lamp indirect ophthalmoscopy with a hand-held 78 diopter lens.

Visual Field Testing. All subjects had automated full threshold visual field testing with the Humphrey Field Analyzer (HFA; Humphrey-Zeiss, Dublin, CA) with program 24-2 or 30-2. The visual field locations in program 30-2 that are not in 24-2 were deleted from the data and display.

Summary Statistics. The HFA perimetry test provides a statistical analysis package referred to as STATPAC 2 to aid the clinician in the interpretation of the visual field results. A STATPAC printout includes the numerical plot of absolute sensitivity at each test point, grayscale plot of interpolated raw sensitivity data, numerical plot and probability plot of total deviation, and numerical plot and probability plot of pattern deviation.[14] Global indices are statistical classifiers tailored to SAP: mean deviation (MD), pattern SD (PSD), short-term fluctuations (SF),[15] corrected pattern SD (CPSD), and glaucoma hemifield test (GHT).[16] The clinician uses these plots and indices to estimate the likelihood of glaucoma from the pattern of the visual field.

Visual Field Presentation to Glaucoma Experts. Two glaucoma experts masked to patient identity, optic nerve status, and diagnosis independently interpreted the perimetry as glaucoma or normal. We elected to compare the human experts, STATPAC, and the machine classifiers with each type of classifier having received equivalent input. The printout given to the glaucoma experts for evaluation was the numerical plot of the total deviation, because that was the format closest to the data supplied to the machine classifiers (absolute sensitivities plus age) that the experts were used to interpreting.

Visual Field Presentation to Machine Classifiers. The input to the classifiers for training and diagnosis included the absolute sensitivity in decibels of each of the 52 test locations (not including two locations in the blind spot) in the 24-2 visual field. These values were extracted from the Humphrey field analyzer using the Peridata 6.2 program (Peridata Software GmbH, Huerth, Germany). Because the total deviation numerical plots used by the experts were derived using the age of the subject, an additional feature provided to the machine classifiers was the subject's age.

Classification

The basic structure of a classifier is input, processor, and output. The input was the visual field sensitivities at each of 52 locations plus age. The processor was a human classifier, such as a glaucoma expert; a statistical classifier, such as the STATPAC global indices; or a machine classifier. The output was the presence or absence of glaucoma.

Supervised learning classifiers learn from a teaching set of examples of input–output pairs; for each pattern of data, the corresponding desired output of glaucoma or normal is known. During supervised learning, the classifier compares its predictions to the target answer and learns from its mistakes.

Data Preprocessing. Some classifiers have difficulty with high-dimension input. Principal component analysis (PCA) is a way of reducing the dimensionality of the data space by retaining most of the information in terms of its variance.[17] The data are projected onto their principal components. The first principal component lies along the axis that shows the highest variance in the data. The others follow in a similar manner such that they form an orthogonal set of basis functions. For the PCA basis, the covariance matrix of the data are computed, and eigenvalues of the matrix are ordered in a decreasing manner.

Statistical Classifiers. Learning statistical classifiers use multivariate statistical methods to distinguish between classes. There are limitations from this approach. These methods assume that a certain form, such as linearity (homogeneity of covariance matrices), characterizes relationships between variables. Failure of data to meet these requirements degrades the classifier's performance. Missing values and the quality of the data may be problematic. With statistical classifiers, such as linear discriminant function, the separation surface configuration is usually fixed.

Linear Discriminant Function. Linear discriminant function (LDF) learned to map the 53-feature input into a binary output of glaucoma and not glaucoma. A separate analysis was done with the 53-dimension full data set reduced to eight-dimension feature set by PCA.

Analysis with STATPAC Global Indices. The global indices, MD, PSD, and CPSD, were tested in the same fashion as the machine classifiers, with receiver operating characteristic (ROC) curves and with sensitivity values at defined specificities. The sensitivity and specificity for the glaucoma hemifield test result were computed by converting the plain text result output, "outside normal limits" vs. "within normal limits" or "borderline," to glaucoma versus normal by combining the "borderline" into the "normal" category.

Machine Classifiers. The attractive aspect of these classifiers is their ability to learn complex patterns and trends in data. As an improvement compared with statistical classifiers, these machine classifiers adapt to the data to create a decision surface that fits the data without the constraints imposed by statistical classifiers.[18] Multilayer perceptrons (MLP), support vector machines (SVM), mixture of Gaussian (MoG), and mixture of generalized Gaussian (MGG) are effective machine classifiers with different methods of learning and reasoning. The following paragraphs describe the training of each classifier type. Readers who want detailed descriptions with references of the machine classifiers will find them in the Appendix.[19–33]

Multilayer Perceptron with Learning by Backpropagation of Error Correction. The multilayer perceptron was set up with the Neural Network toolbox 3.0 of MATLAB. The training was accomplished with the Levenberg-Marquart (LM) enhancement of backpropagation. The data for each of the 53 input nodes were renormalized by removing the mean and dividing by the SD. The input nodes were fed into one hidden layer with 10 nodes activated by hyperbolic tangent functions. The output was a single node with a logistic function for glaucoma (1) and normal (0). The learning rate was chosen by the toolbox itself. Training was stopped early when no further decrease in generalization error was observed in a stopping set. The sensitivity–specificity pairs were plotted as the ROC curve.

