HHS Public Access

Author manuscript

Ophthalmol Glaucoma. Author manuscript; available in PMC 2024 September 01.

Published in final edited form as:

Ophthalmol Glaucoma. 2023 ; 6(5): 466–473. doi:10.1016/j.ogla.2023.03.005.

Forecasting Risk of Future Rapid Glaucoma Worsening Using Early Visual Field, Optical Coherence Tomography, and Clinical Data

Patrick Herbert[1], **Kaihua Hou**[1], **Chris Bradley, PhD**[2], **Greg Hager, PhD**[1], **Michael V. Boland, MD,PhD**[3], **Pradeep Ramulu, MD,PhD**[2], **Mathias Unberath, PhD**[1], **Jithin Yohannan, MD,MPH**[1,2]

[1.]Malone Center For Engineering in Healthcare, Johns Hopkins University

[2.]Wilmer Eye Institute, Johns Hopkins University

[3.]Massachusetts Eye and Ear Infirmary, Harvard Medical School

Abstract

Purpose—Assess whether we can forecast future rapid visual field (VF) worsening using deep learning models (DLMs) trained on early VF, OCT, and clinical data.

Design—Retrospective cohort study.

Subjects—4,536 eyes from 2,962 patients. 263 (5.80%) of eyes underwent rapid VF worsening (MD slope < −1dB/yr across all VFs).

Methods—We included eyes that met the following criteria: 1) followed for glaucoma or suspect status 2) had at least five longitudinal reliable VFs (VF_1, VF_2, VF_3, VF_4, VF_5) 3) had one reliable baseline Optical Coherence Tomography (OCT) scan (OCT_1) and one set of baseline clinical measurements ($Clinical_1$) at the time of VF_1.

We designed a DLM to forecast future rapid VF worsening. The input consisted of spatially oriented total deviation values from VF_1 (including or not including VF_2 and VF_3 in some models) and retinal nerve fiber layer thickness values from the baseline OCT. We passed this VF/OCT stack into a vision transformer feature extractor, the output of which was concatenated with baseline clinical data before putting it through a linear classifier to predict that eye's risk of rapid VF worsening across the five VFs.

Corresponding Author: Jithin Yohannan
Address for Reprints: Wilmer Eye Institute, 600 N Wolfe St, Baltimore MD 21287

Conflict of Interest: No conflicting relationship exists for any author.

Meeting Presentations:
Presented on 03/03/22 at AGS in Nashville, TN.
Presented on 05/03/22 at ARVO in Denver, CO.

We demonstrate that deep learning models can forecast an eye's risk of rapid glaucoma worsening using data from as little as one visit. Adding multimodal data and longitudinal data improves model performance.

We compared the performance of models with differing inputs by computing area under receiver operating curve (AUC) in the test set. Specifically, we trained models with the following inputs: Model V: VF_1; VC: VF_1+ $Clinical_1$; VO: VF_1+ OCT_1; VOC: VF_1+ $Clinical_1$+ OCT_1; V_2: VF_1 + VF_2; V_2OC: VF_1 + VF_2 + $Clinical_1$ + OCT_1; V_3: VF_1 + VF_2 + VF_3; V_3OC: VF_1 + VF_2 + VF_3 + $Clinical_1$ + OCT_1.

Main Outcome Measures—AUC of DLMs when forecasting rapidly worsening eyes.

Results—Model V_3OC best forecasted rapid worsening with an AUC (95% CI) of 0.87 (0.77, 0.97). Remaining models in descending order of performance and their respective AUC [95% CI] were: Model V_3 (0.84 [0.74 to 0.95]), Model V_2OC (0.81 [0.70 to 0.92]), Model V_2 (0.81 [0.70 to 0.82]), Model VOC (0.77 [0.65, 0.88]), Model VO [0.75 [0.64, 0.88], Model VC (0.75 [0.63, 0.87]), Model V (0.74 [0.62, 0.86]).

Conclusion: DLMs can forecast future rapid glaucoma worsening with modest to high performance when trained using data from early in the disease course. Including baseline data from multiple modalities and subsequent visits improves performance beyond using VF data alone.

Keywords

Deep Learning; Glaucoma; Forecasting

It is estimated that approximately 5–10% of glaucoma patients in clinical populations — general population metrics may differ — will worsen rapidly,[1] which can result in significant deterioration in quality of life[3] and increasing disability.[4] With currently employed testing strategies, it often takes several years to identify rapid disease worsening during which time visual function may be lost.[5] Approaches that allow providers to accurately identify rapidly worsening patients earlier in the disease course may allow clinicians to tailor follow-up and treatment regimens to prevent vision loss. Additionally, early risk stratification would allow the healthcare system to target more resources toward high-risk groups which may allow for more efficient and cost-effective utilization of limited health care resources. Furthermore, development of methods that can identify high risk eyes earlier in the disease course may allow researchers to enrich enrollment of such high-risk eyes in clinical trials (e.g., neuroprotection studies) so that treatment effects can be demonstrated with smaller sample sizes and shorter study duration.

