tvst

Article

Development and Comparison of Machine Learning Algorithms to Determine Visual Field Progression

Osamah Saeedi[1], Michael V. Boland[2], Loris D'Acunto[3], Ramya Swamy[1], Vikram Hegde[3], Surabhi Gupta[3], Amin Venjara[4], Joby Tsai[1], Jonathan S. Myers[5], Sarah R. Wellik[6], Gustavo DeMoraes[7], Louis R. Pasquale[8,9], Lucy Q. Shen[2], Yangjiani Li[10], and Tobias Elze[10]

[1] University of Maryland Department of Ophthalmology and Visual Sciences, Baltimore, MD, USA
[2] Massachusetts Eye and Ear, Boston, MA, USA
[3] San Francisco, CA, USA
[4] Princeton, NJ, USA
[5] Wills Eye Hospital, Philadelphia, PA, USA
[6] Bascom Palmer Eye Institute, University of Miami School of Medicine, Miami, FL, USA
[7] Columbia University, New York, NY, USA
[8] Icahn School of Medicine at Mount Sinai, Department of Ophthalmology, New York, NY, USA
[9] Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
[10] Schepens Eye Research Institute of Massachusetts Eye and Ear, Department of Ophthalmology, Harvard Medical School, Boston, MA, USA

Correspondence: Osamah Saeedi, Department of Ophthalmology and Visual Sciences, University of Maryland School of Medicine, 419 W Redwood Street, Suite 470, Baltimore, MD 21201, USA. e-mail: osaeedi@som.umaryland.edu

Received: May 11, 2020
Accepted: April 17, 2021
Published: June 22, 2021

Keywords: glaucoma; visual fields; machine learning

Citation: Saeedi O, Boland MV, D'Acunto L, Swamy R, Hegde V, Gupta S, Venjara A, Tsai J, Myers JS, Wellik SR, DeMoraes G, Pasquale LR, Shen LQ, Li Y, Elze T. Development and comparison of machine learning algorithms to determine visual field progression. Transl Vis Sci Technol. 2021;10(7):27, https://doi.org/10.1167/tvst.10.7.27

Purpose: To develop and test machine learning classifiers (MLCs) for determining visual field progression.

Methods: In total, 90,713 visual fields from 13,156 eyes were included. Six different progression algorithms (linear regression of mean deviation, linear regression of the visual field index, Advanced Glaucoma Intervention Study algorithm, Collaborative Initial Glaucoma Treatment Study algorithm, pointwise linear regression [PLR], and permutation of PLR) were applied to classify each eye as progressing or stable. Six MLCs were applied (logistic regression, random forest, extreme gradient boosting, support vector classifier, convolutional neural network, fully connected neural network) using a training and testing set. For MLC input, visual fields for a given eye were divided into the first and second half and each location averaged over time within each half. Each algorithm was tested for accuracy, sensitivity, positive predictive value, and class bias with a subset of visual fields labeled by a panel of three experts from 161 eyes.

Results: MLCs had similar performance metrics as some of the conventional algorithms and ranged from 87% to 91% accurate with sensitivity ranging from 0.83 to 0.88 and specificity from 0.92 to 0.96. All conventional algorithms showed significant class bias, meaning each individual algorithm was more likely to grade uncertain cases as either progressing or stable ($P \leq 0.01$). Conversely, all MLCs were balanced, meaning they were equally likely to grade uncertain cases as either progressing or stable ($P \geq 0.08$).

Conclusions: MLCs showed a moderate to high level of accuracy, sensitivity, and specificity and were more balanced than conventional algorithms.

Translational Relevance: MLCs may help to determine visual field progression.