Support Vector Machine. The class y for a given input vector, $\mathbf{x}$, was $y(\mathbf{x}) = \text{sign}\,(\sum_{i=1}^{p} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b)$, where b was the bias, and the coefficients α_i were obtained by training the SVM. The SVMs were trained by implementing Platt's sequential minimal optimization algorithm in MATLAB.[34–36] The training of the SVM was achieved by finding the support vector components, $\mathbf{x}_i$ and the associated weights, $\boldsymbol{\alpha}_i$. For the linear function, $K(\mathbf{x},\mathbf{x}_i)$, the linear kernel was $(\mathbf{x} \cdot \mathbf{x}_i)$, and the Gaussian kernel was $\exp(-0.5(\mathbf{x} - \mathbf{x}_i)^2/\sigma^2)$. The penalty used to avoid overfit was $C = 1.0$ for either the linear or Gaussian kernel. With the Gaussian kernel, the choice of σ depended on input dimension, $\sigma \propto \sqrt{53}$ or $\sigma \propto \sqrt{8}$. The output was constrained between 0 and 1 with a logistic regression. If the output value was on the positive side of the decision surface, it was considered glaucomatous; if it was on the negative side of the decision surface, it was considered nonglaucomatous. When generating the ROC curve, scalar output of the SVMs was extracted so that the decision threshold could be varied to obtain different sensitivity–specificity pairs for the ROC curve.

TABLE 1. Comparison of Sensitivities at Particular Specificities and Comparison of Areas under Entire ROC Curve

	Sensitivity at Specificity = 1	Sensitivity at Specificity = 0.9	Specificity of Expts.	Sensitivity of Expts.	ROC Area ± SE
Human Experts on Standard Automated Perimetry					
Expt 1			0.96	0.75	
Expt 2			0.59	0.88	
STATPAC Global Indices					
MD	0.45	0.65			0.837 ± 0.022
PSD	0.61	0.76			**0.884 ± 0.020**
CPSD	**0.64**	0.74			0.844 ± 0.025
GHT	**0.67**				
Statistical Classifier					
LDF	0.32	0.60			0.832 ± 0.023
LDF with PCA	0.48	0.64			0.879 ± 0.018
Machine Classifiers					
MLP	0.25	0.75			0.897 ± 0.017
MLP with PCA	0.54	0.71			0.893 ± 0.018
SVM linear	0.44	0.69			0.894 ± 0.017
SVM linear with PCA	0.51	0.67			0.887 ± 0.018
SVM Gaussian	0.53	0.71			0.903 ± 0.017
SVM Gaussian with PCA	0.57	0.75			0.899 ± 0.017
MoG (QDF)	0.61	0.79			0.917 ± 0.016
MoG (QDF) with PCA	0.67	0.78			0.919 ± 0.016
MoG with PCA	**0.67**	**0.79**			**0.922 ± 0.015**
MGG with PCA	0.01	0.78			0.906 ± 0.022

Boldface indicates high performer within grouping of classifiers.

Mixture of Gaussian and Mixture of Generalized Gaussian. To train the classifier, in general the data are analyzed to determine whether unsupervised learning finds more than one cluster for each of the classes. The assumption is that the class conditional density of the feature set approximates a mixture of normal multivariate densities for each cluster of each class (e.g., glaucoma or not glaucoma). The training is accomplished by fitting mixture of Gaussian densities to each class by maximum likelihood. With the class conditional density modeled as a mixture of multivariate normal densities for each class, Bayes' rule is used to obtain the posterior probability of the class, given the feature set in a new example.

Mixture of Gaussian was performed both with the complete 53-dimension input and with the input reduced to 8 dimensions by PCA. The computational work of training with the 53-dimension input was made manageable by limiting the clusters to one each for normal and glaucoma populations in the teaching set. This limitation yielded performance similar to quadratic discriminant function (QDF). The training was accomplished by fitting the glaucoma and normal populations each with a multivariate Gaussian density. For SAP vectors, $\mathbf{x}$, we computed $P[\mathbf{x}|\overline{G}]$ and $P[\mathbf{x}|G]$. From these conditional probabilities, we could obtain the probability of glaucoma for a given SAP, $\mathbf{x}$, by Bayes rule.

Because of the limited number of patients compared with the dimension of the input space, we also analyzed the data with the feature space reduced to eight dimensions with PCA, which contained >80% of the original variance. The number of clusters in each group, generally two, was chosen to optimize ROC area. The ROC curve was generated by varying the decision threshold.

Training of mixture of generalized Gaussian was similar to that done for MoG, except it was accomplished by gradient ascent on the data likelihood.[37] MGG was trained and tested only with input reduced to eight dimensions by PCA.

Statistical Analysis

Sensitivity and Specificity. The sensitivity (the proportion of glaucoma patients classified as glaucoma) and the specificity (the proportion of normal subjects classified as normal) depend on the placement of the threshold along the range of output for a classifier. To ease the comparison of the classifiers, we have displayed the sensitivity at defined specificities (Table 1).

ROC Curve. An ROC curve was constructed for each of the global indices and each of the learning classifiers (Table 1). The area under the ROC curve, bounded by the ROC curve, the abscissa, and the ordinate, quantified the diagnostic accuracy of a test in a single number, with 1 indicating perfect discrimination and 0.5 signifying discrimination no better than random assignment.

The area under the ROC curve served as a comparison of the classifiers. A number of statistical approaches have been developed for determining a significant difference between two ROC curves.[38–42] The statistical test we used for significant difference between ROC curve areas was dependent on the correlation of the curves (Table 2).[39] Without preselection of the comparisons, there were 45 comparisons of classifiers. For $\alpha = 0.05$, the Bonferroni adjustment required $P \leq 0.0011$ for the difference to be considered significant (Table 2).

The shape of the ROC curve can vary; one curve may compare favorably with another at low specificity but differ at high specificity. To compare the same region of multiple ROC curves, we compared the sensitivities at particular specificities.

