Bayesian Machine Learning Classifiers for Combining Structural and Functional Measurements to Classify Healthy and Glaucomatous Eyes

Christopher Bowd, Jiucang Hao, Itan M. Tavares, Felipe A. Medeiros, Linda M. Zangwill, Te-Won Lee, Pamela A. Sample, Robert N. Weinreb, and Michael H. Goldbaum

PURPOSE. To determine whether combining structural (optical coherence tomography, OCT) and functional (standard automated perimetry, SAP) measurements as input for machine learning classifiers (MLCs; relevance vector machine, RVM; and subspace mixture of Gaussians, SSMoG) improves diagnostic accuracy for detecting glaucomatous eyes compared with using each measurement method alone.

METHODS. Sixty-nine eyes of 69 healthy control subjects (average age, 62.0, SD 9.7 years; visual field mean deviation [MD], −0.70, SD 1.41 dB) and 156 eyes of 156 patients with glaucoma (average age, 66.4, SD 10.2 years; visual field MD, −3.12, SD 3.43 dB) were imaged with OCT (Stratus OCT, Carl Zeiss Meditec Inc., Dublin, CA) and SAP (Humphrey Field Analyzer II with Swedish Interactive Thresholding Algorithm, SITA; Carl Zeiss Meditec Inc.) within 3 months of each other. RVM and SSMoG MLCs were trained and tested on OCT-determined RNFL thickness measurements from 32 sectors (−11.25° each) obtained in the circumpapillary area under the instrument-defined measurement ellipse and SAP pattern deviation values from 52 points from the 24-2 grid, independently and in combination. Tenfold cross-validation was used to train and test classifiers on unique subsets of the full 225-eye data set, and areas under the receiver operating characteristic curve (AUROC) for the classification of eyes in the test set were generated. AUROC results from classifiers trained on OCT and SAP alone and those trained on OCT and SAP in combination were compared. In addition, these results were compared to currently available OCT measurements (mean retinal nerve fiber layer [RNFL] thickness, inferior RNFL thickness, and superotemporal RNFL thickness) and SAP indices (MD and pattern standard deviation [PSD]).

RESULTS. The AUROCs for RVM trained on OCT parameters alone, SAP parameters alone and OCT and SAP parameters combined were 0.809, 0.815, and 0.845, respectively. The AUROCs for SSMoG trained on OCT parameters alone, SAP parameters alone, and OCT and SAP parameters combined were 0.817, 0.841, and 0.869, respectively. Combining techniques using both RVM and SSMoG significantly improved on MLC analysis of OCT, but not SAP, measurements alone. Classification performance using RVM and SSMoG was statistically similar.

CONCLUSIONS. RVM and SSMoG Bayesian MLCs trained on OCT and SAP data can successfully discriminate between healthy and early glaucomatous eyes. Combining OCT and SAP measurements using RVM and SSMoG increased diagnostic performance marginally compared with MLC analysis of data obtained using each technology alone. (Invest Ophthalmol Vis Sci. 2008;49:945–953) DOI:10.1167/iovs.07-1083

Since 1990 (Goldbaum MH, et al. IOVS 1990;31:ARVO Abstract 503), machine learning classifier (MLC) techniques have been applied to optical imaging and visual function measurements to improve glaucoma detection, with results suggesting that these techniques are as good as or better than currently available analysis techniques at classifying eyes as glaucomatous or healthy.1–4 Bowd et al.5 recently showed, using a relatively new supervised (i.e., trained on labeled data, in this case, glaucoma versus healthy) Bayesian MLC called the relevance vector machine (RVM),6,7 that RVM analysis of scanning laser polarimetry (SLP) data is able to classify eyes more successfully than are standard retinal nerve fiber layer (RNFL) thickness measurements such as superior and average RNFL thicknesses. In addition, RVM analysis performed as well as support vector machine (SVM)6,9 analyses at this task. Unlike SVM, RVM incorporates probabilistic output (probability of class membership, e.g., probability of glaucoma defined as membership in the glaucoma training set) through Bayesian inference. Its decision function depends on fewer input variables than SVM, possibly allowing better classification estimates for small data sets with high dimensionality (i.e., a large number of input variables).7 Probabilistic output allows an intuitive interpretation of classifier output and therefore is likely to be more informative for clinical use than is output converted to a nonlinear scale by using postprocessing, similar to SVM.