Several prior studies have attempted to predict future visual field (VF) change using available data. The use of generative models for forecasting future VFs and patterns in their change over time has been explored and shown promising results.[6,7] Some have attempted to use linear models of past mean deviation (MD) to predict future VF loss, but the requirement of a relatively large number of VFs to accurately estimate future VF loss is a major drawback because of the possibility of loss to follow-up as well as lost vision while patients are waiting for the diagnosis of worsening.[8] Additionally, obtaining larger numbers of VFs is not always feasible in real-world clinical settings where treatment decisions often need to be made after just a few visits, making early and accurate identification of rapid progressors difficult. To address this problem, we previously explored the use of a variety of machine learning methods to predict rapid worsening using only baseline VF data[2] and

achieved a modest area under the curve (AUC) of 0.72 for predicting future rapid worsening with baseline VF data alone. While a step in the right direction, for our models to be useful for clinical purposes, we need to achieve better performance (i.e., AUC > 0.8).[9]

We set out to produce better predictive models for risk of glaucoma worsening by exploring several methods to improve our ability to forecast future rapid worsening. First, we included multimodal data (i.e., more than just VF) such as baseline clinical information (i.e., IOP, visual acuity) and baseline structural optical coherence tomography (OCT) data in addition to the baseline VF as inputs into our models. Additionally, we evaluated a new machine learning architecture by employing a transformer based deep learning network. Transformers have recently shown promising results for replacing previous deep learning architectures such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs or LSTMs).[10,11] Third, we explored adding additional longitudinal VF data (one to two visits beyond baseline) to see if there is an improvement in model predictions.

Methods

This study was reviewed and approved by the Johns Hopkins Medicine Institutional Review Board and adheres to the tenets of the Declaration of Helsinki. The need for informed consent was waived due to the retrospective nature of the study.

We included eyes followed from 2013 to 2021 in our clinical database based on the following criteria: 1) followed for glaucoma or suspect status 2) had at least five reliable VFs (VF_1 – VF_5) 3) had one reliable baseline OCT scan (OCT_1) and 4) one set of baseline clinical measurements that included: age, gender, best documented visual acuity (BDVA), and intraocular pressure (IOP) at the time of the first VF ($Clinical_1$). We classified reliable VF data as < 15% false positives and either < 25% false negatives for mild/moderate glaucoma or < 50% false negatives for severe glaucoma.[12] We classified reliable OCTs as those with signal strength greater than or equal to 6 and an average RNFL thickness greater than 30 μm. We set the floor at 30 to account for artifacts in the scan causing thickness values to appear very low (i.e., below the physiological floor of approximately 50 microns on Cirrus OCT).[13,14] All VFs used in this study were done with SITA strategies (not including SITA-SWAP) and 24–2 test patterns. For eyes with more than five VFs in the specified time period, we labeled the VF recorded at baseline (i.e., the same time as its OCT and clinical information) as VF_1. We classified eyes as rapid progressors if their MD slope was worse (more negative) than −1 dB/yr across all VFs.[15] This slope was estimated for each eye using linear regression over time using all of the patient's VF data (e.g., if the patient has 9 VF tests, then the slope was calculated using all 9). The process of using linear regression on MD to quantify sensitivity loss (adjusted for age) has been shown to detect deterioration better than other global indices, particularly in patients with more rapidly progressing glaucoma.[16] Since univariable regression of MD over time may be prone to outliers among the last observations, tests were done using robust regression methods and no improvements were shown to performance, so we used simple linear regression.

Next, we built our transformer based deep learning model to forecast future VF worsening (Figure 1). We spatially oriented the OCT retinal nerve fiber layer (RNFL) thickness data

into a 12×12 grid to match the quadrant or clock hour divisions of the scan it came from. Further, we also radially imputed the total deviation values from the VF in order to fill out an analogous 12×12 grid. We then stacked these three images to form a three-channel image for every eye which became the input to the vision transformer. Anatomically, nerve fiber layers can lie at different orientations in different eyes,[17] and to account for this we applied data augmentation techniques such as random cropping and rotation to the OCT layers of the stacked image. While the spatial relationships between these channels are not reflected perfectly, the vision transformer will still be able to learn the appropriate relations. We combined the output of the vision transformer (64 extracted features) with another vector of 34 features that included clinical information, global features of the VF (i.e., MD, reliability indices), and global features from the OCT (i.e., average RNFL thickness, disc parameters, signal strength). We then input this combined vector into a fully connected classifier whose output was the probability of future rapid worsening.

We trained a total of 8 different models: Model “V” which includes only baseline VF (VF_1) as input; Model “VC” which includes baseline VF and clinical information (VF_1+ $Clinical_1$); “VO” which includes baseline VF and OCT information (VF_1+ OCT_1); “VOC” which includes VF_1 + OCT_1 + $Clinical_1$; “V_2” which includes VF_1 + VF_2; “V_2OC” which includes VF_1 + VF_2 + OCT_1 + $Clinical_1$; “V_3” which includes VF_1 + VF_2 + VF_3; and “V_3OC” which includes VF_1 + VF_2 + VF_3, + OCT_1 + $Clinical_1$. VF data are included in all models since our objective is forecasting future rapid VF worsening. The architecture shown in Figure 1 is for Model VOC. For models such as V_2OC, where VF data beyond baseline was added, it was done so in the form of additional channels to the stacked image. In these cases, time between VFs was also included as input for the model. For models such as VC, we removed the two OCT channels of the stacked image, as well as the global OCT metrics from the clinical information vector.