Cross Validation. The ultimate goal is that a learning classifier should become trained well enough on its teaching examples (apparent error rate) to be able to generalize to new examples (actual error rate). The actual error rate was determined with cross validation. We randomly partitioned the glaucoma patients and the normal subjects each into 10 partitions and combined one partition from the glaucoma patients with one partition from the normal subjects to form each of the 10 partitions of the data set. One partition of the data set became the test set, and the remaining nine partitions of the data set were combined to form the teaching set. During the training of the multilayer perceptron, another set was used as a stopping set to determine when training was complete, and the eight remaining partitions were combined into the teaching set.[43] The training-test process was repeated until each partition had an opportunity to be the test set. Because the classifier was forced to generalize its knowledge on previously unseen data, we determined the actual error rate.

Comparisons

The STATPAC global indices, statistical classifier, and machine classifiers were compared by the area under the entire ROC curve.[39] Glaucoma experts consider the cost of a false positive to be greater than a

TABLE 2. Significance of Difference (*P*) of Compared ROC Area and Correlation Coefficients for Values along the Compared Curves

Classifier		MD	PSD	CPSD	LDF PCA*	MLP	SVM linear	SVM Gauss	MoG(QDF)†	MoG PCA	MGG PCA
ROC Area		**0.837**	**0.884**	**0.844**	**0.879**	**0.878**	**0.894**	**0.903**	**0.916**	**0.922**	**0.906**
MD	*P* value		0.019	0.77	0.018	0.007	**0.0001**	**<0.00005**	**<0.00005**	**<0.00005**	**0.0005**
0.837	Correlation		*0.55*	*0.42*	*0.61*	*0.60*	*0.76*	*0.76*	*0.54*	*0.56*	*0.49*
PSD				0.022	0.83	0.44	0.55	0.19	**0.0009**	0.006	0.18
0.884				*0.73*	*0.48*	*0.55*	*0.56*	*0.69*	*0.88*	*0.72*	*0.60*
CPSD					0.16	0.024	0.034	0.007	**0.0003**	**0.0001**	0.004
0.844					*0.38*	*0.42*	*0.42*	*0.52*	*0.58*	*0.62*	*0.54*
LDF PCA						0.13	0.56	0.020	0.021	0.0055	0.10
0.879						*0.77*	*0.91*	*0.83*	*0.56*	*0.58*	*0.55*
MLP							0.71	0.55	0.17	0.10	0.60
0.878							*0.86*	*0.85*	*0.65*	*0.56*	*0.53*
SVM linear								0.18	0.12	0.048	0.44
0.894								*0.92*	*0.63*	*0.62*	*0.58*
SVM Gauss									0.26	0.15	0.84
0.903									*0.75*	*0.67*	*0.60*
MoG(QDF)										0.68	0.52
0.916										*0.62*	*0.50*
MoG PCA											0.044
0.922											*0.88*

Boldface indicates statistical significance after Bonferroni adjustment ($P < 0.0011$ for $\alpha < 0.05$). Italics signify correlation coefficient.

*PCA indicates that full data set of 53 features was reduced to 8 features by principal component analysis for classifiers that have difficulty with high-dimension data sets.

†MoG(QDF) means MoG with full data set constrained to one cluster each for normal and glaucoma populations, approaching characteristics of quadratic discriminant function.

false negative. A high specificity is desirable because the prevalence of glaucoma is low and progression is very slow. The left-hand end of the ROC curves are of interest when high specificity is desired. Consequently, we also compared the sensitivities at specificities 0.9 and 1.0.

Sensitivity–specificity pairs that did not correspond to specificities 0.9 and 1.0 were indicated on the ROC plots (see Figs. 1 through 3). The classification results of the two glaucoma experts and the glaucoma hemifield test were represented on the ROC plots by single sensitivity–specificity pairs. We compared the sensitivity of the classifiers at specificity 1.0 for comparison with GHT, at specificity 0.995. We also analyzed the false-positive and false-negative visual fields for each of the classifiers.

RESULTS

Normal and Glaucoma Groups

The mean and SD of the visual field MD of the patients in the glaucoma group was -3.9 ± 4.3 dB. Most glaucoma patients had early to moderate glaucoma; only 3 of the 156 glaucoma patients had advanced glaucoma.

Glaucoma Experts

The current results of the comparison of classifiers for SAP are summarized in Table 1 and Figures 1 through 3. Expert 1, analyzing SAP total deviation numeric plots, had a sensitivity of 0.75 and a specificity of 0.96. Expert 2 had a sensitivity of 0.88 and a specificity of 0.59. These values were similar to the best of STATPAC and the best machine classifiers (see Fig. 3 and Table 1).

STATPAC 2 and Statistical Classifiers

The global indices with the highest ROC areas were PSD and CPSD (Fig. 1 and Tables 1 and 2). Correction of PSD for short-term fluctuation (CPSD) resulted in a difference in the area for the entire ROC curve, but it was PSD that had the higher ROC area. Only at specificity 1 was the sensitivity of CPSD greater than PSD, but not significantly. MD had lower area under the ROC curve than CPSD and PSD. There was poor correlation between MD and PSD ($\rho = 0.55$) and between MD and CPSD ($\rho = 0.42$).

GHT is a special case, because it is constrained to a specificity of 0.995. It is therefore best compared with results of all the classifiers at specificity = 1. With our data, GHT had no false positives; hence, its specificity was 1. At specificity 1, the other global indices had sensitivities less than GHT (0.67).

Linear discriminant function is a statistical classifier that is not specifically designed for SAP. The area under the entire ROC curve was similar for LDF (0.832) and MD (0.837). The sensitivity of LDF was less than all the global indices at high specificities (Fig. 1). Reducing the dimension of the feature set to eight by PCA improved ROC area of LDF (0.879), but not quite significantly ($P = 0.0038$, compared with the Bonferroni cutoff of 0.0011). There was poor correlation between LDF with PCA and between PSD and CPSD ($\rho = 0.48$ and 0.38, respectively).