Introduction

Visual fields remain a crucial tool to identify and monitor glaucomatous vision loss. As such, measurement of visual field worsening is critical to both patient care and assessment of glaucoma interventions. However, uncertainty regarding visual field progression may delay effective treatment and stabilization of disease. Over the years, numerous algorithms for visual field progression have been developed. Agreement among these algorithms and in comparison with expert graders can vary widely depending on which dataset they are applied to.[1–3]

Machine learning is a tool that uses artificial intelligence and allows systems to learn from an experience without explicit programming. Machine learning classifiers are computerized programs that use supervised learning to classify a new observation. The use of such classifiers has grown exponentially in recent years with numerous applications in business and medicine.[4] Ophthalmology, with its emphasis on imaging and ancillary testing, provides an ideal test case for the clinical application of machine learning classifiers. Recent work has focused on application of deep learning to detection of diabetic retinopathy, age-related macular degeneration, or glaucomatous-appearing optic nerves from fundus photographs[5,6]

Work using artificial intelligence to classify glaucoma first began in the 1990s with the development of neural networks to distinguish glaucomatous versus normal visual fields.[7–9] Starting in the early 2000s, machine learning classifiers were applied to visual field data to determine progression.[10,11] This prior work used smaller datasets and was generally applied to patients at a single site. Development of machine learning solutions, particularly deep learning generally requires access to "big data."[12] Applied to tens of thousands of visual fields, machine learning could allow for the development of an classifier to robustly determine visual field progression with greater accuracy than current methods. To test this hypothesis, we applied the most commonly used machine learning classifiers to a dataset of 90,713 visual fields from five different institutions. We then assessed their accuracy and precision as compared to results from a panel of glaucoma specialists.

Methods

Data Source: Glaucoma Research Network Visual Field Database

We used the visual field database of the Glaucoma Research Network which includes visual fields from the Wilmer Eye Institute (Baltimore, MD, USA), the Massachusetts Eye and Ear (Boston, MA, USA), the Wills Eye Hospital (Philadelphia, PA, USA), the New York Eye and Ear Infirmary (New York, NY, USA), and the Bascom Palmer Eye Institute (Miami, FL, USA). This dataset has been used in prior work characterizing visual fields.[13–19] This retrospective study was approved by the Institutional Review Boards of the participating institutions and adhered to the tenets of the Declaration of Helsinki. Identifying information from the visual fields was removed, but all other information from each test was retained. No clinical or diagnostic information was available for any of the subjects.

Inclusion/Exclusion Criteria

Of the initial dataset of 831,240 visual fields from 177,172 patients, we included only visual fields that were SITA Standard 24-2 with a white size III stimulus on a white background. Only tests from patients older than age 18 were included. We excluded any tests with 20% or greater fixation losses, false-positive results greater than or equal to 15%, or false-negative errors of N/A or not available. A false-negative rating of N/A indicates that there was an insufficient number of test points eligible for presentation of false-negative catch trials. This happens because either fewer than 6 false negative questions are asked or more than 7% of the test points have a threshold < 0 dB (Carl Zeiss Meditec, personal communication, November 26, 2018). We also excluded fields in which the Glaucoma Hemifield Test noted "abnormally high sensitivity," "general reduction of sensitivity," or "borderline/general reduction." After these exclusion criteria were applied, only eyes with five eligible studies were included in the analysis. Depending on which eyes met eligibility criteria, both eyes were used in some patients and one eye in other patients.

Supervised Machine Learning

To apply machine learning classifiers to a large set of visual fields we defined progression as progression in four of six conventional algorithms (mean deviation [MD] slope, Visual Field Index [VFI] slope, Advanced Glaucoma Intervention Study visual field score, Collaborative Initial Glaucoma Treatment Study visual field score, Pointwise Linear Regression (PLR), and Permutation of Pointwise Linear Regression), which we have termed "majority progression." The specific method of programming these algorithms and determining majority progression is detailed elsewhere[19] and explained briefly here. We assigned each eye a label of "progressing," "stable," or "unclear" based on the following criteria: If four or more

