tvst

Artificial Intelligence

Pointwise Visual Field Estimation From Optical Coherence Tomography in Glaucoma Using Deep Learning

Ruben Hemelings[1,2], Bart Elen[2], João Barbosa-Breda[1,3,4], Erwin Bellon[5], Matthew B. Blaschko[6], Patrick De Boever[2,7,8], and Ingeborg Stalmans[1,9]

[1] Research Group Ophthalmology, Department of Neurosciences, KU Leuven, Leuven, Belgium
[2] Unit Health, Flemish Institute for Technological Research (VITO), Mol, Belgium
[3] Cardiovascular R&D Center – UnIC@RISE, Department of Surgery and Physiology, Faculty of Medicine of the University of Porto, Porto, Portugal
[4] Department of Ophthalmology, Centro Hospitalar e Universitário São João, Porto, Portugal
[5] Department of Information Technology, University Hospitals Leuven, Leuven, Belgium
[6] ESAT-PSI, KU Leuven, Leuven, Belgium
[7] Center for Environmental Sciences, Faculty of Industrial Engineering, Hasselt University, Diepenbeek, Belgium
[8] Department of Biology, University of Antwerp, Wilrijk, Belgium
[9] Ophthalmology Department, UZ Leuven, Leuven, Belgium

Correspondence: Ruben Hemelings, Vito Health, Industriezone Vlasmeer 7, 2400 Mol, Belgium. e-mail: ruben.hemelings@kuleuven.be

Received: December 20, 2021
Accepted: July 4, 2022
Published: August 23, 2022

Keywords: structure–function; visual field; optical coherence tomography; deep learning; convolutional neural network; glaucoma

Citation: Hemelings R, Elen B, Barbosa-Breda J, Bellon E, Blaschko MB, De Boever P, Stalmans I. Pointwise visual field estimation from optical coherence tomography in glaucoma using deep learning. Transl Vis Sci Technol. 2022;11(8):22, https://doi.org/10.1167/tvst.11.8.22

Purpose: Standard automated perimetry is the gold standard to monitor visual field (VF) loss in glaucoma management, but it is prone to intrasubject variability. We trained and validated a customized deep learning (DL) regression model with Xception backbone that estimates pointwise and overall VF sensitivity from unsegmented optical coherence tomography (OCT) scans.

Methods: DL regression models have been trained with four imaging modalities (circumpapillary OCT at 3.5 mm, 4.1 mm, and 4.7 mm diameter) and scanning laser ophthalmoscopy en face images to estimate mean deviation (MD) and 52 threshold values. This retrospective study used data from patients who underwent a complete glaucoma examination, including a reliable Humphrey Field Analyzer (HFA) 24-2 SITA Standard (SS) VF exam and a SPECTRALIS OCT.

Results: For MD estimation, weighted prediction averaging of all four individuals yielded a mean absolute error (MAE) of 2.89 dB (2.50–3.30) on 186 test images, reducing the baseline by 54% (MAEdecr%). For 52 VF threshold values' estimation, the weighted ensemble model resulted in an MAE of 4.82 dB (4.45–5.22), representing an MAEdecr% of 38% from baseline when predicting the pointwise mean value. DL managed to explain 75% and 58% of the variance (R^2) in MD and pointwise sensitivity estimation, respectively.

Conclusions: Deep learning can estimate global and pointwise VF sensitivities that fall almost entirely within the 90% test–retest confidence intervals of the 24-2 SS test.

Translational Relevance: Fast and consistent VF prediction from unsegmented OCT scans could become a solution for visual function estimation in patients unable to perform reliable VF exams.

Introduction

Glaucoma causes retinal ganglion cell (RGC) loss, resulting in structural and functional changes in the visual system. Standard automated perimetry (SAP) is the reference technique to follow functional visual field (VF) loss during glaucoma management.[1,2] Current SAP devices such as the Humphrey Field Analyzer (HFA; Carl Zeiss Meditec, Dublin, CA, USA) have high intrasubject variability and a lengthy examination time.[3,4] Retinal nerve fiber layer (RNFL) thickness

measurements in circumpapillary optical coherence tomography (OCT) scans are a widely accepted surrogate for quantitatively assessing structural retinal damage in patients with glaucoma.[5] However, OCT-measured retinal layer thinning is still far from being an exact quantification of RGC loss.

In a quest for enhanced glaucoma management, the structure–function relationship has been and still is an intensively studied topic.[6–9] Reproducible functional damage typically only becomes noticeable with current SAP when extensive RGC loss has occurred due to SAP measurement limitations.[10] Meanwhile, in advanced disease, SAP tends to detect progressive damage better than RNFL thickness values due to a measurement floor.[11] Machine learning studies that use linear approaches and RNFL thickness to estimate VF values reported limited correspondence, with most proposed regression models relying on assumptions such as log transformation to predict decibel (dB) VF values.[12–14] Deep learning (DL) approaches overcome the need for data transformation because they can model nonlinear functions.[15–19] Initial DL on structure–function focused on the estimation of global VF indices such as mean deviation (MD)[15,16] from OCT-derived information such as RNFL thickness maps.[17–19] OCT-derived layer thinning presents only part of the rich information on retinal structure that raw OCT scans contain. Recent work has shown that unsegmented OCT as an input to DL models has merits in predicting glaucoma detection and VF damage.[20–22]

This study wanted to assess the potential of DL regression models to predict VF information from raw OCT scans collected in a real-life glaucoma clinic population. To achieve this goal, unsegmented SPECTRALIS OCT (Heidelberg Engineering, Heidelberg, Germany) scans were used to train and evaluate customized DL models that estimate both the visual field sensitivity threshold at each location (52 threshold values) and MD as measured by the HFA. A thorough analysis of factors that influence modeling performance is presented.

Methods

Data initially extracted comprised 1643 matched OCT–VF pairs corresponding to 998 eyes of 542 patients who visited the University Hospitals Leuven's glaucoma clinic between 2015 and 2019. This work is part of the larger study on "automated glaucoma detection with deep learning" (study number S60649), approved by the Ethics Committee Research UZ/KU Leuven in November 2017. Informed consent was waived because of the retrospective nature and because patient reidentification was impossible because the link between patient ID and study ID was removed upon data export. The research adhered to the tenets of the Declaration of Helsinki. Inclusion criteria were (1) the availability of a SPECTRALIS OCT (Heidelberg Engineering) scan using the Glaucoma Module Premium Edition (GMPE), containing one scanning laser ophthalmoscopy (SLO) en face image, 24 radial scans, and three circumpapillary rings, and (2) the results of an HFA3 exam with the strategy 24-2 SITA Standard (SS; 52 test points) obtained with the Humphrey Field Analyzer (model 850 v1.3.1.2; Carl Zeiss Meditec, Dublin, CA, USA). Multiple OCT–VF pairs per eye were allowed on the condition that they were generated at unique visits. We used the circumpapillary rings as these scans cover all nerve fibers that pass through the optic nerve, unlike the radial scans. The GMPE protocol generates the circumpapillary rings by averaging over 16 consecutive B-scans, resulting in high-quality scans. The SLO image served as an intrastudy benchmark to quantify the improved modeling performance using circumpapillary rings.