We randomly split patients 80%/10%/10% into training/validation/test sets using a stratified method to ensure that there were equal ratios of classes (rapid worsening eyes and non-rapid worsening eyes) in each set. Though our data is at an eye-level, we ensured that an individual patient could not have eyes in more than one set to prevent any data leakage between sets. We split the data for each of these models in an identical fashion, ensuring that the same training and test sets were used for all models. We evaluated model performance with AUC and precision-recall (PR) curves on the held-out test set and further broke down performance via the use of metrics such as precision, recall, F1-Score, and Youden’s J-Index. Recall and sensitivity are identical, while precision is the fraction of true positives among all positives predicted by the model. F1-Score is the harmonic mean of precision and recall and is useful for combining the two metrics into one. Youden’s J-Index is simply $sensitivity + specificity - 1$ and is often used as a summary measure of the ROC curve or to help to select optimal thresholds for classification. It should be noted that we chose to select thresholds that optimized F1-Score or Youden’s J which led to precision and recall or sensitivity and specificity, respectively, being treated with equal weighting. These may not be the thresholds used if such models are clinically deployed, as clinicians may tend to favor thresholds that result in higher sensitivity to avoid false negative predictions (i.e., avoid missing eyes that are rapid progressors).

We performed several types of sensitivity analyses. In order to estimate the effect of within-subject correlations on the confidence intervals of our model performance metrics, we performed a sensitivity analysis where performance was assessed on a test set that included only one randomly selected eye per patient. Since it is possible that some eyes that may have been at high risk for rapid progression underwent glaucoma surgery to reduce this risk, we performed a sensitivity analysis where we included any patient who had undergone surgery for uncontrolled glaucoma (trabeculectomy, tube shunt, Xen, diode laser) within the time period of the VF data that was labeled as a rapid progressor. We chose these specific surgery types as they are most often performed for eyes at risk for or undergoing rapid progression at our institution. Additionally, we performed a sensitivity analysis where we entirely removed patients who had undergone surgery from our dataset. We compared these model results to the model described above (which only used VF data to define rapid progression).

Results

We included a total of 4,536 eyes from 2,962 patients who underwent VF testing in analysis (Table 2). Among these, 263 (5.80%) eyes rapidly worsened. Compared to non-rapid progressors, eyes with rapid worsening were older, had lower MD, higher PSD, higher LogMAR best documented visual acuity, and lower RNFL thickness. Additionally, rapid progressors tended to have worse VF reliability metrics than non-rapid progressors (i.e., longer test duration, higher rate of FNs and FPs). For all comparisons, $p < 0.05$. LogMAR, MD, Rate of MD Change, and PSD had notably different median (IQR) values compared to their mean (SD) reported in Table 1. Overall, Non-rapid Progressor, and Rapid Progressor [Median (IQR)] values for them are as follows: LogMAR [0.10 (0.18), 0.10 (0.10), 0.10 (0.30)]; MD [−2.19 (4.33), −6.61 (8.87), −2.04 (4.01)]; Rate of MD Change [−0.31 (0.39), −0.07 (0.35), −1.39 (0.73)]; PSD [2.23 (3.00), 2.15 (2.58), 5.72 (6.44)].

Model V_3OC most accurately forecasted future rapid worsening with an AUC (95% CI) of 0.87 (0.77, 0.97), followed by Models V_3, V_2OC, V_2, VOC, VO, VC, V in descending order (Figure 2). That V_3OC performed best is expected given that it used more information to make predictions than any of the other models. The results of the PR curve were analogous to the AUC curve, with V_3OC having the highest precision with an equal recall weighting. Across the board, adding additional information predictably improved performance. Additionally, we performed comparisons of our models to baseline logistic regression models trained on one, two, or three VFs. All three logistic regression models performed around 0.62 AUC with much worse results than their counterparts (Models V_1, V_2, V_3) in every metric.

A more detailed overview of performance metrics for each model may be viewed in Table 2. Each of the summary metrics (AUC, Youden's, F1-Score) continues the trend of improvement with the addition of multimodal data and longitudinal data.

Notably, adding more longitudinal data, in the form of additional VFs, caused statistically significant improvements in AUC ($p < .05$) compared to models that used baseline data alone (Table 3). Any model that included two VFs beyond the baseline (V_3, V_3OC) was significantly better than models that used baseline data alone. Of the models that used

baseline data alone, none were significantly better than each other but the trend of more information associated with more performance still held true.