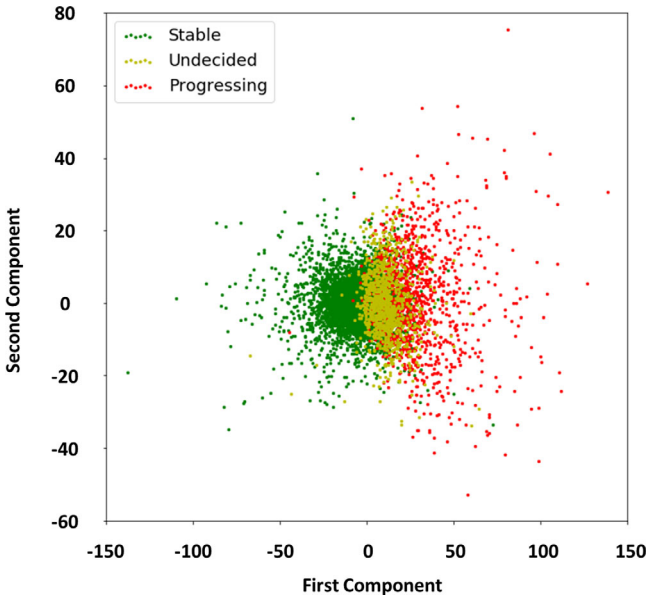

Figure 1. Principal component analysis: Principal component analysis shows that progressing cases (*red*) and stable cases (*green*) are generally segregated with undecided cases between the two (*yellow*).

algorithms classified one eye as progressing, then it was labeled as "progressing." If four or more algorithms determined an eye was stable, then it was labeled as "stable." If an eye was rated as "progressing" by three algorithms and "stable" by the other three algorithms, then it was assigned a label of "unclear." In this way, the majority label acted as a proxy label for visual field progression.

After labeling we analyzed the distribution of the classifiers using principal component analysis (PCA). PCA was applied to the dataset to confirm two distinct regions of stable and progressing eyes with unclear eyes in the middle of the continuum (Fig. 1). The inputs for PCA were the six progression algorithms that result in a six-dimensional space with each dimension being one conventional algorithm. PCA analysis results in a plot of each data point in this six-dimensional space along a single dimension (the first principal component) that contains the maximum variance in the data. When color coded to indicate the majority vote for a data point, we found a clear separation of progressing and stable eyes with uncertain eyes in the middle. This confirmed to us that majority vote of the six progression algorithms was a consistent proxy label that represented a real and important pattern in the data.

We divided the dataset into a training set consisting of 80% of eligible eyes (selected at random) to train the machine learning classifiers and a testing set consisting of 20% of eligible eyes to determine how effective each machine learning classifier was. We then had a panel of three glaucoma specialists (M.B., R.S., and O.S.) grade a subset of 161 eyes of 161 different patients and compared accuracy and performance of the machine learning classifiers as compared to the expert panel. This is further detailed below.

Machine Learning Classifiers

We utilized the following commonly used machine learning classifiers:[20]

- Logistic regression
- Random forest
- Extreme gradient boosting
- Support vector classifier
- Fully connected neural network a (sometimes referred to as multi-layer perceptron)
- Convolutional neural network

Machine learning classifiers used here require a fixed length input. To account for different numbers of visual fields and varying lengths of follow up, visual fields for a given eye were divided into the first and second half, and each location for the first and second half visual fields were averaged over time within each half. In the case of an odd number of visual fields, the points from the middle field were included in the average of the latter "half." The pointwise difference between the two means was then calculated and that was then used as input into each classifier. The dataset was then divided into a training and a test set. The training set included 80% (7745 eyes) of the included eyes and the test set included 20% (2631 eyes). Details on the fully connected neural network parameters, convolutional neural network parameters, and the remainder of the machine learning classifiers can be found in the Supplementary Material. Supplementary Table S1 describes the structure of the fully connected neural network and Supplementary Figure S1 describes the loss by epoch of the fully connected neural network. Supplementary Figure S2 shows the shape normalization of the matrix required for the convolutional neural network and Supplementary Table S2 shows its structure. Supplementary Table S3 lists the hyperparameters for the remaining MLCs.

We deliberately underfit the machine learning classifiers to make them more resilient against minor irregularities in the training data set. The tuning was carried out by assuming that given the large training set, there exists for training set eyes, a normal (i.e., Gaussian) distribution of predicted class probabilities in each predicted class (stable and progressing). We felt that a population of more than 13,000 eyes was sufficiently large to give us a reasonably accurate distribution

within each category. We fit models to the training set using the simplest and least expressive model possible (e.g., for a random forest, the hyperparameter we used was the maximum depth of the trees in the forest). We increased the max-depth hyperparameter until we had a reasonably accurate normal (i.e., Gaussian) distribution of eyes in each of the stable and progressing categories. The accurate normal distribution was determined on the basis of the output of a probability plot plotting the ordered values versus theoretical quantiles.[21] In other words, we chose the simplest model that resulted in a normal distribution of eyes in each class. We did not use the test set at any point for tuning the model.