Findings in some recent studies have suggested that RNFL measurements obtained by optical coherence tomography (OCT) are more sensitive to glaucomatous damage than are SLP measurements.10 In addition, existing studies have shown that structural and functional techniques for detecting glaucoma often identify different glaucoma patients when glaucoma severity is not too advanced.10,11 and that combining structural
and functional techniques can improve glaucoma detection. Based on these results, we chose to assess the ability of RVM trained and tested on OCT RNFL thickness measurements to classify glaucomatous and healthy eyes. In addition, we combined OCT RNFL thickness information with visual sensitivity measurements measured using standard automated perimetry (SAP) to determine whether combining structural and functional information using MLCs improves glaucoma detection. The purpose of this study was first to determine whether combining reasonably easy-to-obtain measurements from glaucomatous eyes, by using complex MLC techniques, would improve diagnostic performance. If performance was improved, these results might prompt the automated combination and analysis of information from structural and functional diagnostic techniques to improve glaucoma detection in the clinic. Second, successful classification of eyes using RVM trained on OCT data would provide more support for MLC use in glaucoma diagnosis with optical imaging. Finally, we introduced a new Bayesian MLC, the subspace mixture of Gaussians (SSMoG), and determined its performance on the same data set.

## METHODS

### Subjects

This observational cross-sectional study included one randomly selected eye from each of 225 study participants (156 patients with glaucoma and 69 healthy control subjects) older than 40 years and enrolled in the University of California, San Diego, Diagnostische Innovationen in Glaucoma Study (DIGS). Each study participant underwent a comprehensive ophthalmic evaluation, including review of medical history, best corrected visual acuity, slit lamp biomicroscopy, intraocular pressure measurement with Goldmann applanation tonometry, gonioscopy, dilated slit lamp fundus examination with a 78-D lens, simultaneous stereoscopic optic disc photography (TRC-88; Topcon Instruments Corp. of America, Paramus, NJ), and SAP using the 24-2 Swedish Interactive Threshold Algorithm (SITA; Humphrey Field Analyzer II, Carl Zeiss Meditec, Inc., Dublin, CA). To be included in the study, participants had to have a best-corrected acuity better than or equal to 20/40, spherical refraction within ±5.0 D, cylinder correction within ±3.0 D, and open angles on gonioscopy. Eyes with coexisting retinal disease, uveitis, or nonglaucomatous optic neuropathy were excluded.

For labeling eyes during classifier training, glaucomatous eyes were defined as either eyes with glaucomatous appearance of the optic disc (i.e., glaucomatous optic neuropathy [GON]) based on masked assessment (described later) of simultaneous stereoscopic optic disc photographs and defined as the presence of neuroretinal rim thinning and/or diffuse or localized RNFL defects indicative of glaucoma or eyes with repeatable (two consecutive) SAP results outside normal limits by pattern standard deviation (PSD; P < 0.5%) or Glaucoma Hemifield Test. For GON criterion, photographs were evaluated by two trained observers who were masked to the date of the photographs and the identity and diagnosis of the study participants. Discrepancies between observers were resolved by adjudication by a third trained observer.

GON eyes with normal visual fields were included in the glaucoma group to assure the representation of early glaucoma in the sample. In addition, ROC curves for discriminating between healthy and glaucomatous eyes defined by field defects when using OCT often are so high that they do not allow improvement (in this case, achievable by the addition of visual field information).

For glaucomatous eyes classified based on visual fields, the first abnormal SAP was on or before the imaging date. Average SAP mean deviation (MD) of the glaucomatous eyes was −3.12 dB (SD 3.45; range, −20.14 to +1.15). Seventy-five percent of the glaucomatous eyes had MD ≥ −4.08 dB, and 90% had MD ≥ −6.57, indicating primarily early glaucoma. The mean age of the patients with glaucoma at the time of VF testing was 64.4 years (SD 10.2; range, 41.3–87.0 years). Intraocular pressure was not part of the inclusion criteria for the glaucoma group.

Healthy eyes were defined as of healthy volunteers with normal-appearing optic discs by examination and no history of intraocular pressure >22 mm Hg. Average SAP MD of the healthy eyes was −0.70 dB (SD 1.41; range, −5.34 to +1.80) and was significantly different from that of glaucomatous eyes (t test; P < 0.0001). The mean age of the healthy participants was 62.0 years (SD 9.7; range, 42.3–85.6) and was significantly younger than that of the patients with glaucoma (t test, P = 0.005); however, the age range was similar between groups.

Healthy eyes were included based on examination results and IOP, not on optic disc appearance by detailed, consensus stereoscopic photographs or visual field results, because we did not want to bias results in favor of either optical imaging (i.e., OCT measurements) or visual function (SAP) measurements (although it is arguable that requiring a healthy examination could increase somewhat the specificity of OCT measurements). In addition, this sample is more representative of a screened population than is a population requiring healthy-appearing optic discs by very detailed photograph assessment and completely normal visual field test results.

This research adhered to the tenets of the Declaration of Helsinki and Health Insurance Portability and Accountability Act guidelines. The University of California, San Diego, Human Research Protection Program approved all methodology.