In all of our sensitivity analyses, the difference between the sensitivity analysis results and our original results were not significantly different. We saw little to no change in any model performance metric when only testing on one randomly selected eye per patient to account for within-patient correlation. When also labeling patients who had undergone surgery as rapid progressors, or removing patients who had undergone surgery from the dataset, there was no significant difference in the performance of models compared to labeling rapid progressors with VF data alone.

Discussion

With just two early VF tests combined with baseline OCT and clinical information, our models were able to achieve an AUC of > 0.80 at detecting future rapid glaucoma worsening. In the case where only a single baseline visit was available, we were able to improve upon our previous results through the introduction of a transformer-based deep learning model after adding additional information beyond VFs (clinical and OCT data). Our study demonstrates using multimodal data (VF, OCT, and clinical information) and longitudinal data (serial VF testing) as model inputs can meaningfully improve the ability to identify patients who will become rapid progressors. Deployment of such models, after careful external and prospective validation, may allow clinicians to better risk-stratify patients to improve treatment outcomes and resource allocation. Furthermore, such forecasting models may enhance the ability to enroll appropriate candidates for clinical trials where change in VF worsening is the main outcome (i.e., neuroprotection studies) by allowing for the selection of a higher percentage of rapid progressors in the trial group.

While prior studies used linear models to predict rapid progressors, our model allows for earlier identification of rapid progressors than linear models, which require a larger amount of serially obtained VFs for identification.[5] Other studies used machine learning methods for detection of VF worsening[19,20] or detection of glaucoma using images of the nerve and OCT,[21,22] but not for the task of forecasting future worsening. There have also been some studies that focused on the application of forecasting future findings in glaucoma, such as using generative deep learning models to predict future VFs, including patterns of change.[3,4] However, since most VFs do not change over time, the task of predicting future VFs by way of minimizing their mean absolute error (MAE) is not difficult for machine learning models — if the model simply predicts no change for all eyes, it will incur low MAE loss. In contrast, identifying the small subgroup of the patient population that will undergo rapid worsening is more difficult for predictive models, as the model must correctly select the small subgroup of eyes that will undergo rapid worsening from the larger cohort.

In our previous work, we saw modest AUCs in the single visual field only task with the best performing model being the support vector machine at 0.72 AUC. Our previous deep learning model also showed a marginally worse AUC of 0.71 compared to random forests. Now, Model V (0.74 AUC) performs better than the previous methodology, even though we used a much smaller dataset for training (n =18,340 in the former dataset vs. n = 3,632

in the current dataset). We previously noticed that adding data from the first two VFs did not increase performance over using only the initial VF; however, our new model shows the capability of using additional VF information to improve accuracy as well. This may be because we used spatially oriented longitudinal VF information in this study compared to a vector of non-spatially oriented VF values for the deep learning model in the previous study. Another possible explanation is the learning effect associated with VF testing — meaning the second or third VFs might be a more accurate representation of the patient's disease state.

Furthermore, in this work, we took a closer look at how each individual piece of the VF, OCT, and clinical dataset impacts model performance with ablation studies. The most notable improvements in performance were through the additional VFs. We saw a slightly larger improvement from adding OCT data than clinical data, which could suggest it was more important to the model, but it is important to note that it was not a statistically significant difference as seen in Table 3. Precision values across the board show at least a threefold increase when compared to our base rate of about 5% of eyes in the cohort rapidly worsening, even seeing as much as sixfold when three visual fields are included.

Additionally, it should be noted that we considered sensitivity and specificity to be of equal weight when calculating the values for Table 2. In practice, this may not be desired in a clinical setting where practitioners could assign more weight to true positives than false positives. However, this could easily be adjusted by decreasing the threshold at which the model predicts a positive rapid progressor in order to identify more borderline patients.

Our present work has several strengths. We were able to gather a relatively large multimodal dataset even with our selection criteria, allowing us to create a sizable training split of about 3600 eyes for the model to learn from. On top of the size and multimodality of the dataset, we were able to gather a real-world clinical population rather than the carefully selected population of a study, which allows for better generalizability of our results to a treated clinical population. Additionally, since patients were followed over a period of years, we were also able to add additional information from their longitudinal follow-up visits. Even so, our study was still limited by several factors. We were only able to use the limited summary measures of RNFL thickness and global metrics for our OCT data. Using raw OCT image data may be able to improve our results. We plan to incorporate this into future work to see if it improves model performance. High-definition OCT images may allow our model to identify OCT features associated with rapid progression that are not captured on numeric RNFL or global OCT metrics. Furthermore, MD is not specific to glaucoma, nor sensitive to localized progression. Defining rapid worsening on MD rates may be impacted by things such as cataract progression and focal VF changes occurring in central or paracentral locations. Generalizability of our model is limited to patients undergoing follow-up and treatment for glaucoma or glaucoma suspect status. Additionally, it is also important to check both the external validity of this study, in which case multi-institutional collaborations may help, and to prospectively validate model performance. This is especially important since, in our cohort, those with glaucoma are already being treated, so learning to identify a more aggressively treated group in order identify them earlier and treat them more aggressively may be a bit redundant. An additional limitation of the model is a lack

of interpretability. While the model shows respectable AUC on this task, it cannot explain its decisions in any meaningful way. Using ablation studies (e.g., in VC we removed OCT data from the model and examined how the results were impacted), we attempted to provide a clearer picture of how each piece of data affects the total results, but it is still important to note that there are no clear feature importance values with transformer based deep learning models such as seen in simpler models (logistic regression, random forest etc.). Looking at the transformer's multi-headed attention for additional explainability is a possible avenue for future work as well. Reducing black box architectural designs will allow clinicians to understand why the model classifies some patients as rapid progressors, which allows for better decision making and the ability to catch errors. Improving model interpretability will likely improve clinician and patient trust in the model outputs.[23]