Precision (positive predictive value), recall (sensitivity), and F1 score (harmonic mean of precision and recall) are commonly used metrics[22] to assess the predictive value of machine learning classifiers. We calculated these values as well as specificity and negative predictive value for each machine learning algorithm when applied to the test set including "unclear" cases and used the weighted average[23] for these metrics to take into account performance on both positive and negatives. We compared the performance of each machine learning classifier in eyes with short-term (< 10 years) versus long-term (>10 years) follow-up (Supplementary Table S4). We then conducted an extensive subanalysis assessing the performance of each machine learning algorithm by severity. Severity was assessed by dividing the testing set into mild (MD > −6 dB) (n = 1854), moderate (−6 dB > MD > −12 dB) (n = 433) and severe (−12 dB > MD) (n = 344) visual fields based on the first visual field for each eye (Supplementary Table S5). We also calculated the Cohen's kappa values between pairs of machine learning algorithms as well as between the majority of six conventional algorithms (Supplementary Table S6).

Expert Grading Panel

Three glaucoma specialists (M.B., R.S., and O.S.) graded a subset of the test set of 161 eyes of 161 different patients, grading them as "stable," "progressing," or "unclear." The graders were masked to the individual algorithms' ratings but had data that would be available in a single field analysis including VFI and MD. Glaucoma progression analysis and MD or VFI regressions were not available to the graders. The graders were masked to the individual algorithms' ratings but had data that would be available in a single field analysis including VFI and MD. Glaucoma progression analysis and MD or VFI regressions were not available to the graders. The 161 eyes were selected from the test set to have a mix of eyes that were progressing, stable, or were "unclear" (three algorithms progressing and three stable). Forty eyes that were progressing, 57 eyes that were stable, and 64 eyes that were "unclear" (as determined by the majority of six conventional algorithms) were chosen. If there was a disagreement between specialists, then the grade that was given by two of the three graders was assigned. If the three graders all differed in their assessment, the assigned label was "unclear." Using the expert panel decision as the "ground truth," the overall weighted average

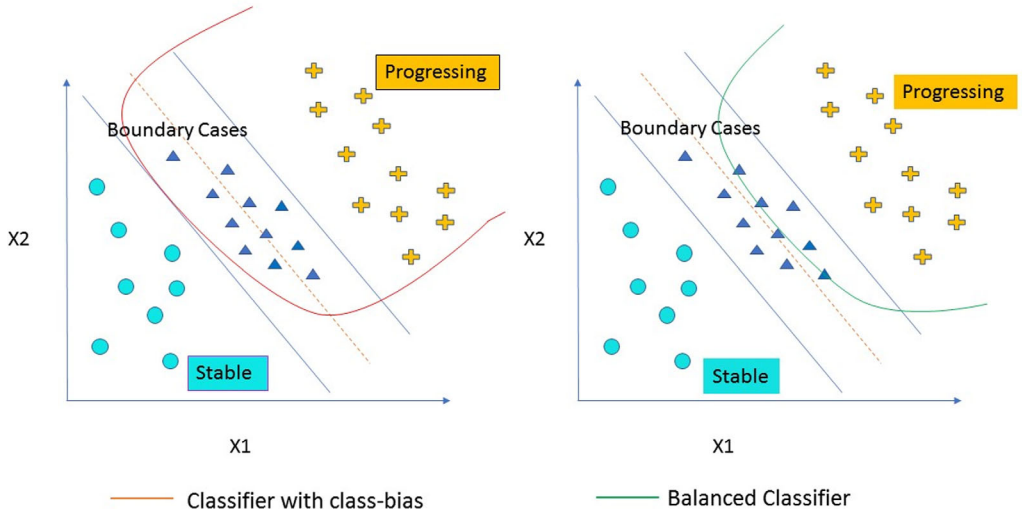

Figure 2. Biased and balanced classifiers: Graphical illustration displaying class bias and overfitting on the left and a balanced classifier on the right. *Yellow crosses* represent progressing cases, *blue dots* symbolize stable cases, and *blue triangles* represent unclear cases. In the biased classifier on the left, the boundary cases are all considered progressing whereas in the balanced classifier on the right, the boundary cases are more evenly split between progressing and stable.