The 1643 OCT–VF pairs of 542 patients were allocated to train, validation, and test sets, accounting for 60%, 20%, and 20% of the patients, respectively. We took care that all data from a single patient were stored in the same partition to avoid overestimating performance. To achieve this, random splitting was performed on anonymized patient ID instead of individual data points. Subsequently, subsets of the VF data of the validation and test data were selected on standard HFA reliability indices limits set by the manufacturer, with false positives (FPs) and fixation losses not exceeding 15% and 20%, respectively.[23] OCT–VF pairs were not excluded based on false negatives, as this is intrinsically linked with the level of glaucomatous VF damage.[24] We did not exclude unreliable visual field data in the training set but used them as data augmentation (OCT images with noisy visual field labels). Statistical group differences between train, validation, and test sets were tested using a χ^2 test at an α of 0.05.[25] Unlike related work, eyes featuring conditions like cataracts and retinal diseases that could influence VF testing were not excluded from the data set. The rationale behind this is to develop a VF estimation system that thrives in all cases encountered at a tertiary glaucoma clinic. We assessed the influence of including these cases in a post hoc sensitivity analysis.

OCT data were extracted in RAW format using Heyex v6.12.1 software (Heidelberg Engineering). The binary files were subsequently processed with heyexReader v0.1.3, a Python package for reading

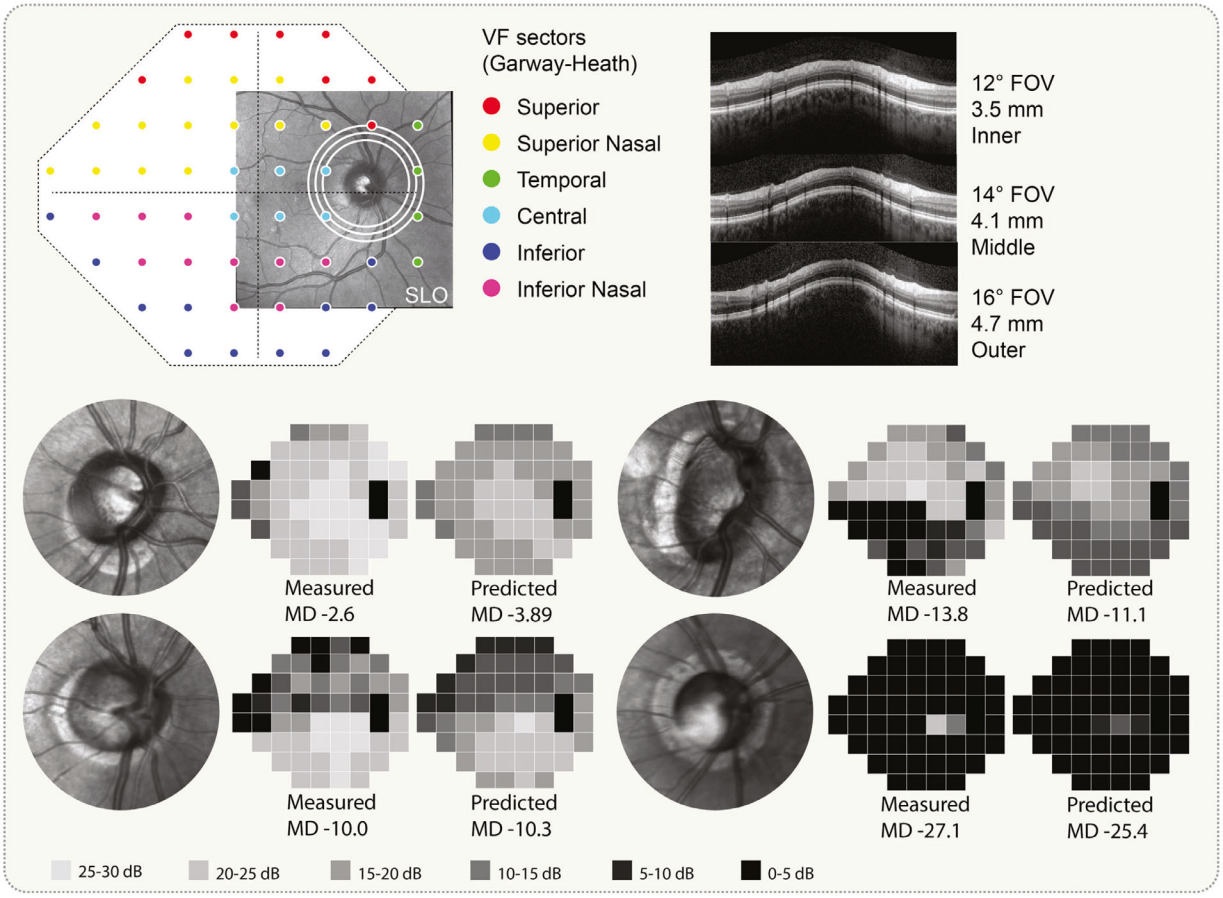

Figure 1. *Top panel*: Overview of imaging modalities, the spatial relationship between structure and analyzed VF points, allocated to Garway-Heath sectors. The 30° SLO covers less than half of VF test locations, still considerably more than the three circumpapillary OCT scans (*white circles*) displayed on the right. *Bottom panel*: Four cases of the independent test set. Each case features (1) an ONH zoom of the original 30° SLO image, (2) measured VF map and MD, and (3) the corresponding predicted VF map and MD. The displayed cases include an example of early glaucoma (*top left*), moderate glaucoma with loss in the superior hemifield (*bottom left*), a myopic eye with severe glaucomatous loss in the inferior hemifield (*top right*), and severe glaucoma with only a small central island remaining (*bottom right*).

Heyex OCT files. The three circumpapillary RNFL rings (3.5 mm, 4.1 mm, and 4.7 mm) and SLO were extracted as lossless image files with dimensions 768 × 496 and 1536 × 1536, respectively. We obtained VF data with HFA3 that were analyzed in PeriData v3.5.7 (PeriData Software GmbH, Hürth, Germany). Pointwise sensitivity threshold values were extracted from the individual patients' printouts using an optical character recognition (OCR) tool developed for this task. OCR output was manually verified on 10% of the data, matching perfectly with actual threshold values. These values were paired with global indices such as MD that were exported as a comma-separated value text file by PeriData. Two VF test points were discarded in all analyses, as these are on the anatomical blind spot (see Fig. 1), resulting in 52 threshold values to model.

Four DL models were trained using the 3.5 mm, 4.1 mm, and 4.7 mm circumpapillary rings and SLO images for MD estimation and an additional four DL models for 52 threshold values' estimation. We compared the single models with ensemble (weighted averaged) predictions. The optimal weights of the ensemble strategy were retrieved using a grid search strategy with a step size of 0.05. A total of 80 (4 models, 20 possible weight values) combinations were assessed on the validation set. We selected the Xception[26] architecture pretrained on ImageNet,[27] as this is a well-established convolutional neural network (CNN) that outperforms most ResNets while featuring fewer parameters. This study did not rely on transfer learning, with no model layers frozen throughout the whole training process. Publicly available standard CNN frameworks are typically trained on ImageNet for classification of 1000 classes and need to be altered for custom regression purposes. The Xception encoder was followed by a global average pooling and convolution operation, avoiding any fully connected layers to minimize overfitting.[28] The final convolution

operation had either 1 or 52 filters, depending on the target (MD or 52 threshold values), featuring a linear activation to allow for regression. Models were trained using mean squared error (MSE) loss, optimized using Adam[29] with a starting learning rate of $1e^{-4}$. The latter was reduced to 75% of its value after 10 epochs without improving the validation loss. Each epoch featured 300 training steps of batches containing four preprocessed and augmented images. This configuration struck a balance between computing memory constraints and ensuring the model encountered all training OCT–VF pairs in a single epoch (1200 data pairs per epoch). Single-channel circumpapillary rings were slightly upscaled to 768 × 512, adding 12 pixels in image height, to obtain image dimensions that are a multiple of 32 for optimal convergence. SLO images were downscaled to 512 × 512, as the original image dimensions were too memory intensive. The pixel intensity values were rescaled between 0 and 1 by a simple division operation. Augmentation included horizontal flipping, elastic deformation, and random erasing.[30] Model training and evaluation were performed using Keras[31] v2.2.4, Tensorflow[32] v1.12.0, in a Python 3.6.7 environment running on a server with six GTX 1080 Ti and two TITAN V graphics processing units.