If our model is validated externally and prospectively and deployed for clinical and research use, there would be numerous benefits. First, faster identification of patients at high risk for rapid worsening would enable real-world clinicians to make a more informed decision about treatment or follow-up plans to reduce the risk of serial VF loss (i.e., more frequent follow-up and earlier testing in high-risk eyes). Importantly, use of such models would also allow clinicians to counsel patients on their risk of vision loss from glaucoma and emphasize the importance of follow-up and treatment adherence. Moreover, the results of the model could be used to better allocate healthcare resources by enabling low risk patients to follow with general ophthalmologists/properly trained optometrists and higher risk patients to follow with glaucoma specialists.

Application of the model to future clinical trials may also greatly reduce sample size requirements. The FDA has proposed a 7 dB difference in MD values between treatment and placebo groups as a meaningful difference in VF loss.[24] As 90% of patients with treated glaucoma progress slowly, if patients are non-selectively enrolled in a clinical trial, the sample size requirements to show treatment effect become burdensome.[25] However, enriching enrollment for high-risk eyes could dramatically reduce sample size requirements. For instance, both our V_2OC and VOC models achieve precision of >0.33 and the V_3OC model achieves a precision of 0.49. If such models were used to select eyes to be included in clinical trials, this would result in a three-to-five-fold enrichment over the base rate of approximately 10% rapidly worsening eyes and a substantial reduction in sample size and trial costs.

In summary, we have developed a deep learning model that can identify glaucoma patients at high risk for future rapid worsening using early VF data and baseline clinical and OCT information. Training our model using multimodal and longitudinal data improves model performance over using baseline VF data alone. Our best models achieve clinically useful AUC > 0.8, though more than 1 VF is required to achieve this level of predictive accuracy. This study represents an important step forward in the development of automated methodologies for early the detection of high-risk glaucoma patients which, with further validation, may improve clinical care as well as clinical trial design.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Financial Support:

5 K23 EY032204-02 (JY); Unrestricted Grant from Research to Prevent Blindness (JY); The sponsor or funding organization had no role in the design or conduct of this research.

Abbreviations:

VF	Visual Field
DLM	Deep Learning Model
OCT	Optical Coherence Tomography
MD	Mean Deviation
RNFL	Retinal Nerve Fiber Layer
ViT	Vision Transformer
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
AUC	Area Under the Curve
BDVA	Best Documented Visual Acuity
IOP	Intraocular Pressure
MAE	Mean Absolute Error

References

1. Chauhan BC, Malik R, Shuba LM, Rafuse PE, Nicolela MT, Artes PH. Rates of glaucomatous visual field change in a large clinical population. Invest Ophthalmol Vis Sci. 2014;55(7):4135–4143. doi:10.1167/iovs.14-14643 [PubMed: 24917147]
2. Shuldiner SR, Boland MV, Ramulu PY, et al. Predicting eyes at risk for rapid glaucoma progression based on an initial visual field test using machine learning. PLOS ONE. 2021;16(4):e0249856. doi:10.1371/journal.pone.0249856 [PubMed: 33861775]
3. Hirooka K, Sato S, Nitta E, Tsujikawa A. The Relationship Between Vision-related Quality of Life and Visual Function in Glaucoma Patients. J Glaucoma. 2016;25(6):505–509. doi:10.1097/IJG.0000000000000372 [PubMed: 26766401]
4. Odden JL, Mihailovic A, Boland MV, Friedman DS, West SK, Ramulu PY. Assessing Functional Disability in Glaucoma: The Relative Importance of Central Versus Far Peripheral Visual Fields. Invest Ophthalmol Vis Sci. 2020;61(13):23. doi:10.1167/iovs.61.13.23
5. Chauhan BC, Garway-Heath DF, Goñi FJ, et al. Practical recommendations for measuring rates of visual field change in glaucoma. Br J Ophthalmol. 2008;92(4):569–573. doi:10.1136/bjo.2007.135012 [PubMed: 18211935]