of accuracy, precision (positive predictive value), recall (sensitivity), specificity, negative predictive value, and F1 score (harmonic mean of precision and recall) were calculated for the each conventional and machine learning algorithm for the stable and progressing cases. "Unclear" cases were not evaluated in this analysis but subsequently analyzed separately to check for class bias, as explained below.

Given that the output of conventional algorithms is binary, and the output of the machine learning classifiers includes three potential values (progressing, not progressing, unclear), we used the weighted average of the progressing and stable cases for the machine learning classifiers. This was done for accuracy, sensitivity, specificity, the false-positive rate, false-negative rate, and the F1 score. Receiver operating characteristic curves could not be calculated given that the machine learning classifier output was not binary.

To check for class bias, or overfitting of an algorithm, the χ^2 statistic was used to determine if the unclear cases were split evenly between progressing and stable cases. For example, a classifier with class bias might classify a large proportion of unclear cases as progressing, when one would expect that the unclear or "boundary cases" would split evenly between progression and stable (Fig. 2). We aimed to create balanced classifiers that did not overfit the data in this fashion.

Results

A total of 90,713 visual fields of 13,156 eyes of 9143 patients met the inclusion criteria. The average age was 67.1 ± 12.3 years. The data set included 6479 (49.2%) right eyes and 6677 (50.8%) left eyes. The average mean deviation of all visual fields was −5.6 ± 6.3 dB. Each eye had an average of 6.9 ± 2.4 visual fields mean follow up of 6.3 ± 2.6 years, and an average of 1.25 ± 1.35 fields per year. A total of 4669 eyes (35.5%) had five fields in the dataset, 7351 eyes (55.9%) had six to 10 fields, and 1136 (8.6%) had more than 10 fields. Detailed information about the features of this dataset can be found in the previously published paper using this dataset.[19] In at least four of six algorithms, 11.7% of eyes progressed. After completion of training, each machine learning classifier was compared to the majority progression as determined by six conventional algorithms. Table 1 shows the Precision, Recall, AUC, and F1 score for each machine learning classifier when applied to the testing dataset. Supplementary Tables S4 and S5 show the performance of each machine learning classifier stratified by length of follow up and severity respectively. Supplementary Table S4 shows that the machine learning classifiers generally performed similarly in short-term and long-term follow-up. Supplementary Table S5 shows that that machine learning classifiers perform better in mild glaucoma (MD ≥ −6.0 dB) than in more severe glaucoma. Supplementary Table S6 shows Cohen's kappa values between pairs of machine learning classifiers, as well as between the majority of six conventional algorithms. Machine learning classifiers have a relatively high degree of concordance with each other (Kappa range, 0.50–0.75).

Results of Grading by Expert Panel

Of the 161 cases graded by the expert panel, 53 (32.9%) were stable, 68 (42.2%) were progressing, and 40 (24.8%) were graded as "unclear." The average age of these selected eyes was 64.0 ± 12.4 years, average MD was −7.5 ± 6.4 dB, and the average number of visual fields was 6.7 ± 2.1. For the 121 that were labeled as progressing or stable, the weighted average of accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and χ^2 values to determine whether the classifiers were balanced were calculated and listed in Table 2. The majority progression by the six conventional algorithms yielded an accuracy of 83%, whereas the machine learning classifiers had an accuracy of 87% to 90% with high values of precision (PPV), recall (sensitivity), and F1 score.