### Optical Coherence Tomography

Study participants underwent ocular imaging with the commercially available StratusOCT (software ver. 4.0; Carl Zeiss Meditec, Inc., Inc.). The StratusOCT measures RNFL thickness by using a low-coherence light source projected onto the retina. This measurement beam’s reflected light source is compared to the reflectance of a reference beam reflected from a reference mirror at a known position, to determine the thickness of the retina. A proprietary edge-detection algorithm is used to define the posterior border of the RNFL (the anterior border is defined by the large difference in reflectance along the vitreoretinal interface). Details of this technique have been described. The ‘fast’ RNFL scanning option was used, and measurements from the three individual RNFL scans were combined automatically by OCT software to create the mean image used for analysis. Only well-centered scans with a signal strength ≥7 were included. In addition, all scans were subjectively evaluated to confirm that the RNFL detection algorithm followed closely the apparent anterior and posterior RNFL borders.

RNFL thickness measurements were obtained for 32 circumpapillary sectors of approximately 11.25° each, by using the standard 3.4-mm scan diameter. The 32 sectors were constructed by averaging the available 256 A-scan measurements provided by StratusOCT export software by groups of eight. Thirty-two sectors were chosen based on results from a previous study. Sector 1 was located temporally, with sectors 8 and 24, located superiorly and inferiorly, respectively (i.e., results were normalized to a right eye).

### Standard Automated Perimetry

SAP threshold measurements, total deviation (TD) measurements and PSD measurements at each of 52 test points (excluding those representing the blind spot) were obtained. The visual field tests that first defined abnormality was never the test used for MLC analysis. All visual fields used to classify (i.e., label) eyes and to train the RVMs were reliable, with false positives and fixation losses ≤25%.

### Relevance Vector Machine

The RVM learning classifier is a Bayesian model that provides probabilistic predictions (e.g., probability of glaucoma based on the training examples) through Bayesian inference. Its decision function depends on fewer sample vectors (i.e., is more sparse) than previously reported SVM, because SVM minimizes the training error under the constraint of
maximum smoothness, requiring more decision points.\textsuperscript{6} The benefit of a sparser classifier is that its results are more generalizable (i.e., it decreases overfitting). In classification, RVM outputs probabilities of class membership in contrast to SVM, which expresses a nonprobabilistic value that is then related to the decision threshold. RVMs point probabilistic output provides a conditional distribution that allows a more intuitive expression of uncertainty in the prediction.\textsuperscript{17}

The RVM was implemented using a commercial software algorithm (SparseBayes ver. 1.0 algorithm; Microsoft Research, Cambridge, UK; for Matlab; The MathWorks, Natick, MA). The kernel width was optimized by a grid search with width $= \sqrt{(2 \cdot a \cdot n)}$ of input variables, where $a$ is 0.5, 0.8, 1.0, 1.2, 1.4. The $a$ that offered the largest AUROC was chosen.

**Subspace Mixture of Gaussians**

SSMoG is a “committee” or “mixture of experts” MLC that, like RVM, provides a Bayesian-based probabilistic output based on labeled input. Committee machines combine individual estimators or “experts” (in this case Gaussian process regression estimators), each modeling different regions of the input space (in this case, aspects of OCT, SAP, and combined measurements), with the goal of improving generalization performance.\textsuperscript{18} Modeling with MoG fits the different regions of input space with Gaussian surfaces. SSMoG is a dimension-reducing MoG method that allows the use of MoG classifiers on high dimension data by iteratively sampling with replacement multiple random subsets of the full-dimensional data set, thereby eliminating the need for pre-processing dimension reduction. In this study, each feature subset of the full dataset was composed of eight features. The sampling and training procedure was repeated 100 times, with 10-fold cross-validation. The final output of SSMoG was accomplished by averaging the output of the 100 MoG classifiers. To “optimize” SSMoG, we defined the final output as the combination of the best (defined as largest AUROC) 20 MoGs by averaging. SSMoG was implemented (i.e., coded) with commercial software (Matlab; The MathWorks). For each MoG, the number of mixtures ranged from one to four, and the best number was chosen.

**Analyses**

First, the classifiers were trained and tested on both the 32-dimension (52 sector) OCT data and 52- or 53-dimension SAP (threshold, TD, PSD) data (52 SAP inputs, 53 when thresholds were considered because participant age also was included), independently. Next, they were trained and tested on OCT data in conjunction with each type of SAP data individually, resulting in 84 (TD, PSD) or 85 (threshold) input dimensions (e.g., 52 OCT inputs + SAP inputs). We combined OCT data with several types of SAP data because of recent evidence that thresholds measurements (as we have used previously) are not necessarily optimal input for neural network analyses.\textsuperscript{19}

OCT results were compared with SAP results and results obtained when OCT and SAP data were combined. These results also were compared to OCT mean RNFL thickness, inferior RNFL thickness and superior RNFL thickness, and SAP MD and PSD, to compare RVM performance to the performance of currently available parameters from OCT and SAP.