6. Wen JC, Lee CS, Keane PA, et al. Forecasting future Humphrey Visual Fields using deep learning. PLOS ONE. 2019;14(4):e0214875. doi:10.1371/journal.pone.0214875 [PubMed: 30951547]
7. Berchuck SI, Mukherjee S, Medeiros FA. Estimating Rates of Progression and Predicting Future Visual Fields in Glaucoma Using a Deep Variational Autoencoder. Sci Rep. 2019;9(1):18113. doi:10.1038/s41598-019-54653-6 [PubMed: 31792321]
8. Yousefi S, Goldbaum MH, Balasubramanian M, et al. Learning from data: recognizing glaucomatous defect patterns and detecting progression from visual field measurements. IEEE Trans Biomed Eng. 2014;61(7):2112–2124. doi:10.1109/TBME.2014.2314714 [PubMed: 24710816]
9. Mandrekar JN. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. Journal of Thoracic Oncology. 2010;5(9):1315–1316. doi:10.1097/JTO.0b013e3181ec173d [PubMed: 20736804]
10. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. ICLR. 2021.
11. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.
12. Yohannan J, Wang J, Brown J, et al. Evidence-based Criteria for Assessment of Visual Field Reliability. Ophthalmology. 2017;124(11):1612–1620. doi:10.1016/j.ophtha.2017.04.035 [PubMed: 28676280]
13. Mwanza JC, Kim HY, Budenz DL, et al. Residual and dynamic range of retinal nerve fiber layer thickness in glaucoma: comparison of three OCT platforms. Invest Ophthalmol Vis Sci. 2015;56:6344–6351. [PubMed: 26436887]
14. Sung MS, Heo H, Park SW. Structure-function Relationship in Advanced Glaucoma After Reaching the RNFL floor. J Glaucoma. 2019;28:1006–1011. [PubMed: 31567911]
15. Chauhan BC, Malik R, Shuba LM, Rafuse PE, Nicolela MT, Artes PH. Rates of glaucomatous visual field change in a large clinical population. Invest Ophthalmol Vis Sci. 2014;55(7):4135–4143. doi:10.1167/iovs.14-14643 [PubMed: 24917147]
16. Gardiner SK, Demirel S. Detecting change using standard global perimetric indices in glaucoma. Am J Ophthalmol. 2017;176:148–156. doi:10.1016/j.ajo.2017.01.013 [PubMed: 28130041]
17. Bak E, Lee KM, Kim M, Oh S, Kim SH. Angular Location of Retinal Nerve Fiber Layer Defect: Association With Myopia and Open-Angle Glaucoma. Invest Ophthalmol Vis Sci. 2020;61(11):13. doi:10.1167/iovs.61.11.13
18. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–845. [PubMed: 3203132]
19. Yousefi S, Kiwaki T, Zheng Y, et al. Detection of Longitudinal Visual Field Progression in Glaucoma Using Machine Learning. Am J Ophthalmol. 2018;193:71–79. doi:10.1016/j.ajo.2018.06.007 [PubMed: 29920226]
20. 1. Dixit A, Yohannan J, Boland MV. Assessing Glaucoma Progression Using Machine Learning Trained on Longitudinal Visual Field and Clinical Data. Ophthalmology. 2021;128(7):1016–1026. doi:10.1016/j.ophtha.2020.12.020 [PubMed: 33359887]
21. Muhammad H, Fuchs TJ, De Cuir N, et al. Hybrid Deep Learning on Single Wide-field Optical Coherence tomography Scans Accurately Classifies Glaucoma Suspects. J Glaucoma. 2017;26(12):1086–1094. doi:10.1097/IJG.0000000000000765 [PubMed: 29045329]
22. Cerentini A, Welfer D, Cordeiro d'Ornellas M, Pereira Haygert CJ, Dotto GN. Automatic Identification of Glaucoma Using Deep Learning Methods. Stud Health Technol Inform. 2017;245:318–321. [PubMed: 29295107]
23. Montesano G, Quigley HA, Crabb DP. Improving the Power of Glaucoma Neuroprotection Trials Using Existing Visual Field Data. Am J Ophthalmol. 2021;229:127–136. doi:10.1016/j.ajo.2021.04.008 [PubMed: 33905747]
24. Weinreb RN, Kaufman PL. The glaucoma research community and FDA look to the future: a report from the NEI/FDA CDER Glaucoma Clinical Trial Design and Endpoints Symposium. Invest Ophthalmol Vis Sci. 2009;50(4):1497–1505. doi:10.1167/iovs.08-2843 [PubMed: 19321793]

25. Ploug T, Sundby A, Moeslund TB, Holm S. Population Preferences for Performance and Explainability of Artificial Intelligence in Health Care: Choice-Based Conjoint Survey. J Med Internet Res. 2021;23(12):e26611. doi:10.2196/26611 [PubMed: 34898454]