The coefficient of determination (R^2), Pearson's r (r), mean absolute error (MAE), and MSE were calculated to evaluate model performances. R^2, MAE, and MSE were computed using the scikit-learn library[33] and r using the NumPy library.[34] The best model configurations (one for MD, one for threshold values) were selected on the highest R^2 metric and evaluated with the independent test set. We computed a baseline MAE for validation and test sets by predicting the mean MD or threshold value, equivalent to a model that obtains an R^2 score of zero. A fourth evaluation metric was defined as the reduction (if any) of the MAE baseline by the DL model, denoted as MAEdecr%. The latter represents a more interpretable value as opposed to MAE for interstudy comparison. The 95% confidence intervals (CIs) were obtained through bootstrap sampling (5000 iterations). The model trained on SLO images offered a way to assess the added value of OCT (with complete retinal layer information) over en face retinal imaging.

We analyzed MAEdecr% per individual VF point location and VF sector (Garway-Heath et al.[35]) to verify which VF regions are better modeled than others. In addition, MAE was also reported on the sector level, stratified by VF loss severity.[36] In an attempt to compare model performance with two previous studies on pointwise structure–function modeling,[12,14] the 90% CI VF sensitivity threshold predictions were compared against empirically established HFA 24-2 SS test–retest variability published by Artes et al.[3] Finally, we performed a sensitivity analysis to reveal what factors affected performance. Examined factors included OCT scan quality, history of retinal disease, signs of cataracts, VF reliability indices, and high myopia.

Results

Study Sample

Study sample characteristics are presented in Table 1. The average MD was −7.58 dB (ranging from −33.8 to +2.0 dB), which is expected considering the data from a tertiary glaucoma clinic. After filtering on standard reliability indices, 1390 OCT–VF pairs were eligible for model training and evaluation. All characteristics were similar between train, validation, and test sets, except for spherical equivalent ($P = 0.0468$). The baseline MAE values for MD and 52 pointwise sensitivity threshold estimation were 7.20 dB and 8.31 dB for the validation set and slightly lower for the test set (6.32 dB and 7.76 dB, respectively; see the last column of Table 2).

MD Estimation

The customized CNN model featuring an Xception encoder explained up to 72% (95% CI, 0.63–0.78) of the variance (= R^2) in the validation set of 198 OCT–VF pairs. The MAE of 3.40 dB obtained using the CNN trained on 3.5-mm (inner) scans reduced the baseline MAE by 53%. The performances of single models were similar between circumpapillary scan diameters (R^2 from 0.66 to 0.72, MAE from 3.76 to 3.40, $P > 0.05$). These OCT-trained models significantly outperformed the model trained with SLO images ($P < 0.05$), with the latter explaining 47% (0.33–0.58) of the variance and decreasing the MAE baseline by 34%. Normalized ensembling weights were 0.41, 0.12, 0.32, and 0.15. The weighted averaging of predictions of the four single models resulted in improved evaluation metrics: $R^2 = 0.75$ (+0.03), $r = 0.87$ (+0.02), MAE = 3.25 dB (−0.15), and MAEdecr% = 55% (+0.02). This ensemble model had similar results on the independent test set of 186 OCT–VF pairs (R^2 = 0.75, r = 0.87, MAE = 2.89 dB, MAEdecr% = 54%). Table 2 (second row of each cell) gives a detailed overview of MD estimation results. Figure 1 contains HFA-measured labels and CNN-predicted MD values of four cases of the test set spanning a variety of VF severity levels.

Table 1. Study Sample Characteristics

Characteristic	Train	Val	Test	Total
OCT–VF pairs, *n*	1006	198	186	1390
Eyes, *n*	598	137	131	866
Patients, *n*	325	84	88	497
Age, y	55.6 ± 20	57.1 ± 17	58.5 ± 17	55.8 ± 19
Sex, F/M	0.51/0.49	0.43/0.57	0.53/0.47	0.50/0.50
MD data available	1006/1006 (100)	198/198 (100)	186/186 (100)	1390/1390 (100)
MD, dB	−7.55 ± 7.55	−7.95 ± 8.92	−7.37 ± 7.93	−7.58 ± 7.80
MD ≥ −6 dB	600 (60)	121 (61)	108 (58)	829 (60)
−6 dB > MD > −12 dB	181 (18)	29 (15)	31 (17)	241 (17)
MD ≤ −12 dB	225 (22)	48 (24)	47 (25)	320 (23)
SphEq data available	799/1006 (79)	152/198 (77)	152/186 (84)	1103/1390 (79)
SphEq, D	−2.17 ± 2.73	−2.52 ± 2.72	−2.42 ± 3.17	−2.25 ± 2.79
+1 D ≤ SphEq	85 (11)	19 (13)	19 (13)	123 (11)
+1 D > SphEq > −1 D	190 (24)	13 (9)	36 (24)	239 (22)
−1 D ≥ SphEq > −6 D	441 (55)	105 (69)	74 (49)	620 (56)
−6 D ≥ SphEq	83 (10)	15 (10)	23 (15)	121 (11)
IOP data available	638/1006 (63)	119/198 (60)	125/186 (67)	882/1390 (63)
Max IOP, mm Hg	24.08 ± 8.25	24.16 ± 7.71	22.98 ± 7.57	23.94 ± 8.09
vCDR data available	882/1006 (88)	173/198 (87)	166/186 (91)	1221/1390 (88)
vCDR estimate	0.69 ± 0.22	0.69 ± 0.20	0.69 ± 0.20	0.69 ± 0.21
OCT scan quality	23.55 ± 4.36	23.82 ± 4.37	23.63 ± 4.22	23.60 ± 4.34

Values are presented as number (%) or mean ± SD unless otherwise indicated. D, diopters; IOP, intraocular pressure; SphEq, spherical equivalent; vCDR, vertical cup-disc ratio.