Areas under the ROC curves (AUROCs) for classifying eyes as healthy or glaucomatous were determined for each classifier and each standard parameter from OCT and SAP. Significant differences in AUROCs were determined using the method of De Long et al.\textsuperscript{20}

**Training and Testing MLCs**

Tenfold cross-validation was used to train and test RVM and SSMoG classifiers, to avoid training and testing on the same data. First, glaucomatous and healthy eyes were randomly divided into 10 approximately equal, exhaustive, and mutually exclusive subsets. Next, classifiers trained on 9 subsets combined were subsequently tested on the 10th subset. This process was repeated 10 times, with each subset serving as the test set one time, so that each tested eye was never part of its training set and was tested only once. The test results from all eyes were then used to plot the bias-corrected ROC curve. Sensitivities at 75% and 90% specificities, arbitrarily chosen to represent moderate and high specificity, respectively, were also reported, although these values can be estimated from visual inspection of ROC curves.

**Classifier Optimization**

Because the dimensionality of the data sets (number of features) is relatively large but the size of the data sets (number of samples) is relatively small, there is some uncertainty about the location of the separating surface. To alleviate the “curse of dimensionality” (reduced classifier performance caused by the forced inclusion of irrelevant parameters in the solution set),\textsuperscript{17} we used backward elimination, which we found to be a slight improvement over forward selection for optimization,\textsuperscript{1,2,5} to reduce the data dimension for all RVM analyses. We started with the full-dimensional feature set (e.g., 32 RNFL thickness measurements + 52 SAP PSD measurements) and sequentially deleted the features that least improved the performance of the feature set until performance began to decline (i.e., number of features included at the peak of an AUROC [P] versus number of features [x] plot, called “peaking”).

Similar to the application of 10-fold cross-validation to minimize bias in the testing and training of the full-dimension RVM and SSMoG, we used cross-validation to minimize bias in the test sets used during the process of optimizing the RVM feature sets. The data were randomly divided into five approximately equal-sized subsets. Four of the five subsets were used as feature-selection sets to determine the optimized feature-set, selected based on the maximum AUROC (by “peaking”), using 10-fold cross-validation. The optimized feature set was then trained on the initial four subsets, resulting in a single classifier. Next, this classifier was tested on the remaining single subset, and a bias-corrected AUROC was generated. This sequence was repeated five times, with each partition serving as the test set one time, resulting in five unbiased estimates of the AUROC. This technique for optimization is described in greater detail elsewhere.\textsuperscript{21}

Several optimized feature sets were investigated for RVM (and SSMoG) using each data set tested, and we report the results from the optimized feature set for each data set that resulted in the largest AUROC only, although differences in AUROC were not always significant among sets.

**Results**

OCT RNFL thickness measurements, SAP global indices (MD, PSD) and outputs from all MLCs differed significantly between healthy and glaucomatous eyes (all $P \leq 0.0001$; Table 1).

**Classification**

**Relevance Vector Machine.** Because AUROCs for optimized RVMs tend to be slightly (if not always significantly) higher than nonoptimized AUROCs,\textsuperscript{1,2,5} nonoptimized results are shown in Table 2 only, whereas optimized results are reported both in the table and the text. The AUROC for optimized RVM using OCT data alone (16 features in peak feature set) was 0.806. This value was similar ($P \geq 0.05$) to AUROCs for OCT mean thickness (0.777), OCT inferior thickness (0.770), SAP MD (0.775), and SAP PSD (0.772). AUROCs for optimized RVM using SAP data alone were 0.809 for thresholds (20 features), 0.815 for TDS (12 features) and 0.802 for PDs (10 features). No significant differences were found among any of these values and OCT and SAP standard indices, except for OCT superior thickness (AUROC = 0.680). The AUROC for OCT superior thickness was significantly smaller than AUROCs for all optimized RVM results (all $P \leq 0.007$).

Combining OCT and SAP data (OCT+Thr, 24 features; AUROC = 0.845) for analysis with optimized classifiers signif-
significantly increased the AUROC compared with OCT-measured mean and inferior RNFL thicknesses, SAP MD, and RVM when, using OCT data alone (all comparisons, \( P < 0.05 \)). Combining OCT with other SAP results (OCT+SAP TD, AUROC = 0.789; and OCT+SAP PD, AUROC = 0.789) resulted in no differences compared with noncombined OCT and SAP RVM results. AUROCs for the best performing OCT-based RVM, SAP-based RVM and combination SAP/OCT based RVM are shown in Figure 1.

Sensitivities at fixed specificities of 75% and 90% are shown in Table 2 and are generally higher when OCT and SAP data are combined.

**Subspace Mixture of Gaussians.** SSMoG tended to perform better on SAP data alone, and RVM tended to perform better on OCT data alone (Table 2). SS MoG optimized, noncombined AUROCs for OCT and SAP ranged from 0.801 (SAP PD) to 0.841 (SAP TD) with all comparisons among these results \( P > 0.05 \). Combining OCT and SAP measurements and optimizing results increased AUROCs to as high as 0.869 (OCT combined with SAP TD). This value was significantly greater than the optimized AUROC for OCT (0.817, \( P < 0.05 \)), but not greater than the “best” optimized AUROC for SAP (i.e., SAP TD). AUROCs for the best-performing OCT-based SS MoG, SAP-based SS MoG, and combination SAP/OCT-based SS MoG are shown in Figure 2.