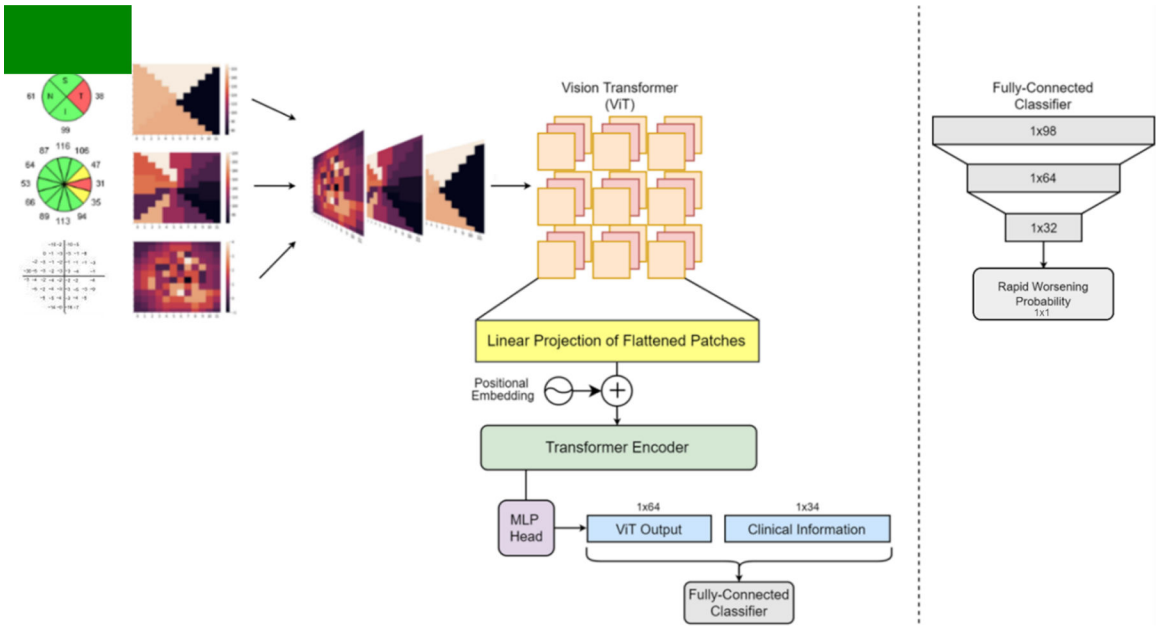

Figure 1.

An overview of our process and vision transformer model. The vision transformer architecture follows the same process as described by Dosovitskiy et al. 2021 in their original paper.[10]

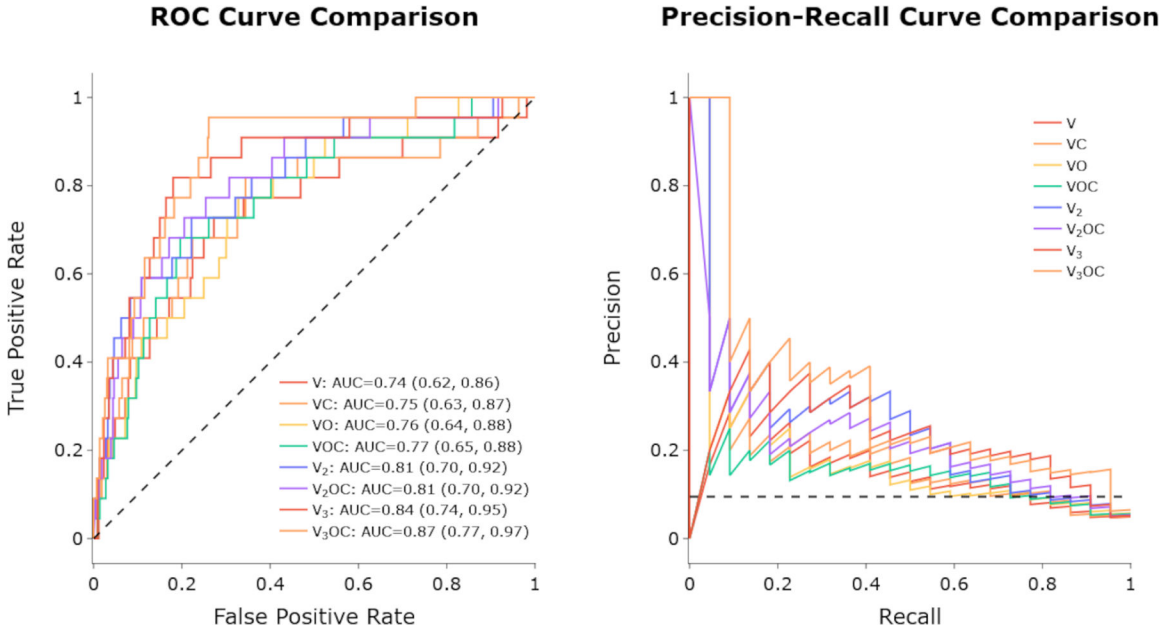

Figure 2.

ROC curves (left) and PR curves (right) of the different models. Curves are color-coded in a rainbow fashion, starting at red for V and up to violet for V_3OC.

Table 1.

Demographic, Clinical, Visual Field, and OCT Information of Eyes Included in Training, Validation, and Test Set

Demographic, clinical, VF, and OCT information of all eyes included in the study. While we report Mean (SD) here, distributions weren't always Gaussian and we discuss Median (IQR) for LogMAR, MD, Rate of MD Change, and PSD in the text.