Machine Classifiers

PCA did not improve the ROC areas for MLP, SVM linear, or SVM Gaussian (Table 1). These classifiers were able to learn and classify from high-dimension data. Mixture of Gaussian and Mixture of Generalized Gaussian are less efficient with high-dimension input. Reducing the dimensionality of the input by PCA permitted two clusters for glaucoma and one cluster for normal, which allowed the full capabilities of these classifiers to manifest. MoG with PCA had higher area under the ROC curve (0.922) than MoG constrained to QDF (0.917) with the full data set, yet it was MoG constrained to QDF that was significantly higher than PSD ($P = 0.0009$), because there was higher correlation between the curves for PSD and MoG constrained to QDF. Removing age from the data set lowered the area under the curve for MoG constrained to QDF by 0.008 (from 0.917 to 0.909). Though MoG with PCA reported a higher sensitivity (0.673) at specificity 1 than GHT (0.667), these values were similar.

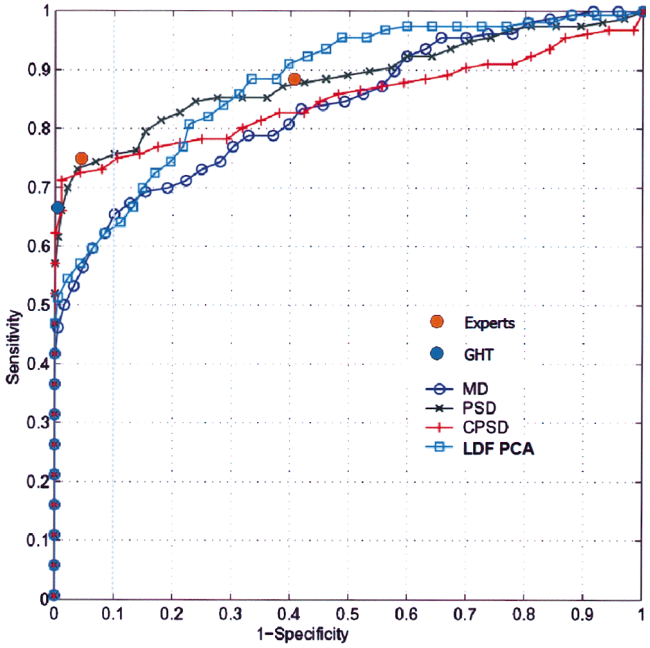

FIGURE 1. ROC curves for global indices from STATPAC and another statistical classifier, LDF with PCA. The glaucoma experts and GHT are indicated by single sensitivity–specificity pairs denoted by *solid circles.*

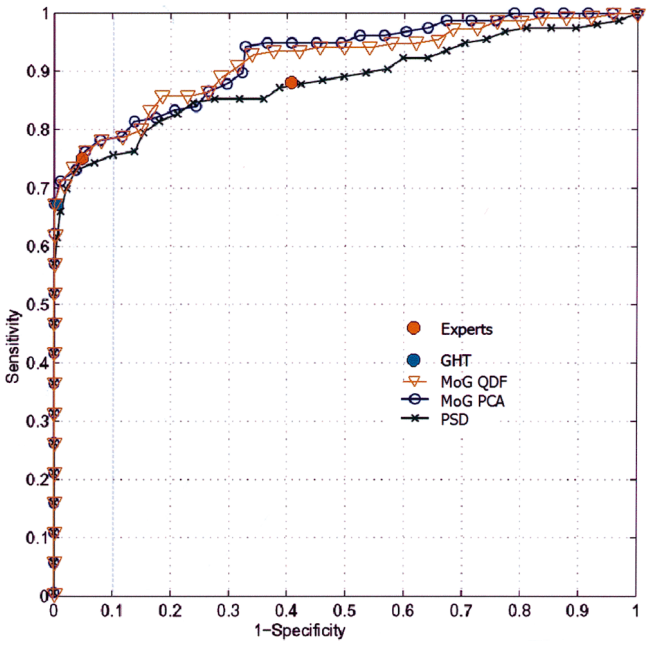

FIGURE 3. ROC curves comparing the best of the machine classifiers with the best of the STATPAC global indices. The glaucoma experts and GHT are superimposed on the curves as described in Figure 1.

MGG analysis was done only with PCA, because of the complexity of this analysis with the full data set input. There was one cluster for each class. The two MoG curves and the MGG curve were similar between specificities 0.9 and 1, and all three had higher sensitivities than the other machine classifiers in this range (Fig. 2 and Table 1).

Errors

The best expert had a specificity of 0.96. Therefore, we evaluated the incorrect classifications by the best of each type of classifier (MoG, expert 1, and PSD) at specificity 0.96. Table 3 demonstrates visual field characteristics of the eyes with GON that were misclassified as normal (false negatives) by the best classifier of each type at specificity 0.96. There was no significant difference in the number of false negatives of each classifier (41, 39, and 41, respectively). Of these, 37 (90%), 37 (95%), and 40 (98%), respectively, had visual fields characterized as normal; there was no significant difference in these values. The means of the mean deviation, number of total deviation locations with probability $< 5\%$, number of pattern deviation locations with probability $< 5\%$, and number of contiguous pattern deviation locations with probability $< 5\%$ were all within the range considered clinically normal and were similar for each classifier. The concordance of false negatives was 0.94 between MoG and PSD, 0.92 between MoG and expert 1, and 0.94 between expert 1 and PSD. Thirty-four fields were misclassified by all three classifiers.

Table 4 displays the visual field characteristics of the normal eyes that were misclassified as glaucomatous (false positives). All the misclassified fields had visual fields characterized as normal, except for one normal field misclassified by PSD that had characteristics of early field loss. The means of the STATPAC plots described above were all in the normal range and were similar between classifiers. The concordance of the false positives was 0.96 between MoG and PSD, 0.98 between MoG and expert 1, and 0.96 between expert 1 and PSD. Three fields were misclassified by all three classifiers.