Pointwise Sensitivity Threshold Estimation

The model trained using 4.7-mm scans (outer) resulted in the best results on the validation set, explaining 57% (0.48–0.63) of the variance (R^2) across the 52 points. The models' performances were similar (R^2 from 0.54 to 0.57, MAE from 4.98 to 5.10, $P > 0.05$) among circumpapillary rings. The best single-scan model (4.7 mm) lowered the MAE baseline to 5.10

Table 2. Quantitative Results for All Models Trained for the Estimation of 52 Threshold Values (First Row of Each Cell) and MD (Second Row of Each Cell)

Modality	Target	R^2 (95% CI)	Pearson *r* (95% CI)	MSE (95% CI)	MAE (dB) (95% CI) (MAEdecr%)
Baseline (validation)	52 points	0.00	0.00	109.57 (93.29–127.23)	8.31 (7.67–8.99)
	MD	0.00	0.00	79.09 (62.25–97.50)	7.20 (6.49–7.96)
Inner, 3.5 mm	52 points	0.55 (0.46–0.62)	0.75 (0.70–0.80)	49.26 (41.49–57.52)	4.98 (4.53–5.45) (40)
	MD	0.72 (0.63–0.78)	0.85 (0.80–0.89)	22.24 (17.09–27.85)	3.40 (2.95–3.86) (53)
Middle, 4.1 mm	52 points	0.54 (0.45–0.61)	0.75 (0.70–0.80)	49.38 (41.96–57.40)	5.12 (4.68–5.57) (38)
	MD	0.66 (0.54–0.75)	0.83 (0.77–0.89)	26.61 (20.24–33.50)	3.76 (3.28–4.27) (48)
Outer, 4.7 mm	52 points	0.57 (0.48–0.63)	0.77 (0.72–0.82)	49.02 (39.83–54.41)	5.10 (4.72–5.51) (39)
	MD	0.70 (0.60–0.78)	0.84 (0.78–0.89)	23.54 (17.18–30.69)	3.42 (2.96–3.90) (53)
SLO	52 points	0.39 (0.31–0.46)	0.66 (0.59–0.71)	66.36 (55.52–77.50)	5.82 (5.28–6.39) (30)
	MD	0.47 (0.33–0.58)	0.70 (0.59–0.79)	41.52 (31.58–52.49)	4.77 (4.19–5.37) (34)
Circle scans (weighted average)	52 points	0.59 (0.51–0.65)	0.79 (0.74–0.83)	44.50 (37.62–51.87)	4.89 (4.48–5.30) (41)
	MD	0.74 (0.65–0.80)	0.86 (0.81–0.90)	20.72 (15.57–26.53)	3.27 (2.84–3.72) (55)
Circle scans, SLO (weighted average)	52 points	**0.59** (0.52–0.65)	**0.79** (0.75–0.83)	**44.02 (37.30–51.33)**	**4.88** (4.47–5.30) (**41**)
	MD	**0.75 (0.67–0.80)**	**0.87** (0.82–0.91)	**20.12 (15.07–25.39)**	**3.25** (2.83–3.70) (**55**)
Baseline (test)	52 points	0.00	0.00	101.59 (84.97–119.91)	7.76 (7.12–8.43)
	MD	0.00	0.00	62.48 (48.98–77.32)	6.32 (5.66–7.01)
Test set (186 images)	52 points	0.58 (0.51–0.63)	0.79 (0.75–0.82)	42.35 (36.21–49.93)	4.82 (4.45–5.22) (38)
	MD	0.75 (0.67–0.81)	0.87 (0.83–0.91)	15.73 (11.35–21.06)	2.89 (2.50–3.30) (54)

The first section features results on the validation set (198 images), for which the best results are set in bold. Best model setup on validation data was subsequently used to obtain results on the independent test set (186 images), selected on best R^2. MAE and MSE baseline for validation and test data were computed through the constant prediction of the mean value (threshold value, MD).

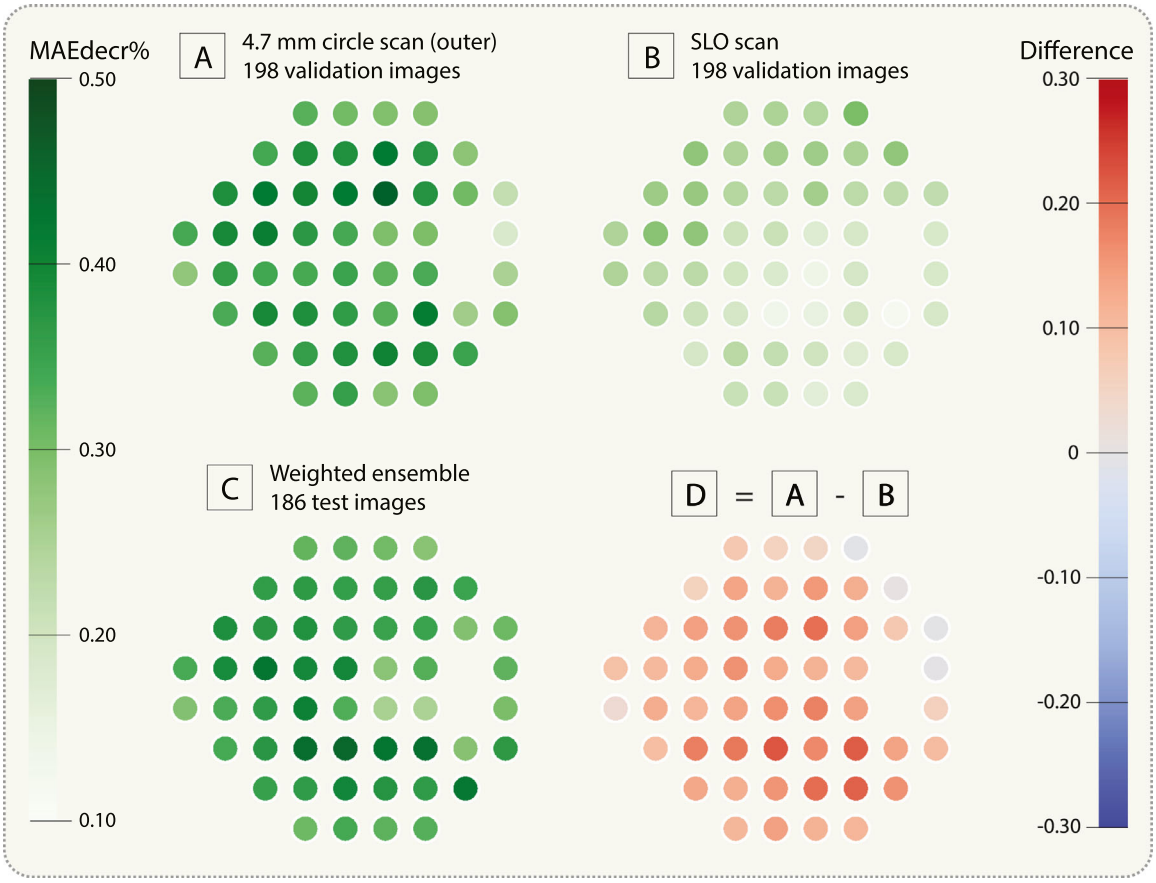

Figure 2. (A) MAEdecr% values for 52 VF threshold values obtained using the model trained on 4.7 mm (outer) OCT scans. MAEdecr% is the decrease in percentage from the baseline MAE, with the latter obtained when always predicting the pointwise mean. (B) Similar to panel A, but model trained using en face SLO images, to compare with as a baseline. (C) Final MAEdecr% values obtained on the test set, using the weighted averaged predictions of the four CNNs trained using OCT scans and SLO images. (D) The difference between panels A and B, indicating the superior VF modeling performance of OCT scans across the majority of VF test locations.

dB, representing a reduction of 39% (MAEdecr%). Similar to what was observed in the MD estimation experiments, the SLO-trained model reports significantly lower metrics ($R^2 = 0.39$, [0.31–0.46], $P < 0.05$). Normalized ensembling weights were 0.24, 0.17, 0.45, and 0.14. The weighted average of the predictions of the models scored $R^2 = 0.59$ (+0.02), $r = 0.79$ (+0.02), MAE = 4.88 dB (−0.10), and MAEdecr% = 41% (+0.01). Results on the test set were similar, with $R^2 = 0.58$, $r = 0.79$, MAE = 4.82 dB, and MAEdecr% = 38%.