	Overall	Non-Rapid Progressors	Rapid Progressors
Demographics			
No. of Eyes (%)	4536 (100%)	4273 (94.2%)	263 (5.80%)
Mean No. of VFs / Eye (SD)	9.82 (4.80)	9.95 (4.82)	7.76 (3.9)
Median No. of VFs / Eye (IQR)	8.00 (6.00,13.00)	9.00 (6.00,13.00)	6.00 (5.00,9.00)
No. of Patients (% Male)	2962 (43.6%)	2808 (43.3%)	239 (49.8%)
Mean Age of Patients - years (SD)	65.87 (12.40)	65.66 (12.40)	69.30 (11.94)
	Mean (SD)	**Mean (SD)**	**Mean (SD)**
Clinical			
IOP - mmHg	16.12 (4.64)	16.06 (4.56)	16.99 (5.75)
LogMAR Best Documented Visual Acuity	0.10 (0.15)	0.09 (0.14)	0.18 (0.23)
Visual Fields (at first VF unless specified)			
Mean Time Between VFs - years	1.00 (0.82)	1.01 (0.82)	0.80 (0.74)
Rate of MD Change (all VFs), dB / Year	−0.13 (0.74)	−0.03 (0.59)	−1.70 (1.08)
MD - decibels	−3.42 (4.29)	−3.15 (4.04)	−7.83 (5.61)
PSD - decibels	3.76 (3.23)	3.59 (3.12)	6.51 (3.77)
Test Duration minutes	5.74 (1.16)	5.68 (1.14)	6.59 (1.20)
Percent False Negative	3.31 (4.88)	3.14 (4.72)	5.97 (6.51)
Percent False Positive	2.74 (3.11)	2.71 (3.09)	3.11 (3.33)
OCT: Mean (SD)			
RNFL Thickness - μm	78.60 (28.69)	78.78 (28.74)	75.56 (27.72)
Vertical Cup-to-Disc Ratio	0.64 (0.15)	0.63 (0.15)	0.72 (0.14)
OCT Signal Strength	0.77 (0.10)	0.77 (0.10)	0.75 (0.09)

Note: Rapid Progressor is defined as visual field Mean Deviation slope < −1dB/yr.

SD = Standard Deviation; MD = Mean Deviation; VF = Visual Field; FN = False Negative; FP = False Positive, RNFL = Retinal Nerve Fiber Layer

Table 2.

Comparison of Model Results

A comparison of all model performance statistics across the different models. Youden's J-Index, Precision, Recall, and F1-Score were calculated only for the positive class, i.e. for rapid progressors only. Sensitivity, Specificity, and Youden's were calculated at the optimal point for an equally weighted sensitivity and specificity using Youden's J-Index. Precision and Recall were calculated for maximum F1-Score.

	Model V	**Model VC**	**Model VO**	**Model VOC**	**Model V_2**	**Model V_2OC**	**Model V3**	**Model V3OC**
AUC (95% CI)	0.74 (0.62, 0.86)	0.75 (0.63, 0.87)	0.76 (0.64, 0.88)	0.77 (0.65, 0.88)	0.81 (0.70, 0.92)	0.81 (0.70, 0.92)	0.84 (0.74, 0.95)	0.87 (0.77, 0.97)
Sensitivity (95% CI)	0.73 (0.68, 0.77)	0.82 (0.78, 0.85)	0.77 (0.73, 0.81)	0.68 (0.64, 0.72)	0.73 (0.68, 0.77)	0.73 (0.68, 0.77)	0.82 (0.78, 0.85)	0.95 (0.93, 0.97)
Specificity (95% CI)	0.73 (0.68, 0.77)	0.66 (0.61, 0.70)	0.67 (0.63, 0.72)	0.80 (0.76, 0.84)	0.78 (0.74, 0.82)	0.79 (0.75, 0.83)	0.82 (0.78, 0.85)	0.74 (0.70, 0.78)
Youden's J-Index	0.45	0.47	0.44	0.49	0.51	0.52	0.64	0.69
Precision	0.20	0.21	0.18	0.16	0.33	0.26	0.32	0.39
Recall	0.41	0.45	0.45	0.55	0.45	0.45	0.41	0.41
F1-Score	0.27	0.29	0.25	0.25	0.38	0.33	0.36	0.40

Table 3.

Intermodel AUC Comparisons (p-values)

A comparison of AUC scores between models to determine if performance differences were statistically significant using the DeLong Test.[20]

	Model V	**Model VC**	**Model VO**	**Model VOC**	**Model V_2**	**Model V_2OC**	**Model V_3**	**Model V_3OC**
Model V		0.56	0.33	0.36	0.03	0.17	0.01	0.005
Model VC			0.78	0.46	0.13	0.25	0.03	0.006
Model VO				0.72	0.09	0.26	0.03	0.005
Model VOC					0.23	0.33	0.06	0.003
Model V_2						0.91	0.29	0.08
Model V_2OC							0.47	0.07
Model V_3								0.26
Model V_3OC								