Table 1. Performance Metrics for Machine Learning Classifiers as Compared to Majority Progression of Six Conventional Algorithms in the Testing Dataset

	Sensitivity (Recall)	Specificity	Positive Predictive Value (Precision)	Negative Predictive Value	F1 Score
Random forest	0.79	0.94	0.88	0.91	0.82
Logistic regression	0.85	0.94	0.87	0.94	0.86
Extreme gradient boosting	0.87	0.92	0.84	0.94	0.85
Support vector classifier	0.79	0.95	0.88	0.92	0.83
Convolutional neural network	0.78	0.93	0.87	0.90	0.81
Fully connected neural network	0.84	0.95	0.88	0.94	0.85

Table 2. Performance of All Algorithms on Fields Graded by Panel Of Glaucoma Specialists

	Accuracy	Sensitivity (Recall)	Specificity	Positive Predictive) Value (Precision)	Negative Predictive Value	F1 Score	Proportion Undecided Classified As Progressing	Distribution of "Unclear" Values
Mean deviation	0.73	0.57	0.91	0.91	0.62	0.72	0.35	$P < 0.01$
VFI	0.95	0.97	0.92	0.94	0.96	0.95	0.9	$P < 0.01$
AGIS	0.69	0.47	0.98	0.97	0.59	0.68	0.08	$P < 0.01$
CIGTS	0.69	0.49	0.96	0.94	0.59	0.68	0.13	$P < 0.01$
PLR	0.93	1	0.83	0.88	1	0.92	0.95	$P < 0.01$
PoPLR	0.84	0.78	0.92	0.93	0.77	0.84	0.38	$P = 0.01$
Majority of six conventional algorithms	0.83	0.71	0.97	0.96	0.76	0.81	0	$P < 0.01$
Logistic regression	0.90	0.88	0.92	0.91	0.88	0.90	0.45	$P = 0.37$
Extreme gradient boosting	0.89	0.88	0.90	0.90	0.88	0.89	0.4	$P = 0.20$
Support vector classifier	0.91	0.88	0.94	0.94	0.87	0.91	0.43	$P = 0.80$
Convolutional neural network	0.89	0.83	0.96	0.95	0.84	0.89	0.5	$P = 0.08$
Fully connected neural network	0.87	0.83	0.92	0.91	0.84	0.87	0.33	$P = 0.17$
Random forest	0.90	0.88	0.92	0.91	0.88	0.90	0.55	$P = 0.39$

AGIS, Advanced Glaucoma Intervention Study; CIGTS, Collaborative Initial Glaucoma Treatment Study; PoPLR, Permutation of Pointwise Linear Regression.

Individual algorithms such as VFI and PLR showed high degrees of accuracy, sensitivity, and specificity but showed some evidence of class bias with a disproportionate number of boundary cases classified as progressing. All machine learning classifiers classify the boundary values in an expected proportional distribution ($P \geq 0.08$), whereas using the majority of six conventional algorithms, the χ^2 value is statistically significant and deviates from the expected distribution ($P \leq 0.01$). The dataset of expert panel results is available as supplemental material.

Discussion

Machine learning classifiers yielded close to 90% accuracy and high sensitivity and specificity for determining glaucoma progression. Furthermore, although all conventional algorithms showed class bias, the machine learning classifiers were balanced, with a proportional amount of borderline cases classified as progressing or stable. Specifically, VFI slope and PLR showed the greatest accuracy and sensitivity but might "overcall" progression, which would limit their applicability to other datasets. This study serves as further evidence that machine learning classifiers can be applied to visual field progression and potentially result in equivalent or better accuracy than existing algorithms that are more balanced and thus potentially applicable to a wider dataset. In the future, machine learning may aid clinicians in determining which patients show progression of visual field.

Our goal in modeling visual field progression with multiple machine learning classifiers was not to come up with a perfect classifier for the proxy label we used. Rather, we aimed to create a robust classifier while accounting for an imperfect label that may incorrectly classify a small number of samples. We did this by tuning our hyperparameters to deliberately underfit the noisy proxy label.