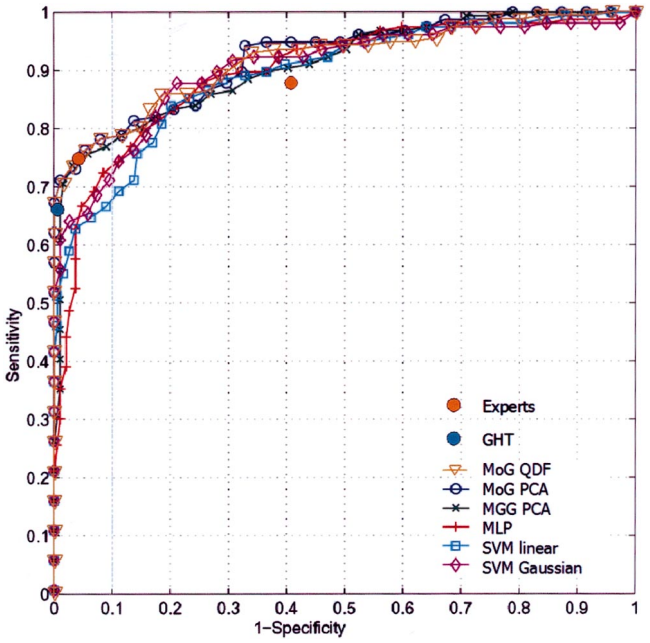

FIGURE 2. ROC curves for the best of each type of machine classifier. The glaucoma experts and GHT are superimposed on the curves as described in Figure 1.

DISCUSSION

The new machine classifiers are quite effective in interpreting SAP, because they compare favorably with current classification methods with STATPAC and because they performed at least as well as trained glaucoma experts. Several factors may further improve the performance of the classifiers relative to human experts and STATPAC: (1) A larger normative group, such as that used in STATPAC may improve the classifiers discriminative performance. (2) It is possible that the classifiers might have done even better if they had been compared with general ophthalmologist, optometrists, or residents who are

TABLE 3. False Negatives out of 156 GON at Specificity 0.96 with Best of Each Classifier Type

	MoG (n = 41)	Expt. 1 (n = 39)	PSD (n = 41)	All Three (n = 34)
Mean Deviation	−0.80 ± 1.49*	0.74 ± 1.58	0.82 ± 1.67	0.52 ± 1.36
No. of points $P < 5\%$ total deviation	3.95 ± 4.89	4.03 ± 5.87	4.49 ± 7.06	2.82 ± 3.91
No. of points $P < 5\%$ pattern deviation	3.02 ± 3.55	2.36 ± 1.97	2.32 ± 1.85	2.09 ± 1.94
No. of contiguous points pattern deviation $P < 5\%$	1.80 ± 2.46	1.26 ± 1.27	1.20 ± 1.19	1.15 ± 1.23

Values are means ± SD, with no. of GON called normal in parentheses.

less familiar with grading visual fields, but this remains to be studied. (3) The normal subjects in this study did not have visual field experience for the most part; whereas, most of the patients with GON had such prior experience. It is well documented that learning effects can occur during the first two visual fields.

The appearance of the optic nerve was the indicator for glaucoma. Issues concerning the shape of the optic nerve head could have affected the training of the classifiers, and idiosyncrasies of optic nerve head shape in the study sample might have impacted the representativeness of the classifiers. For example, if diffuse rim loss is underrepresented in the study sample and diffuse rim loss is associated with diffuse field loss, then the true discriminatory potential of the MD (relative to PSD or CPSD) might have been underestimated in the study.

The better performance of the global indices in STATPAC compared with LDF demonstrates the benefit of designing classifiers specifically for the data from SAP. The machine classifiers are general classifiers that are not optimized for SAP data. Nevertheless the MoG constrained to QDF and the MoG with PCA each significantly outperformed the global indices as measured by area under the entire ROC curve. The MoG classifiers functioned no better than PSD in the high-specificity region. The two MoG classifiers gave the same results as the GHT at the usual specificity of the GHT test.

The differences between the individual machine classifiers are greatest at high specificities, a property considered desirable for glaucoma diagnosis. Because most of the difference between the two MoG curves and the rest of the machine classifiers was in the high-specificity region, we can infer that the difference between these curves was due mostly to the separation of the curves in the high-specificity region.

Though age minimally improves the learning and diagnosis with the machine classifiers, it is uncertain how age contributes. It is possible that age combines with the visual field locations in a manner similar to the way age transforms the absolute numerical plot to the total deviation numerical plot. It is equally plausible that the classifier simply adjusts for the mean age of 50 for normal population and 62.3 in the glaucoma population.

The 34 false negatives of 156 and the 3 false positives of 189 that were misclassified by all three classifiers representing the best of each classifier type may be close to the minimal error attainable from visual fields, given that the gold standard for glaucoma in this study is GON. Analysis of the patterns of the fields misdiagnosed by the three classifiers indicates that the false negatives or false positives appear normal.