Sensitivities at fixed specificities of 75% and 90% are also shown in Table 2 and, similar to RVM results, are generally higher when OCT and SAP data are combined.

**Probabilistic Results**

Both RVM and SS MoG outputs are in the form of a probability of belonging to the glaucoma group of the training set. For the nominally best-performing (by largest AUROC area) RVM combination of OCT and SAP measurements (i.e., optimized OCT RNFL thickness measurements combined with SAP thresholds), average mean output was 0.461 (SD 0.230; range, 0.048–0.953; 90% CI = 0.407–0.512) for healthy eyes and 0.798 (SD 0.234; range, 0.200–1.00; 90% CI = 0.761–0.854) for glaucomatous eyes (\( P < 0.001 \)). Figure 3 is a histogram that shows the number of healthy and glaucomatous eyes that fell into each 10% probability bin based on RVM output. Seventy-two percent of healthy eyes (sensitivity = 0.72), and 81% of glaucomatous eyes were assigned a probability of glaucoma \( \geq 50\% \) (specificity = 0.72), and 81% of glaucomatous eyes were assigned a probability of glaucoma \( > 50\% \) (sensitivity = 0.81).

For the combined SS MoG with the largest AUROC (optimized OCT RNFL thickness measurements combined with SAP TD measurements), average mean output was 0.415 (SD 0.157; range, 0.188–0.747; 90% CI = 0.377–0.415) for healthy eyes and 0.747 (SD 0.224; range, 0.188–1.00; 90% CI = 0.711–0.782) for glaucomatous eyes (\( P < 0.001 \)). Similar to Figure 3, Figure 4 shows the number of healthy and glaucomatous eyes that fell into each 10% probability bin based on SS MoG output. Sixty-three percent of healthy eyes were assigned a probability of glaucoma \( \geq 50\% \) (specificity = 0.65) and 82% of glaucomatous eyes were assigned a probability of glaucoma \( > 50\% \) (sensitivity = 0.82). Both Figures 2 and 3 show considerable overlap in output between the healthy and glaucoma groups.

### Table 1. Comparison of OCT and SAP Indices and MLC (RVM, SS MoG) Output between Healthy and Glaucomatous Eyes

<table>
<thead>
<tr>
<th>Technique</th>
<th>Healthy Eyes, SD</th>
<th>Glaucomatous Eyes, SD</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>OCT mean RNFL thickness (( \mu m ))</td>
<td>98.1, 9.8</td>
<td>83.1, 15.6</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT inferior RNFL thickness (( \mu m ))</td>
<td>130.7, 25.9</td>
<td>102.7, 29.3</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT superior RNFL thickness (( \mu m ))</td>
<td>126.8, 22.0</td>
<td>108.9, 28.4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SAP MD (dB)</td>
<td>-0.70, 1.41</td>
<td>-3.12, 3.43</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SAP PSD (dB)</td>
<td>1.73, 0.75</td>
<td>3.65, 5.19</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT RVM full (32 features)</td>
<td>0.532, 0.194</td>
<td>0.770, 0.212</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SAP Thr RVM full (53 features)</td>
<td>0.541, 0.204</td>
<td>0.777, 0.240</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT TD RVM full (52 features)</td>
<td>0.513, 0.226</td>
<td>0.770, 0.241</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT PD RVM full (52 features)</td>
<td>0.530, 0.251</td>
<td>0.750, 0.237</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT RVM optimized (16 features)</td>
<td>0.513, 0.190</td>
<td>0.768, 0.219</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SAP Thr RVM optimized (20 features)</td>
<td>0.498, 0.226</td>
<td>0.779, 0.237</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SAP TD RVM optimized (12 features)</td>
<td>0.481, 0.235</td>
<td>0.785, 0.243</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SAP PD RVM optimized (10 features)</td>
<td>0.484, 0.287</td>
<td>0.796, 0.244</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT SS MoG</td>
<td>0.509, 0.163</td>
<td>0.744, 0.232</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SAP Thr SS MoG</td>
<td>0.381, 0.182</td>
<td>0.654, 0.242</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SAP TD SS MoG</td>
<td>0.374, 0.179</td>
<td>0.642, 0.247</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SAP PD SS MoG</td>
<td>0.273, 0.193</td>
<td>0.523, 0.281</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT SS MoG optimized</td>
<td>0.475, 0.195</td>
<td>0.758, 0.238</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SAP Thr SS MoG optimized</td>
<td>0.397, 0.196</td>
<td>0.714, 0.244</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SAP TD SS MoG optimized</td>
<td>0.332, 0.187</td>
<td>0.660, 0.262</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SAP PD SS MoG optimized</td>
<td>0.235, 0.202</td>
<td>0.536, 0.304</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT+Thr RVM full (85 features)</td>
<td>0.483, 0.218</td>
<td>0.788, 0.225</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT+TD RVM full (84 features)</td>
<td>0.490, 0.225</td>
<td>0.792, 0.240</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT+PD RVM full (84 features)</td>
<td>0.523, 0.216</td>
<td>0.773, 0.258</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT+Thr RVM optimized (24 features)</td>
<td>0.461, 0.227</td>
<td>0.798, 0.254</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT+TD RVM optimized (29 features)</td>
<td>0.465, 0.230</td>
<td>0.787, 0.241</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT+PD RVM optimized (24 features)</td>
<td>0.479, 0.227</td>
<td>0.786, 0.227</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT+Thr SS MoG</td>
<td>0.349, 0.147</td>
<td>0.692, 0.256</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT+TD SS MoG</td>
<td>0.411, 0.149</td>
<td>0.670, 0.250</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT+PD SS MoG</td>
<td>0.298, 0.139</td>
<td>0.583, 0.254</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT+Thr SS MoG optimized</td>
<td>0.390, 0.162</td>
<td>0.738, 0.242</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT+TD SS MoG optimized</td>
<td>0.415, 0.157</td>
<td>0.747, 0.224</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>OCT+PD SS MoG optimized</td>
<td>0.310, 0.147</td>
<td>0.647, 0.208</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