Application of different machine learning methods is commonly used in other fields and allows for comparison of different algorithms and the ultimate choice of an optimal algorithm.[24,25] Ultimately, the machine learning classifiers, while more balanced, had comparable accuracy to other algorithms. Also, like other conventional algorithms, machine learning classifiers tended to perform better in mild glaucoma than in moderate or severe glaucoma. This could be potentially improved with a larger dataset of visual fields, as well as better labeling of the data. Notably machine learning classifiers have a higher degree of concordance (Kappa range, 0.50–0.75) with each other than conventional algorithms, which we reported had relatively poor agreement with each other (Kappa range, 0.12–0.52), which we have shown in a previously published article using these data.[19]

Prior work assessing the role of machine learning in visual field interpretation generally included smaller sample sizes of patients with more clinical data. In a series of 628 eyes, Goldbaum et al.[10] compared the machine learning classifier, Progression of Patterns (PoP) to VFI, MD, and the glaucoma progression analysis (GPA) and found that it was comparable or better than these algorithms and concluded that machine learning may aid clinicians in detecting visual field progression. Yousefi et al.[26] applied variants of the PoP algorithm to 167 eyes and showed that they were significantly more sensitive than permutation of pointwise linear regression, and linear regression of MD and VFI. Similar to our analysis, these studies used only data from the visual field as input into the respective algorithms, but used associated disc photos to aid in labeling progression. Our work complements this prior work by using a different approach to labeling (majority progression of four algorithms). Although visual fields in our dataset come from heterogeneous disease states as compared to prior work, this shows the potential applicability of machine learning classifiers to "real-world" data. The larger number of eyes and visual fields in our dataset provides an advantage over prior work.

This study incorporated only information available in the visual field and did not include clinical information such as intraocular pressure, optic nerve assessment, biometric properties such as central corneal thickness or axial length, or history of ocular and systemic diseases. This information, if added to the machine learning classifiers may improve accuracy and performance. Given that we used the difference in pointwise means between the average of the first and second half of visual fields, it is possible that the number of tests or length of follow up could influence the performance of classifiers. One potential reason that the convolutional neural network did not perform better than the other MLCs is a relatively small feature set. Another potential limitation is that the expert panel decision may not be completely accurate, and it is possible that a very sensitive algorithm could detect progression in eyes that the expert panel deemed "unclear."

Machine learning holds great promise in glaucoma management and guiding the outcomes of glaucoma clinical trials. Machine learning classifiers may be useful for patients enrolled in trials across the spectrum of disease severity, but more study is needed to test these assertions. The development of more accurate and precise algorithms aided by machine learning

could be used in tele-glaucoma consultation,[27] as well as an aid for practitioners in the clinic to determine glaucoma progression. Ultimately this may further help to stratify patients as low or high risk and determine appropriate intervals of follow-up.

Acknowledgments

Supported by an NIH Career Development Award (K23EY025014) (O.S.), NIH grants R21EY030142 (T.E.), R01EY030575 (T.E.); and NIH/NEI R01 grants EY015473 (L.R.P.) and R01 EY025253 (G.DeM.); NEI Core Grant P30EY003790 (T.E.), the BrightFocus Foundation (T.E.), the Lions Foundation (T.E.), the Grimshaw-Gudewicz Foundation (T.E.), Research to Prevent Blindness and the Alice Adler Fellowship (T.E.), Research to Prevent Blindness (G.DeM.), and the Harvard Glaucoma Center of Excellence (L.Q.S.).

Disclosure: **O. Saeedi,** None; **M.V. Boland,** None; **L. D'Acunto,** None; **R. Swamy,** None; **V. Hegde,** None; **S. Gupta,** None; **A. Venjara,** None; **J. Tsai,** None; **J.S. Myers,** None; **S.R. Wellik,** None; **G. DeMoraes,** None; **L.R. Pasquale,** None; **L.Q. Shen,** None; **Y. Li,** None; **T. Elze,** None