It is difficult to compare our results with other efforts at automated diagnosis from visual fields, because the study populations were different and other studies used human interpretation of visual fields as a gold standard for the diagnosis of glaucoma.[8,44,45] Spenceley et al.[45] reported sensitivity of 0.65 and 0.90 at specificities 1.0 and 0.96, respectively, with MLP; the MLP was taught which fields were glaucomatous and which were normal from an interpretation of the fields by an observer. We obtained sensitivities of 0.67 and 0.73 at these specificities with MoG, our best classifier; the machine classifiers were taught which fields were glaucomatous and which were normal from an interpretation of the optic nerve for the presence of GON by the consensus of two observers. Researchers using pattern recognition methodology consider that an indicator other than the test being evaluated should be used as a gold standard for teaching the classifiers. Also, if the human interpretation of the visual field is used as the indicator for teaching the classifier, the classifier cannot exceed the human interpreter in accuracy. With GON as the indicator of disease, we found that the MoG machine classifiers generated ROC curves that were higher than the sensitivity–specificity pairs from glaucoma experts. Other studies used MLPs for automated diagnosis.[5,8,44,45] We found that the ROC curves of the MoG machine classifiers were higher than the curve for MLP, particularly in the high-specificity region. This observation implies that the MoG classifiers perform better than the MLP used in previous reports.

After long-term experience with SAP, glaucoma experts have learned how to interpret the results. The glaucoma experts performed well, but no better than the machine classifiers. There are newer perimetric tests, such as short-wavelength automated perimetry (SWAP),[46,47] and frequency-doubling technology perimetry (FDT),[48,49] with which clinicians and researchers have less experience. It is likely that machine classifiers will be able to learn from these data and exceed the ability of glaucoma experts in interpreting these tests.

This report describes our success at identifying new machine classifiers that compare favorably with the current interpreters of standard automated perimetry. The benefits that refined information from machine classifiers may add to the plots and indices that STATPAC offers in a clinical setting may can be addressed in future studies. There are methods that can improve even more the performance of the machine classifiers in interpreting perimetry and in extracting information from the perimetry. Classification may be improved by finding better data to distinguish the classes, by identifying better classifiers, or by optimizing the process. The newer perimetry tests,

TABLE 4. False Positives at Specificity 0.96 with Best of Each Classifier Type

	MoG (n = 8)	Expt. 1 (n = 8)	PSD (n = 8)	All Three (n = 3)
Mean deviation	−1.13 ± 1.77	−1.91 ± 0.82	−1.63 ± 1.28	−2.29 ± 0.68
No. of points $P < 5\%$ total deviation	7.13 ± 6.01	8.32 ± 5.83	7.88 ± 6.13	12.00 ± 6.00
No. of points $P < 5\%$ pattern deviation	6.13 ± 4.45	5.75 ± 4.43	6.63 ± 4.10	10.00 ± 4.58
No. of contiguous points pattern deviation $P < 5\%$	2.63 ± 2.20	3.00 ± 1.69	3.00 ± 1.85	4.67 ± 1.53
Normal field	8	8	7	3

Values are means ± SD, with no. of GON called normal in parentheses.

SWAP and FDT, are examples of efforts to improve the data set. This report describes the identification of the best classifiers for SAP with our data set. In a separate report we will describe the optimization of the process that results from identifying the most useful visual field locations and from removing the noncontributing field locations. In addition, these methods may need adjustment for different patient populations, and validations in a variety of settings will be needed.

Our experience with machine learning classifiers indicates that there is additional useful information in visual field tests for glaucoma. Machine classifiers are able to discover and use perimetric information not obvious to experts in glaucoma. There are a number of applications in ophthalmic research to which classifier methodology could be applied.

APPENDIX

Multilayer Perceptron

The multilayer perceptron (MLP) is one of the most popular architectures among other neural networks for its efficient training by error backpropagation.[19–21] The MLP has been successfully applied to a wide class of problems, such as face recognition[22] and character recognition.[23] The architecture is a universal feed-forward network; the input layer and output layer of nodes are separated by one or more hidden layers of nodes. The hidden layers act as intermediary between the input and output layers, enabling the extraction of progressively useful information obtained during learning. The activation function of each neuron uses a sigmoid function to approximate a threshold or step. The use of a continuous sigmoid function instead of a step function enables the generation of an error function for correcting the weights. The sigmoid function may be logistic or hyperbolic tangent.

During learning, there are two passes through the layers of the network. In the forward pass, the data in the input source nodes are weighted by the connections, summed, and transformed by the activation function. This process continues up the layers to the output node, where the values generated are compared with the desired output. The error signal is passed backward to reinforce or inhibit each weight. Each sample in the teaching set is similarly processed. The procedure is repeated for the entire teaching set, descending the error surface until there is an acceptably low total error rate for the stopping set. The ability of the network to generalize what it has learned is tested with a set of data different from the teaching set.

Support Vector Machine

Support Vector Machines (SVMs) are a new class of learning algorithms that are able to solve a variety of classification and regression (model fitting) problems.[24,25] They exploit statistical learning theory to minimize the generalization error when training a classifier. SVMs have generalized well in face recognition,[26] text categorization,[27] recognition of handwritten digits,[28] and breast cancer diagnosis and prognosis.[29]

For a two-class classification problem, the basic form of SVM is a linear classifier, $f(u) = \text{sign}(u) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$, where $\mathbf{x}$ is the input vector, $\mathbf{w}$ is the adjustable weight vector, $\mathbf{w}^T\mathbf{x} + b = 0$ is the hyperplane decision surface, $f(u) = -1$ designates one class (e.g., normal) and $f(u) = 1$ the other class (e.g., glaucoma). For linearly separable data, the parameters $\mathbf{w}$ and b are chosen such that the margin ($\propto 1/|\mathbf{w}|$) between the decision plane and the training examples is at maximum. This results in a constrained quadratic programming (QP) problem in search for the optimal weight $\mathbf{w}$.