MLC output indicates assigned probability of glaucoma. Thr, threshold.
Although it is intuitive to expect that an RVM/SSMoG disease/no-disease cutoff of 0.50 is ideal, this is not necessarily the case, because the criterion for a meaningful cutoff is dependent on the desired sensitivity and specificity pairing (diagnostic utility). When we arbitrarily set cutoffs based on a 0.90 specificity (best OCT/SAP RVM probability cutoff 0.801; best OCT/SAP SSMoG probability cutoff 0.642) sensitivities of RVM and SSMoG were 0.63 and 0.68, respectively (see Table 2). Although sensitivities decreased as specificity was increased to 0.90 (see values above), diagnostic utility increased with positive likelihood ratios (LR; sensitivity/1-specificity, LR) increasing from 2.90 to 6.28 (i.e., from a small to a moderate effect on posttest probability of disease) for RVM and from 2.20 to 6.80 for SSMoG. Using this 0.90 specificity-based cutoff, agreement for healthy/glaucoma classification was observed between RVM and SSMoG in 87% of eyes tested (0.742; 95% CI 0.655–0.830), indicating good classification agreement between techniques. Of the 28 (13%) eyes not classified the same by both techniques, 14 were assigned probabilities within 0.11 of each other. Table 3 shows LR values for several different, arbitrarily chosen specificities and therefore more completely illustrates the effect of varying disease/no-disease cutoffs.

**DISCUSSION**

In the present study, combining OCT and SAP measurements using MLCs resulted in a trend of increased AUROCs for discriminating between healthy and glaucomatous eyes, compared to measurements from either modality alone. The optimized RVM and SSMoG classifiers demonstrated superior performance in distinguishing between healthy and glaucomatous eyes, with AUROCs ranging from 0.777 to 0.848 for RVM and 0.787 to 0.869 for SSMoG. The classification agreement between techniques was high, with 87% of eyes classified correctly.

**FIGURE 1.** ROC curves for best optimized RVM analysis of OCT data, best optimized RVM analysis of SAP data, and best optimized RVM analysis of combined OCT and SAP data. AUROCs are shown in parentheses.
pared to using each measurement technique alone. Optimization of the classifiers permitted significant improvements of RVM for OCT/H11001 SAP thresholds, and of SSMoG for OCT/H11001 SAP/TD, over the best MLC results of OCT data alone. Accuracy of MLC analyses of OCT and SAP parameters improved on standard (currently available) parameters for each technique. These results agree with those of several studies of OCT,23 SAP,2 CSLO,1 and SLP5 and indicate that RVM and the relatively new SSMoG analyses show promise for diagnostic classification of complex imaging and visual field data. In addition, the less computationally demanding SSMoG analysis performed similarly to RVM for classification. Both techniques provide intuitive probability of group membership as output, which is more desirable than results generated using SVMs and multilayer perceptrons, for instance.

Outputs from RVM and SSMoG indicated that most of the glaucomatous eyes were assigned a high probability of belonging to the glaucoma training group (the distribution was negatively skewed; Figs. 3, 4). The distributions of probabilities for healthy eyes, on the other hand, were less skewed, probably a result of the often-reported wide range of structural and functional measurements in healthy eyes and thus not surprising. We also demonstrated the fact that $P > 0.50$ is not optimal for classifying eyes as glaucomatous when high specificity and informative LRs are desired.