Noticeable differences can be observed in Figure 2A when inspecting the individual threshold values. The model trained on 4.7-mm circle scans reached high MAEdecr% values (range, 19%–46%) for all 52 VF test points, with the highest values recorded in superior and inferior nasal VF sectors, corresponding to nasal step locations, and the lowest in the temporal VF, corresponding to the temporal wedge location. The SLO-trained model yielded lower MAEdecr% values (range, 12%–30%), especially in the inferior VF area (Fig. 2B). Figure 2D illustrates this contrast, with differences up to 22% recorded between the two models. The best model using the weighted averaged predictions of four models equivalently reached high MAEdecr% values on all 52 test points (range, 25%–45%; Fig. 2C). The lowest and highest values were recorded in the central and inferior nasal VF, respectively. These findings corroborate the sectoral analysis in Table 3. The MAE baseline was the most elevated in both inferior nasal and superior nasal VF sectors, indicating more variance in those locations. The model explained most of the variance in those sectors with R^2 equal to 0.64 and 0.60, respectively. Superior VF sector obtained the highest MAEdecr% (51%). The lowest variance was explained in the central sector ($R^2 = 0.52$).

Figure 3 visualizes the individual threshold prediction performance in the same boxplot style as Zhu et al.[12] and Guo et al.[14] The graph plots the

Table 3. Metrics on the Six Visual Field Sectors as Described by Garway-Heath et al.,[35] Computed on the Test Set Using the Weighted Ensemble Model

Sector	R^2	Pearson *r*	MAE (dB)	MAE Baseline (MAEdecr %)
Central	0.52 (0.46–0.57)	0.77 (0.72–0.81)	4.84 (4.33–5.39)	7.50 (35)
Temporal	0.55 (0.44–0.62)	0.77 (0.69–0.83)	4.41 (4.01–4.85)	6.30 (30)
Inferior	0.56 (0.46–0.63)	0.77 (0.70–0.82)	5.09 (4.66–5.55)	7.99 (36)
Inferior nasal	0.64 (0.57–0.70)	0.83 (0.78–0.87)	4.67 (4.20–5.17)	8.11 (42)
Superior	0.54 (0.45–0.62)	0.76 (0.70–0.81)	4.89 (4.48–5.32)	7.37 (51)
Superior nasal	0.60 (0.52–0.67)	0.80 (0.75–0.84)	4.79 (4.32–5.28)	8.09 (41)

MAE baseline was obtained by always predicting the sector threshold mean.

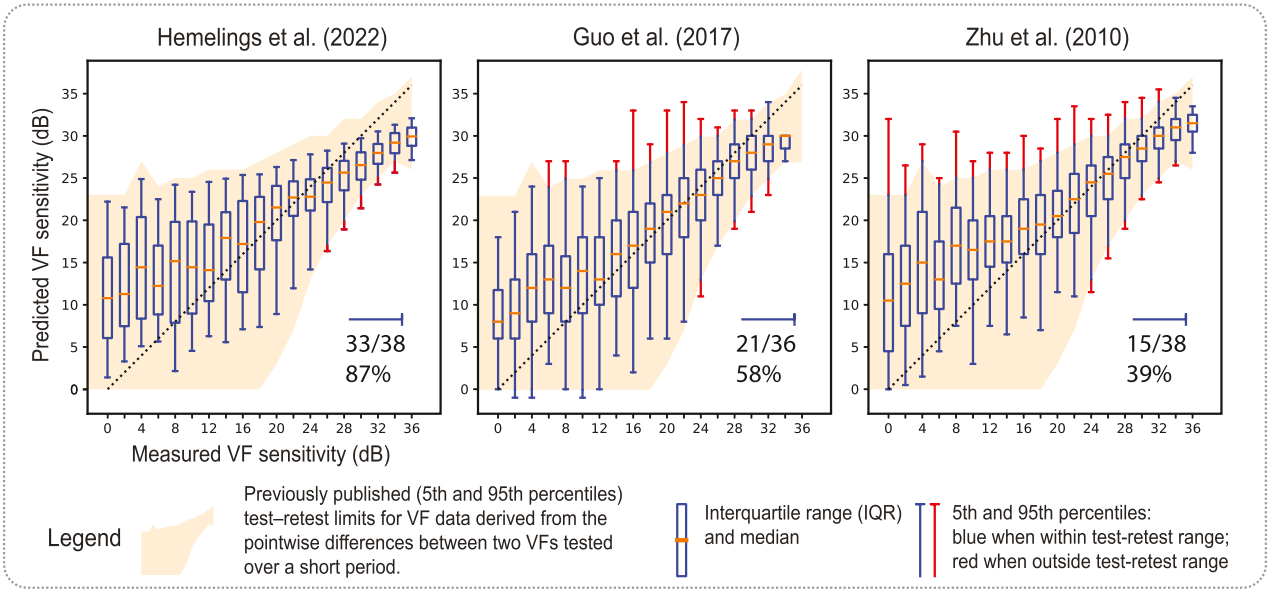

Figure 3. Comparative overview of three original studies (current, Guo et al.,[14] and Zhu et al.[12]) that report on the relationship between measured and predicted VF threshold values, stratified by sensitivity (step size of 2 dB). The error ranges obtained by our approach leveraging DL are smaller than previous non-DL studies. Thirty-three of 38 whiskers are located within the 90% CI test–retest limits reported by Artes et al.[3]

SAP-measured dB values against the predicted dB values at an interval of 2 dB. The largest prediction errors occurred in VF points with low sensitivity values in all three studies. However, the variability of predictions by our CNN was significantly more consistent with test–retest CI: 33 of 38 boxplot whiskers fall within the 90% CI determined by Artes et al.[3] Using a two-sample *z*-test for proportions yields a *P*-value of 0.00256, showing significant improvement over the previous result of 58% of whiskers within the shaded region.

The Garway-Heath VF sector with the largest MAE differed with VF severity level, as can be deducted from Table 4. Average MAE was largest in the superior VF sector (4.01 dB) for individuals with mild VF loss. The largest MAE can be found in the opposite inferior sector (5.19 dB) with moderate VF loss. Finally, advanced VF loss resulted in the central sector having the largest MAE (8.89 dB).

The sensitivity analysis presented in Table 5 reveals that the MAEdecr% remained mostly stable across subsets of the test set compared to the MAE value. The exclusion of examined factors did not lead to significant improvements in MAEdecr% values. MAEdecr% for MD improved the most when excluding high myopia or OCT data with scan quality inferior to 20 (MAEdecr% from 54% to 57%). For pointwise VF estimation, the highest MAEdecr% was obtained when filtering out OCT–VF pairs that feature an FP value of more than 10% (MAEdecr% from 38% to 39%).