After training, $\mathbf{w} = \sum_{i=1}^{p} \alpha_i y_i \mathbf{x}$, where p is the number of support vectors, α_i is the contribution from the support vector $\mathbf{x}_i$, and y_i is the training label. The output of the SVM is $u(\mathbf{x}) = \sum_{i=1}^{p} \alpha_i y_i \mathbf{x}_i^T\mathbf{x} + b$. Instead of a hard (glaucoma or not glaucoma) decision function, we convert the SVM output $u(\mathbf{x})$ into a probabilistic one, using a logistic transformation.[36]

In a more general setting, the training for classification of SVMs is accomplished by non-linear mapping of the training data to a high dimensional space, $\vec{\varphi}(\mathbf{x})$, where an optimal hyperplane can be found to minimize classification errors.[30] In this new space, the classes of interest in the pattern classification task are more easily distinguished. Although the separating hyperplane is linear in this high dimensional space induced by the non-linear mapping, the decision surface found by mapping back to the original low-dimensional input space will not be linear any more. As a result, the SVMs can be applied to data that are not linearly separable.

For good generalization performance, the SVM complexity is controlled by imposing constraints on the construction of the separating hyperplane, which results in the extraction of a fraction of the training data as support vectors. The subset of the training data that serves as support vectors thereby represents a stable characteristic of the data. As such they have a direct bearing on the optimal location of the decision surface. The hyperplane will attempt to split the positive examples from the negative examples. The system recognizes the test pattern as normal or glaucoma from the sign of the calculated output and thereby classifies the input data. After the nonlinear mapping and training, the output of SVM is given by $u(\mathbf{x}) \sum_{i=1}^{p} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b$, where $K(\mathbf{x}_i, \mathbf{x}) = \vec{\varphi}^T(\mathbf{x})\, \vec{\varphi}(\mathbf{x}_i)$ and is called the kernel function. A full mathematical account of the SVM model is described by Vapnik.[24]

Mixture of Gaussian

Mixture of Gaussian (MoG) is a special case of committee machine.[31] In committee machines, a computationally complex task is solved by dividing it into a number of computationally simple tasks.[32] For the supervised learning, the computational simplicity is achieved by distributing the learning task among a number of "experts" that divide the input space into a set of subspaces. The combination of experts makes up a committee machine. This machine fuses knowledge acquired by the experts to arrive at a decision superior to that attainable by any one expert acting alone. In the associative mixture of Gaussian model (MoG), the experts use self-organized learning (unsupervised learning) from the input data to achieve a good partitioning of the input space. Each expert does well at modeling its own subspace. The fusion of their outputs is combined with supervised learning to model the desired response.

Mixture of Generalized Gaussian

Whereas the conditional probability densities for some problems are Gaussian, in others the data may distribute with heavier tails or may even be bimodal. It would be undesirable to model these problems with Gaussian distributions. With the development of generalized Gaussian mixture model,[37] we are able to model the class conditional densities with higher flexibility, while preserving a comprehension of the statistical properties of the data in terms of means, variances, kurtosis, etc. This just-evolved approach was developed at the Salk Institute computational neurobiology laboratory. The independent component analysis mixture model can model various distributions, including uniform, Gaussian, and Laplacian. It has been demonstrated in real-data experiments that this model generally improves classification performance over the standard Gaussian mixture model.[33] The mixture of generalized Gaussians (MGG) uses the same mixture model as MoG. However, each cluster is now described by a linear combination of non-Gaussian random variables.