Based on the results of other studies, we were somewhat surprised that our results indicated only modest improvements when structural and functional measures were combined. First, the Ocular Hypertension Treatment Study24 and the European Glaucoma Prevention Study25 recently suggested that a similar percentage of suspect eyes develop glaucoma when first conversion is defined based on optic disc abnormality or on visual field abnormality, suggesting to us that OCT and SAP would detect abnormalities in a significant percentage of different eyes with early-stage glaucoma. Next, other studies have suggested that combining imaging and visual field data results in improved diagnostic accuracy for glaucoma detection.12–14 Our results indicate that there is some dependency between the structural and functional measurements. Nevertheless, the significant improvement in performance of optimized MLCs trained on combined data compared to single tests alone demonstrates the benefit of combining the tests.

An early study by Caprioli12 combined optical imaging data (obtained by a Rodenstock Instruments [Munich, Germany] computerized image analysis system) and SAP data (obtained by Humphrey [Carl Zeiss Meditec, Inc.] or Octopus [Haag Streit, Köniz, Switzerland] perimetry) from 54 healthy and 185 glaucomatous eyes in a linear discriminant analysis model.
Classification performance of the combined model (sensitivity = 0.90, specificity = 0.76, and accuracy = 0.87) was better than models incorporating only structural indices (sensitivity = 0.88, specificity = 0.35, and accuracy = 0.76) and only functional indices (sensitivity = 0.99, specificity = 0.06, and accuracy = 0.77). Using the same data set, Brigatti et al.26 combined structural and functional measurements using multilayer perceptron neural networks and demonstrated similar results. Performance of the combined model (sensitivity = 0.90, specificity = 0.84, and accuracy = 0.88) was better than models incorporating only structural indices (sensitivity = 0.87, specificity = 0.56, and accuracy = 0.80) and only functional indices (sensitivity = 0.84, specificity = 0.86, and accuracy = 0.84).

More recently, a study from our group indicated that combining imaging (current generation OCT, CSLO, and SLP) and visual field measurements (frequency doubling technology perimeter, FDT) increased sensitivity for detecting eyes with glaucoma compared with using either technique in isolation, with no significant decrease in specificity. In this study of 101 eyes, adding FDT pattern SD measurements increased sensitivity as much 21% compared with using imaging alone.

In a study using MLCs to combine CSLO and perimetry results, Mardin et al.13 combined CSLO-measured rim volume, rim area, and cup shape measurements (global and sectoral) with Octopus perimeter–measured MD; corrected loss variance (global and sectoral) by using several classification techniques (MLC and statistical); and compared the diagnostic accuracy of these techniques to the same techniques using CSLO and perimetry results alone in 88 healthy and 88 glaucomatous (average Octopus MD = 7.13 dB) eyes. The largest AUROC reported was 0.976 for combined CSLO and perimetry measurements with a “double-bagging” technique (optimal parameters are selected by using linear discriminant analysis and included in a bootstrap aggregate, or “bagging,” regression tree). This result was an improvement over double-bagging of CSLO measurements alone (AUROC = 0.855), but not over double-bagging of perimetry measurements alone (AUROC = 0.98). Results using other classifiers were similar. These results were probably due in part to inclusion criteria that required normal visual fields in healthy eyes and abnormal visual fields in glaucomatous eyes.

Finally, in a very recent study from our laboratory,27 combining CSLO (Heidelberg Retina Tomograph, HRT II) and short-wavelength automated perimetry (SWAP) by using RVM resulted in improved accuracy compared with RVMs using each test alone for classifying healthy eyes (n = 68) and those with glaucoma (n = 144, average SAP MD = −5.1 dB) defined based on the same criteria used in the present study. In this study, AUROCs for the best-performing optimized RVM analyses of CSLO parameters, SWAP parameters and combined CSLO and

### Table 3. Sensitivities and Positive Likelihood Ratios

<table>
<thead>
<tr>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Cutoff</th>
<th>RVM Positive Likelihood Ratio</th>
<th>Sensitivity</th>
<th>Cutoff</th>
<th>SSMoG Positive Likelihood Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.75</td>
<td>0.78</td>
<td>0.631</td>
<td>3.10 (2.04, 4.71)</td>
<td>0.80</td>
<td>0.517</td>
<td>3.21 (2.11, 4.86)</td>
</tr>
<tr>
<td>0.85</td>
<td>0.70</td>
<td>0.748</td>
<td>4.66 (2.63, 8.24)</td>
<td>0.75</td>
<td>0.578</td>
<td>5.0 (2.85, 8.83)</td>
</tr>
<tr>
<td>0.90</td>
<td>0.63</td>
<td>0.801</td>
<td>6.28 (3.06, 13.00)</td>
<td>0.68</td>
<td>0.642</td>
<td>6.80 (3.32, 14.00)</td>
</tr>
<tr>
<td>0.96</td>
<td>0.56</td>
<td>0.859</td>
<td>14.00 (4.41, 45.00)</td>
<td>0.65</td>
<td>0.677</td>
<td>16.00 (5.12, 52.00)</td>
</tr>
</tbody>
</table>

Sensitivities and positive likelihood ratios (LR+[95% CI]) at various specificities (arbitrarily selected) for “best” RVM (OCT+SAP threshold optimized) and SSMoG (OCT+SAP TD optimized) models combining OCT and SAP data.
SWAP parameters were 0.878, 0.780, and 0.925, respectively. These values tended to be higher than those reported in the present study.