References

1. Heijl A, Bengtsson B, Chauhan BC, et al. A comparison of visual field progression criteria of 3 major glaucoma trials in early manifest glaucoma trial patients. *Ophthalmology.* 2008;115:1557–1565.
2. Birch MK, Wishart PK, O'Donnell NP. Determining progressive visual field loss in serial Humphrey visual fields. *Ophthalmology.* 1995;102:1227–1234; discussion 1234–1235.
3. Katz J, Congdon N, Friedman DS. Methodological variations in estimating apparent progressive visual field loss in clinical trials of glaucoma treatment. *Arch Ophthalmol*. 1999;117:1137–1142.
4. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology.* 2018;286:800–809.
5. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA.* 2016;316:2402–2410.
6. Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA.* 2017;318:2211–2223.
7. 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.
8. Brigatti L, Hoffman D, Caprioli J. Neural networks to identify glaucoma with structural and functional measurements. *Am J Ophthalmol*. 1996; 121:511–521.
9. Lietman T, Eng J, Katz J, Quigley HA. Neural networks for visual field analysis: how do they compare with other algorithms? *J Glaucoma.* 1999;8:77–80.
10. Goldbaum MH, Lee I, Jang G, et al. Progression of patterns (POP): a machine classifier algorithm to identify glaucoma progression in visual fields. *Invest Ophthalmol Vis Sci.* 2012;53:6557–6567.
11. Lin A, Hoffman D, Gaasterland DE, Caprioli J. Neural networks to identify glaucomatous visual field progression. *Am J Ophthalmol*. 2003;135:49–54.
12. Rodriguez F, Scheinker D, Harrington RA. Promise and perils of big data and artificial intelligence in clinical medicine and biomedical research. *Circ Res.* 2018;123:1282–1284.
13. Wang M, Pasquale LR, Shen LQ, et al. Reversal of glaucoma hemifield test results and visual field features in glaucoma. *Ophthalmology.* 2018;125:352–360.
14. Wang M, Shen LQ, Boland MV, et al. Impact of natural blind spot location on perimetry. *Sci Rep.* 2017;7:6143.
15. Wang M, Shen LQ, Pasquale LR, et al. An artificial intelligence approach to detect visual field progression in glaucoma based on spatial pattern analysis. *Invest Ophthalmol Vis Sci*. 2019;60:365–375.
16. Bommakanti N, De Moraes CG, Boland MV, et al. Baseline age and mean deviation affect the rate of glaucomatous vision loss. *J Glaucoma*. 2020;29:31–38.
17. Wang M, Tichelaar J, Pasquale LR, et al. Characterization of central visual field loss in end-stage glaucoma by unsupervised artificial intelligence. *JAMA Ophthalmol*. 2020;138(2):190–198.
18. Wang M, Shen LQ, Pasquale LR, et al. Artificial Intelligence Classification of Central Visual Field Patterns in Glaucoma. *Ophthalmology*. 2019;S0161-6420(19):32329–32332.
19. Saeedi O, Elze T, D'Acunto L, et al. Agreement and Predictors of Discordance of Six Visual Field Progression Algorithms. *Ophthalmology*. 2019;126:822–828.

20. Kotsiantis S.B. Supervised machine learning: A review of classification techniques. In Maglogiannis I.G. (Ed.), *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies.* Fairfax, VA: IOS Press, Inc. 2007:3–24.
21. scipy.stats.probplot. The SciPy Community. Available at https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.probplot.html . Accessed January 8, 2019
22. Powers D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *J Mach Learn Technol.* 2011;2(1):37–63
23. Manning CD, Prabhakar R, Hinrich S. *Introduction to information retrieval*, volume 1. Cambridge, UK. Cambridge University Press. 2008
24. Petersen ML, LeDell E, Schwab J, et al. Super learner analysis of electronic adherence data improves viral prediction and may provide strategies for selective HIV RNA monitoring. *J Acquir Immune Defic Syndr.* 2015;69(1):109–118.
25. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. *Lancet Respir Med.* 2015;3:42–52.
26. Yousefi S, Balasubramanian M, Goldbaum MH, et al. Unsupervised gaussian mixture-model with expectation maximization for detecting glaucomatous progression in standard automated perimetry visual fields. *Transl Vis Sci Technol.* 2016; 5(3):2.
27. Verma S, Arora S, Kassam F, et al. Northern Alberta remote teleglaucoma program: clinical outcomes and patient disposition. *Can J Ophthalmol*. 2014;49:135–140.