References

1. Chauhan BC, Drance SM, Douglas GR. The use of visual field indices in detecting changes in the visual field in glaucoma. *Invest Ophthalmol Vis Sci.* 1990;31:512–520.
2. Flammer J, Drance SM, Zulauf M. Differential light threshold. Short- and long-term fluctuation in patients with glaucoma, normal controls, and patients with suspected glaucoma. *Arch Ophthalmol.* 1984;102:704–706.
3. Wild JM, Searle AET, Dengler-Harles M, O'Neill EC. Long-term follow-up of baseline learning and fatigue effects in automated perimetry of glaucoma and ocular hypertensive patients. *Acta Ophthalmol.* 1991;69:210–216.
4. Goldbaum MH, Sample PA, White H, Weinreb RN. Discrimination of normal and glaucomatous visual fields by neural network [ARVO Abstract]. *Invest Ophthalmol Vis Sci.* 1990;31:S503. Abstract nr 2471.
5 Goldbaum MH, Sample PA, White H, et al. Interpretation of automated perimetry for glaucoma by neural network. *Invest Ophthalmol Vis Sci.* 1994;35:3362–3373.
6. Mutlukan E, Keating K. Visual field interpretation with a personal computer based neural network. *Eye.* 1994;8:321–3.
7. Brigatti L, Nouri-Mahdavi K, Weitzman M, Caprioli J. Automatic detection of glaucomatous visual field progression with neural networks. *Arch Ophthalmol*. 1997;115:725–728.
8. Brigatti L, Hoffman BA, Caprioli J. Neural networks to identify glaucoma with structural and functional measurements. *Am J Ophthalmol.* 1996;121:511–521.
9. Henson DB, Spenceley SE, Bull DR. Artificial neural network analysis of noisy visual field data in glaucoma. *Artif Intell Med.* 1997; 10:99–113.
10. Dybowski R, Weller P, Chang R, Gant V. Prediction of outcome in critically ill patients using artificial neural network synthesized by genetic algorithm. *Lancet.* 1996;347:1146–1150.
11. Baxt WG, Skora J. Prospective validation of artificial neural network trained to identify acute myocardial infarction. *Lancet.* 1996; 347:12–15.
12. Bottaci L, Drew PJ, Hartley JE, et al. Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. *Lancet.* 1997;350:469–472.
13. Bickler-Bluth M, Trick GL, Kolker AE, Cooper DG. Assessing the utility of reliability indices for automated visual fields. Testing ocular hypertensives. *Ophthalmology.* 1989;96:616–619.
14. Anderson D, Patella V. *Automated Static Perimetry.* 2nd ed. New York, NY: Mosby; 1999.
15. Flammer J, Niesel P. Die Reproduzierbarkeit perimetrischer Untersuchungsergebnisse. *Klin Monatsb Augenheil.* 1984;184:374–376.
16. Åsman P, Heijl A. Glaucoma hemifield test. Automated visual field evaluation. *Arch Ophthalmol.* 1992;110:812–819.
17. Karhunen K. Uber lineare methoden in der Wahrscheinlichkeitsrechnung. *Annales Academiae Scientiarium Fennicae, Series AI: Mathematica-Physica.* 1947;37:3–79.
18. Rumelhart DE, McClelland JL (eds.). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* Volume 1, Cambridge, MA: MIT Press;1986.
19. Rumelhart DE, Hinton G, Williams R. Learning representations of back-propagation errors. *Nature.* 1986;323:533–536.
20. Werbos PJ. *Backpropagation: Past and Future.* IEEE Int Conf on Neural Networks. New York: IEEE n 88CH2632–8; 1988:343–353.
21. Broomhead DS, Lowe D. Multivariable functional interpolation and adaptive networks. *Complex Syst.* 1988;2:321–355.
22. Samal A, Iyengar PA. Automatic recognition and analysis of human faces and facial expression: a survey. *Pattern Recognition.* 1992; 25:65–77.
23. Baird HS. Recognition technology frontiers. *Pattern Recognition Lett.* 1993;14:327–334.
24. Vapnik V. *Statistical Learning Theory.* New York: Wiley; 1998.
25. Vapnik VN. *The Nature of Statistical Learning Theory.* 2nd ed. New York: Springer; 2000.
26. Osuna E, Freund R, Girosi F. Training support vector machines: an application to face detection. Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE 97CB36082; 1997:130–136.
27. Dumais ST, Platt J, Heckerman D, Sahami M. Inductive learning algorithms and representations for text categorization. In: Gardarin G, French J, Pissinou N, Makki K, Bouganim L, eds. Proceedings of CIKM '98 7th International Conference on Information and Knowledge Management. New York; 1998:148–155.
28. LeCun Y, Jackel LD, Bottou L, et al. Learning algorithms for classification: a comparison on handwritten digit recognition. Neural Networks: Proceedings of the CTP-PBSRI Joint Workshop on Theoretical Physics. Singapore: World Scientific; 1995:261–276.
29. Mangasarian OL, Street WN, Wolberg WH. Breast cancer diagnosis and prognosis via linear programming. *Oper Res.* 1995;43:570–577.
30. Burges CJC. A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discovery.* 1998;2:121–167.
31. Haykin S. *Neural Networks: A Comprehensive Foundation.* 2nd ed. Upper Saddle River, NJ: Prentice-Hall; 1999.
32. Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. Adaptive mixtures of local experts. *Neural Comput.* 1991;3:79–87.
33. Lee T-W, Lewicki MS, Sejnowski TJ. Unsupervised classification with non-Gaussian sources and automatic context switching in blind signal separation. *IEEE Trans Pattern Analysis Machine Intell.* 2000;22:1078–1089.
34. Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK. A fast iterative nearest point algorithm for support vector machine classifier design. *IEEE Trans Neural Networks*. 2000;11:1124–1136.
35. Platt JC. Sequential minimal optimization: a fast algorithm for training support vector machines. In: Scholkopf B, Burges CJC, Smola AJ, eds. *Advances in Kernel Methods—Support Vector Learning.* Cambridge, MA: MIT Press; 1998:185–208.
36. Platt JC. Probabilistic output for Support Vector Machines and comparisons to regularized likelihood methods. In: Smola A, Bartlett P, Schölkopf B, Schuurmans E, eds. *Advances in Large Margin Classifiers.* Cambridge, MA: MIT Press; 2000:61–74.
37. Lee T.-W. Lewicki MS. The generalized Gaussian mixture model using ICA. In: Pajunen P, Karhunen J, eds. Proceedings of the Second International Workshop on Independent Component Analysis and Blind Signal Separation (ICA '00). Helsinki, Finland; 2000: 239–244.
38. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem.* 1993;39:561–577.
39. DeLong ER, DeLong DM, Clark-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1998;44:837–845.
40. Hanley JA, McNeil BJ. Method for comparing the area under two ROC curves derived from the same cases. *Radiology.* 1983;148: 839–843.
41. Hanley JA. Receiver operating characteristic (ROC) methodology: the state of the art. *Crit Rev Diag Imaging.* 1989;29:307–335.
42. McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Medical Decision Making.* 1984;4:137–150.
43. Stone M. Cross-validatory choice and assessment of statistical predictions. *J Roy Statist Soc Ser.* 1974;36:111–147.
44. Leitman T, Eng J, Katz J, Quigley HA. Neural networks for visual field analysis: how do they compare with other algorithms? *Glaucoma.* 1999;8:77–80.
45. Spenceley SE, Henson DB, Bull DR. Visual field analysis using artificial neural networks. *Ophthal Physiol Opt.* 1994;14:239–248.
46. Johnson CA, Adams AJ, Casson EJ, Brandt JD. Blue-on-yellow perimetry can predict the development of glaucomatous visual field loss. *Arch Ophthalmol.* 1993;111:645–650.
47. Sample PA, Weinreb RN. Color perimetry for assessment of primary open-angle glaucoma. *Invest Ophthalmol Vis Sci.* 1990;31: 1869–1875.
48. Johnson C, Samuels SJ. Screening for glaucomatous visual field loss with frequency-doubling perimetry. *Invest Ophthalmol Vis Sci.* 1997;28:413–425.
49. Maddess T, Henry GH. Performance of nonlinear visual units in ocular hypertension and glaucoma. *Clin Vis Sci.* 1992;7:371–383.