Several other studies have investigated MLC analysis of OCT data. The first such study demonstrated that multilayer perceptron classifiers trained on 25 available OCT RNFL and optic disc parameters were able to discriminate between healthy eyes (n = 100) and those with early glaucoma (n = 89, average SAP MD = −2.7 dB) with an AUROC of 0.821 (sensitivity = 0.51 at set 0.90 specificity). This value was increased to 0.874 (sensitivity = 0.67 at set 0.90 specificity) when principal components analysis was used to reduce the input data to three parameters (based on AUROC [x] by number of features [y] plot peaking). In another study of 71 healthy and 64 glaucomatous eyes (average SAP MD = −4.3) by the same group, self-organizing maps and decision trees were used to establish rules for classifying eyes as healthy or glaucomatous based on 25 available OCT RNFL and optic disc measurements. A self-organizing map identified the two (of five) most informative parameter clusters (i.e., features; defined based on the largest mapped distance between healthy and glaucomatous eyes), and the three top-ranked parameters from each cluster were included in a decision tree that, after “pruning,” resulted in a three-branch tree that yielded 0.73 sensitivity, 0.92 specificity, and 0.83 accuracy. In a third study, the ability of several statistical and MLCs were investigated for discriminating between 42 healthy and 47 glaucomatous eyes (glaucomatous eyes were approximately one-half early glaucoma with SAP MD = −6 dB, average MD = −2.18; and approximately one-half advanced glaucoma with MD < −6 dB, average MD = −10.9 dB). An SVM trained on the eight (of 38) OCT parameters with highest correlation with visual fields most accurately classified eyes with AUROC of 0.981 and sensitivity of 0.92 at a fixed specificity of 0.95. Although the results of these studies are not directly comparable to those of our study because of differences in classifiers used, patient demographics, and glaucoma severity, they further stress the usefulness of MLCs for glaucoma diagnosis.

In the present study, we used backward elimination to decrease (i.e., optimize) the dimensionality of RVM data sets. This technique identifies individual parameters that most strongly contribute to the derived classification model. In previous similar studies we reported the specific parameters included in reduced-feature optimized data sets as a means to describe important parameters and regions of interest in the peripapillary retina and visual fields. This information has been omitted from the present study because selecting these regions can be somewhat misleading because of strong correlations (i.e., dependence) among adjacent peripapillary and visual field locations that can result in one important location being excluded by the inclusion of another related, equally important location. In addition, it is possible that in more advanced glaucoma, or when using a larger training set, different sectors would be identified as most important for the classification task.

AUROC results from our MLC techniques may be somewhat overestimated because we used cross-validation instead of truly independent training and test sets. Although RVM and SSMoG were trained and tested on different data, each data set was generated from the same rather homogeneous pool. This fact may exaggerate somewhat the differences in classification ability between MLCs and standard OCT and SAP parameters, although we were careful to employ separate training and test sets at all steps in training, testing, and optimizing (when applicable) our MLCs. In addition, establishing classifier performance, per se, was not the main purpose of this study; rather, we sought to determine whether combining structural and functional measurements in a novel way could improve classification accuracy and therefore relative accuracy was most important.

It is also likely that our reference standard criteria affected AUROC results. It is clear based on the range of SAP MD present in the healthy eyes, that some eyes with glaucomatous visual field damage were included in this group, and it is likely that some healthy eyes with suspicious optic discs were included in the glaucoma group. These situations combined very likely resulted in smaller overall AUROCs. However, we chose these criteria to decrease, as much as is practical, the bias in favor of either structural or functional tests and, again, establishing classifier performance, per se, was not the main purpose of the study. In addition, it is possible that other MLCs could more accurately separate our data. Our decision to use these particular classifiers was based on experience gained from our prior studies. We chose to use RVM because it has performed well with imaging data in the past and we chose to use SSMoG because MoG has performed well with visual field data in the past.

Overall, the current results suggest that RVM and SSMoG classifiers trained on OCT and/or SAP measurements can successfully discriminate between healthy and glaucomatous eyes. In addition, techniques designed to optimize the MLCs can improve their performance. Classification by RVM and SSMoG trained on OCT and/or SAP measurements is better than classification based on individual OCT parameters and SAP global indices. Combining OCT and SAP measurements using these MLC techniques improved performance marginally compared to the same classifiers trained on OCT and SAP data alone, although AUROCs for combined analysis displayed a consistent increasing trend.

References