Table 4. Pointwise MAE Aggregated on the Global Level and Six Visual Field Sectors as Described by Garway-Heath et al.,[35] Stratified by Three VF Severity Groups in the Test Set

	MAE		
Sector	Early VF Loss (MD ≥ -6 dB)	Moderate VF Loss (-6 dB $>$ MD >-12 dB)	Advanced VF Loss (MD ≤ -12 dB)
All	3.582	4.561	7.849
Central	3.126	4.640	**8.927**
Temporal	3.597	3.246	7.059
Inferior	3.762	**5.226**	8.047
Inferior nasal	3.235	3.850	8.512
Superior	**4.210**	4.576	6.672
Superior nasal	3.517	4.905	7.640

The largest MAE per severity group is highlighted in bold, indicating the best modeling performance by the ensembled CNN.

Table 5. Post Hoc One-at-a-Time Sensitivity Analysis to Assess Influence of Certain Input Factors on the Error Term for the Test Set

Subset	No. of OCT–VF Pairs	MD MAE, dB / MAEdecr%	52 Points MAE, dB / MAEdecr%
All (FL $\leq$20, FP $\leq$15)	186	2.89/54	4.82/38
Effect of other ocular disease			
Excluding high myopia ($\leq$–6 D)[a]	125	2.69/**57**	4.65/38
Excluding history of cataract	130	2.63/46	4.49/23
Excluding other types of glaucoma (other than POAG, NTG)	173	2.75/56	4.70/38
Excluding history of retinal diseases	161	2.88/53	4.95/34
Effect of OCT scan quality			
Excluding scans with quality $<$20	149	2.65/**57**	4.64/38
Excluding scans with quality $<$15	181	2.82/55	4.74/38
Effect of visual field reliability indices			
Excluding HFA FL $<$10	135	2.70/55	4.51/38
Excluding HFA FP $<$10	179	2.79/56	4.74/**39**
Excluding HFA FN $<$10	148	2.74/55	4.67/37

The best performance (largest MAEdecr%) per column is highlighted in bold. Baseline error is the MAE obtained when predicting the mean MD or pointwise value. FL, fixation loss; FN, false negative; FP, false positive; NTG, normal tension glaucoma; POAG, primary open angle glaucoma.

[a]Not all OCT–VF pairs have a SphEq label assigned; hence, omission of pairs might be due to missing values.

Discussion

This study is the first to regress all 24-2 VF sensitivity threshold values and MD from unsegmented OCT images of a real-life glaucoma clinic population. The weighted ensemble managed to explain 75% of the variance in MD estimation, on par with the current state of the art.[15,20] Pointwise 24-2 VF predictions fell almost completely (87% of whiskers) within the empirically determined 90% CIs of test–retest setups. Our data-driven modeling approach overcomes the need for retinal layer segmentation and prior assumptions on the structure–function relationship. Furthermore, the model displayed robustness against challenging cases that feature other conditions than glaucoma.

The advantages of omitting a mandatory retinal layer segmentation processing step are twofold. First, it alleviates potential segmentation errors because of bad scan quality[37,38] or critically thinned RNFL (floor effect) in severe glaucoma cases. Second, the models allow the extraction of relevant information from

other RNFL parameters and retinal layers besides the commonly used RNFL thickness values. Previous work hinted that OCT reflectance data might be more informative for glaucoma than conventional RNFL thickness values.[39,40] Christopher et al.[15] verified that original voxel information from the RNFL layer resulted in CNN models with higher MD estimation performance than models trained using RNFL thickness values (0.70 and 0.63 in R^2 score, respectively). Our SLO-trained model for MD estimation gave similar results to the SLO model in Christopher et al.[15] ($R^2 = 0.48$ [0.41–0.54] vs. our $R^2 = 0.47$ [0.33–0.58]). Our best model (weighted ensemble of four models trained on three types of circumpapillary rings and SLO) explains 75% (0.67–0.81) of the MD variance in the test set, whereas the best setup from Christopher et al.[15] using average RNFL OCT voxel intensity explains 70% (0.64–0.74). A formal interstudy comparison is not possible because the evaluation metrics depend on data set characteristics such as sample MD (−7.6 vs. −5.2 dB in their glaucoma subset).

Another recent study by Yu et al.[16] describes the prediction of global VF indices using a three-dimensional (3D) CNN that takes the complete volumetric OCT scans of the ONH (optic nerve head) and macula as inputs. The authors report a Pearson correlation coefficient of 0.86 (0.83–0.89) for MD, which is comparable to the correlation of 0.87 (0.83–0.91) of our ensemble model. Again, direct comparison is difficult because of different data (sample MD of −2.1 dB in their study). Furthermore, the use of correlation metrics provides no indication of prediction error but aims to quantify the linearity between the VF measurements and predictions. The high memory demands of 3D CNNs forced the authors to compromise on OCT scan resolution: the original cubes were downsized to 32% of their original size. By doing so, they introduced a risk of unintentionally removing fine-grain structural features. Our two-dimensional CNN setup preserved the original image width of all OCT scans (768 A-scans).

Wong et al.[41] compared several machine learning approaches for the estimation of MD from RNFL thickness. They obtain MAEdecr% rates up to 26% using gradient-boosted trees in an external test set, which is half of the MAEdecr% reported in the current study (54%). Of note, their baseline MAE is 4.09 dB, whereas the test set described here features a baseline MAE of 6.32 dB.

We report a competitive R^2 of 58% on individual threshold values of HFA 24-2 SITA Standard VF exams from OCT data. Similar to MD analyses, OCT-trained models on threshold values significantly outperformed their SLO counterpart. Although not significant, the circumpapillary scans with a larger diameter explained more variance in the validation data, with R^2 increasing from 0.54 to 0.57. This is plausible because OCT data at 16° intersection potentially offer more information on individual RGC axons as they lie further apart from each other with increasing distance from the ONH. Combining the four models through weighted prediction averaging gave a correlation of five percentage points higher than the 0.74 reported by Guo et al.[14] The latter authors had their results using nine-field OCT data covering 60° of the retina, whereas an area of 16° around the ONH was sufficient in our case. The three main differences between our approach and that of Guo et al.[14] are (1) the use of CNNs versus support vector machines (SVMs), (2) raw OCT scans versus thickness values of RGC complex layers, and (3) a larger study sample (863 vs. 86 eyes). The recent study by Park et al.[18] describes a similar approach featuring RGC complex layer segmentation. However, they employed an Inception-v3 CNN instead of an SVM to predict the 24-2 VF map, reporting a root mean squared error (RMSE) of 4.79 dB across the 24-2 map. A follow-up study by Shin and colleagues[42] investigated the advantage of thickness maps generated by swept-source OCT (SS-OCT) versus spectral domain OCT (SD-OCT). RMSE was significantly lower at 4.51 dB for SS-OCT when compared to 5.29 dB for SD-OCT. Finally, Lazaridis et al.[43] recently developed an ensemble model for VF estimation that features both a CNN as well as a variational autoencoder. In their study, incorporation of a downsized raw circumpapillary ring, in addition to RNFL thickness information, decreased the MAE by 22%. This adds further evidence that raw OCT information holds unique information relevant to pointwise VF estimation.

The ensemble model could model specific VF points and sectors better than others. We recorded three out of the five lowest MAEdecr% values in the temporal VF sector in the validation set, while the five best were all in the superior nasal VF sector. These findings match the superior nasal step scotoma location that is typically affected early in glaucoma development.[44] Lower performance in both central and temporal VF sectors could be due to damage that occurs solely in later disease stages. This result is in line with the findings regarding the sectors featuring the lowest MAE baseline (Table 3), showing lower variance in ground-truth threshold values. Christopher et al.[15] predicted sectoral pattern deviation (which is derived from threshold values) using DL approaches, equally getting the best performance in superior nasal VF ($R^2 = 0.67$) and inferior nasal VF ($R^2 = 0.60$).

This study did not exclude eyes with nonglaucomatous conditions that could influence VF results. The main application was to construct a model that can be used in the glaucoma clinic, complementary to existing perimetry solutions. In our view, the best way to assess its feasibility is by using real-life hospital data that impose several challenges next to glaucoma. We exposed the CNN to training data that contain VF loss potentially resulting from nonglaucomatous conditions such as cataracts and retinal disease. This is different from related work, in which nonglaucomatous VF loss is typically excluded from analysis. Even with these additional challenges presented in our work, model performance is extremely high (R^2 of 0.75 and 0.59 for MD and 52 threshold values, respectively), adding proof that the combination of raw OCT data and data-driven CNNs fosters good potential in automated VF estimation. The sensitivity analysis given in Table 5 revealed the robustness of the CNN in challenging cases.

VF testing suffers from intrasubject variability, complicating the diagnosis of glaucoma progression.[45] The best way of assessing VF reliability is through test–retest setups. In the Ocular Hypertension Treatment Study, VF abnormalities were not confirmed in 86% of the original reliable VF exams.[46] As Guo et al.[14] state, it becomes harder to assess actual performance improvements in VF modeling from OCT, given that the ground-truth VF is noisy. Lazaridis and colleagues[43] reported an R^2 of 88% between single VF sensitivity values and a median VF sensitivity computed over up to 10 visits within 3 months. Artes et al.[3] computed repeat VF on 49 glaucomatous eyes, and they published 5th and 95th limits for VF threshold values. Their confidence intervals provide additional evidence that VF points with lower recorded dB values hold more variability than those with higher dB values. Two studies on 24-2 VF estimation from OCT provide a comparison of measured versus predicted VF threshold values, which can be placed next to the empirical 90% CI of Artes et al.[3] The recent DL study by Lazaridis et al.[43] was not included in the comparison as their boxplots from Figure 2D used an α level of 0.05. Figure 3 showcases the prediction variability for all three studies (current, Guo et al.,[14] and Zhu et al.[12]). The 90% CI in the current study shows that no systematic overprediction of dB values occurred, with 33 of 38 whiskers (87%) falling within the shaded area. This represents a significant improvement over the previous result of 58% by Guo et al.[14] ($P = 0.00256$). The interquartile ranges of the boxplots for threshold values smaller than 10 dB seem larger in the current study than the ones in Guo et al.,[14] which is most likely because of the challenging sample in our study. These results confirm that future performance improvement in the current model will be hard to detect, as almost all predictions fall within the empirically determined VF ranges. In such a context of noisy ground truth, it is preferable to adhere to the empirically determined CI instead of focusing on MAE.

This study comes with strengths and limitations. We have trained and validated our method on data sourced from the same hospital, a single type of OCT device for a single type of VF exam. We should further validate our trained models on external OCT–VF data available in the public domain. We envisage this will soon be possible considering the widespread interest in DL in glaucoma management.[47] In this study, we have taken all precautionary measures to prevent overfitting: no fully connected layers in CNN and a single use of the independent test set. Next to unknown generalizability, we provide no analysis on explainability. The move from OCT segmentation parameters to complete data-driven modeling eliminates the risk of segmentation errors but comes at the cost of model decision transparency. Our one-at-a-time sensitivity analysis provided additional insights on the importance of study sample data but did not allow for interaction between factors. Finally, we did not automatically estimate pattern standard deviation (PSD), a map that highlights localized scotomas by accounting for generalized VF loss. PSD is deemed extremely relevant in glaucoma management, as clinicians can focus on VF loss related to the neurodegenerative disease. PSD was not explicitly modeled in the current study, as this information can be obtained using the patient's predicted raw sensitivities and age, comparable to current perimetry solutions.

Deep learning can estimate global and pointwise VF sensitivities that fall almost entirely within the 90% test–retest confidence intervals of the 24-2 SS test. Fast and consistent VF prediction from unsegmented OCT could become a surrogate solution for visual function estimation in patients who cannot perform reliable VF exams.

Acknowledgments

Supported by the Research Group Ophthalmology, KU Leuven and VITO NV (to RH) and the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program. No outside entities have been involved in the study design; in the collection, analysis, and interpretation of data; in the writing of the manuscript; or in the decision to submit the manuscript for publication.

This article was presented at the Association for Research in Vision and Ophthalmology Annual Meeting (ARVO2021) virtual conference, May 1st to May 7th, 2021.

Disclosure: **R. Hemelings**, None; **B. Elen**, None; **J. Barbosa-Breda**, None; **E. Bellon**, None; **M.B. Blaschko**, None; **P. De Boever**, None; **I. Stalmans**, None

References

1. Prum BE, Rosenberg LF, Gedde SJ, et al. Primary Open-Angle Glaucoma Preferred Practice Pattern Guidelines. *Ophthalmology*. 2016;123(1):P41–P111.
2. European Glaucoma Society. European Glaucoma Society Terminology and Guidelines for Glaucoma, 4th Edition—Part 1. *Br J Ophthalmol*. 2017;101(4):54.
3. Artes PH, Iwase A, Ohno Y, Kitazawa Y, Chauhan BC. Properties of perimetric threshold estimates from full threshold, SITA standard, and SITA fast strategies. *Invest Ophthalmol Vis Sci*. 2002;43(8):2654–2659.
4. Gardiner SK, Swanson WH, Goren D, Mansberger SL, Demirel S. Assessment of the reliability of standard automated perimetry in regions of glaucomatous damage. *Ophthalmology*. 2014;121(7):1359–1369.
5. Banegas SA, Antón A, Morilla A, et al. Evaluation of the retinal nerve fiber layer thickness, the mean deviation, and the visual field index in progressive glaucoma. *J Glaucoma*. 2016;25(3):e229–235.
6. Gardiner SK, Johnson CA, Cioffi GA. Evaluation of the structure-function relationship in glaucoma. *Invest Ophthalmol Vis Sci*. 2005;46(10):3712–3717.
7. Ferreras A, Pablo LE, Garway-Heath DF, Fogagnolo P, García-Feijoo J. Mapping standard automated perimetry to the peripapillary retinal nerve fiber layer in glaucoma. *Invest Ophthalmol Vis Sci*. 2008;49(7):3018–3025.
8. Leite MT, Zangwill LM, Weinreb RN, Rao HL, Alencar LM, Medeiros FA. Structure-function relationships using the Cirrus spectral domain optical coherence tomograph and standard automated perimetry. *J Glaucoma*. 2012;21(1):49–54.
9. Malik R, Swanson WH, Garway-Heath DF. 'Structure–function relationship' in glaucoma: past thinking and current concepts. *Clin Experiment Ophthalmol*. 2012;40(4):369–380.
10. Kerrigan-Baumrind LA, Quigley HA, Pease ME, Kerrigan DF, Mitchell RS. Number of ganglion cells in glaucoma eyes compared with threshold visual field tests in the same persons. *Invest Ophthalmol Vis Sci*. 2000;41(3):741–748.
11. Medeiros FA, Zangwill LM, Bowd C, Mansouri K, Weinreb RN. The structure and function relationship in glaucoma: implications for detection of progression and measurement of rates of change. *Invest Ophthalmol Vis Sci*. 2012;53(11):6939–6946.
12. Zhu H, Crabb DP, Schlottmann PG, et al. Predicting visual function from the measurements of retinal nerve fiber layer structure. *Invest Ophthalmol Vis Sci*. 2010;51(11):5657–5666.
13. Zhang X, Bregman CJ, Raza AS, De Moraes G, Hood DC. Deriving visual field loss based upon OCT of inner retinal thicknesses of the macula. *Biomed Opt Express*. 2011;2(6):1734–1742.
14. Guo Z, Kwon YH, Lee K, et al. Optical coherence tomography analysis based prediction of Humphrey 24-2 visual field thresholds in patients with glaucoma. *Invest Ophthalmol Vis Sci*. 2017;58(10):3975–3985.
15. Christopher M, Bowd C, Belghith A, et al. Deep learning approaches predict glaucomatous visual field damage from OCT optic nerve head en face images and retinal nerve fiber layer thickness maps. *Ophthalmology*. 2020;127(3):346–356.
16. Yu HH, Maetschke SR, Antony BJ, et al. Estimating global visual field indices in glaucoma by combining macula and optic disc OCT scans using 3-dimensional convolutional neural networks. *Ophthalmol Glaucoma*. 2021;4(1):102–112.
17. Hashimoto Y, Asaoka R, Kiwaki T, et al. Deep learning model to predict visual field in central 10° from optical coherence tomography measurement in glaucoma. *Br J Ophthalmol*. 2021;105(4):507–513.
18. Park K, Kim J, Lee J. A deep learning approach to predict visual field using optical coherence tomography. *PLoS ONE*. 2020;15(7):e0234902.
19. Mariottoni EB, Datta S, Dov D, et al. Artificial intelligence mapping of structure to function in glaucoma. *Transl Vis Sci Technol*. 2020;9(2):19.
20. Christopher M, Proudfoot JA, Bowd C, et al. Deep learning models based on unsegmented OCT RNFL circle scans provide accurate detection of glaucoma and high resolution prediction of visual field damage. *Invest Ophthalmol Vis Sci*. 2020;61(7):1439.
21. Hemelings R, Elen B, Barbosa Breda J, Blaschko MB, De Boever P, Stalmans I. Convolutional neural network predicts visual field threshold values from optical coherence tomography scans. *Invest Ophthalmol Vis Sci*. 2021;62(8):1022.

22. Thompson AC, Jammal AA, Berchuck SI, Mariottoni EB, Medeiros FA. Assessment of a segmentation-free deep learning algorithm for diagnosing glaucoma from optical coherence tomography scans. *JAMA Ophthalmol*. 2020; 138(4):333–339.
23. Humphrey Field Analyzer 3 (HFA3). Instructions for Use - Models 830, 840, 850, 860, 2660021166-131 *Rev A*. 2018-11, Section 6-12, Manufacturer: Zeiss; 88, https://www.manualslib.com/manual/1548100/Zeiss-Humphrey-Field-Analyzer-3.html#manual. Accessed 15 november 2021.
24. Bengtsson B, Heijl A. False-negative responses in glaucoma perimetry: indicators of patient performance or test reliability? *Invest Ophthalmol Vis Sci*. 2000;41(8):2201–2204.
25. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *J Am Stat Assoc*. 1952;47(260):583–621.
26. Chollet F. Xception: Deep learning with depthwise separable convolutions. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu; 2017:1800–1807, doi:10.1109/CVPR.2017.195.
27. Deng J, Dong W, Socher R, Li LJ, L Kai, F-F Li. ImageNet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE; 2009:248–255.
28. Xu Q, Zhang M, Gu Z, Pan G. Overfitting remedy by sparsifying regularization on fully-connected layers of CNNs. *Neurocomputing*. 2019;328:69–74.
29. Kingma D, Ba J. Adam a method for stochastic optimization. 2015, *ArXiv14126980 Cs*, http://arxiv.org/abs/1412.6980. Accessed May 17, 2020.
30. Zhong Z, Zheng L, Kang G, Li S, Yang Y. Random erasing data augmentation. Published online November 16, 2017, doi:10.48550/arXiv.1708.04896.
31. Chollet F and others. Keras. GitHub. 2015, https://github.com/fchollet/keras.
32. Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. *ArXiv160304467 Cs*, http://arxiv.org/abs/1603.04467. Accessed August 22, 2020.
33. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
34. Harris CR, Millman KJ, SJ Walt, et al. Array programming with NumPy. *Nature*. 2020;585(7825): 357–362.
35. Garway-Heath DF, Poinoosawmy D, Fitzke FW, Hitchings RA. Mapping the visual field to the optic disc in normal tension glaucoma eyes. *Ophthalmology*. 2000;107(10):1809–1815.
36. Hodapp E, Parrish RK, Anderson DR. Clinical decisions in glaucoma. 1st ed. Mosby-Year Book: St. Louis; In: 1993.
37. Asrani S, Essaid L, Alder BD, Santiago-Turla C. Artifacts in spectral-domain optical coherence tomography measurements in glaucoma. *JAMA Ophthalmol*. 2014;132(4):396–402.
38. Liu Y, Simavli H, Que CJ, et al. Patient characteristics associated with artifacts in Spectralis optical coherence tomography imaging of the retinal nerve fiber layer in glaucoma. *Am J Ophthalmol*. 2015;159(3):565–576.e2.
39. Belghith A, Medeiros FA, Bowd C, Weinreb RN, Zhuowen T, Zangwill LM. A novel texture-based OCT enface image to detect and monitor glaucoma. *Invest Ophthalmol Vis Sci*. 2016;57(12).
40. Leung CKS. Retinal nerve fiber layer (RNFL) optical texture analysis (ROTA) for evaluation of RNFL abnormalities in glaucoma. *Invest Ophthalmol Vis Sci*. 2018;59(9):3497.
41. Wong D, Chua J, Bujor I, et al. Comparison of machine learning approaches for structure–function modeling in glaucoma [published online June 21, 2022]. *Ann N Y Acad Sci*.
42. Shin J, Kim S, Kim J, Park K. Visual field inference from optical coherence tomography using deep learning algorithms: a comparison between devices. *Transl Vis Sci Technol*. 2021;10(7):4.
43. Lazaridis G, Montesano G, Afgeh SS, et al. Predicting visual fields from optical coherence tomography via an ensemble of deep representation learners. *Am J Ophthalmol*. 2022;238:52–65.
44. Phelps CD, Hayreh SS, Montague PR. Visual fields in low-tension glaucoma, primary open angle glaucoma, and anterior ischemic optic neuropathy. In: Greve EL, Heijl A, eds. *Fifth International Visual Field Symposium: Sacramento, October 20–23, 1982*. Sacramento: Springer; 1983:113–124.
45. 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.
46. Keltner JL, Johnson CA, Quigg JM, Cello KE, Kass MA, MO Gordon. Confirmation of visual field abnormalities in the Ocular Hypertension Treatment Study. Ocular Hypertension Treatment Study Group. *Arch Ophthalmol Chic Ill 1960*. 2000;118(9):1187–1194.
47. Devalla SK, Liang Z, Pham TH, et al. Glaucoma management in the era of artificial intelligence. *Br J Ophthalmol*. 2020;104(3):301–311.