Predicting visual fields from optical coherence tomography via an ensemble of deep representation learners

Georgios Lazaridis , Giovanni Montesano , Saman Sadeghi Afgeh , Jibran Mohamed-Noriega , Sebastien Ourselin , Marco Lorenzi , David F. Garway-Heath

PII: S0002-9394(21)00663-2
DOI: https://doi.org/10.1016/j.ajo.2021.12.020
Reference: AJOPHT 12113

To appear in: *American Journal of Ophthalmology*

Received date: August 3, 2021
Revised date: December 23, 2021
Accepted date: December 27, 2021

Please cite this article as: Georgios Lazaridis , Giovanni Montesano , Saman Sadeghi Afgeh , Jibran Mohamed-Noriega , Sebastien Ourselin , Marco Lorenzi , David F. Garway-Heath , Predicting visual fields from optical coherence tomography via an ensemble of deep representation learners, *American Journal of Ophthalmology* (2022), doi: https://doi.org/10.1016/j.ajo.2021.12.020

Predicting visual fields from optical coherence tomography via an ensemble of deep representation learners

Georgios Lazaridis,[1,2] Giovanni Montesano[1,3], Saman Sadeghi Afgeh,[4] Jibran Mohamed-Noriega,[1,5] Sebastien Ourselin,[6] Marco Lorenzi,[7] David F. Garway-Heath.[1]

[1]NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom

[2]Centre for Medical Image Computing, University College London, London, United Kingdom

[3]Optometry and Visual Sciences, City, University of London, London, UK

[4]Data Science Institute, London, City University, United Kingdom

[5]Departamento de Oftalmología, Hospital Universitario, UANL, México

[6]School of Biomedical Engineering and Imaging Sciences, King's College London, London, United Kingdom

[7]Université Côte d'Azur, Inria Sophia Antipolis, Epione Research Project, France

Corresponding author information: Georgios Lazaridis, University College London, UK

Email: rmaplaz@ucl.ac.uk

Abstract

Purpose To develop and validate a deep learning (DL) method of predicting visual function from spectral domain optical coherence tomography (SDOCT) derived retinal nerve fiber layer thickness (RNFLT) measurements and corresponding SDOCT images.
Design Development and evaluation of diagnostic technology.
Methods Two DL ensemble models to predict pointwise VF sensitivity from SDOCT images (model 1 – RNFLT profile only; model 2 – RNFLT profile plus SDOCT image), and two reference models were developed. All models were tested in an independent test-retest dataset comprising 2181 SDOCT/VF pairs; the median of ~10 VFs per eye was taken as the best available estimate (BAE) of the true VF. The performance of single VFs predicting the BAE VF was also evaluated.
Participants Training dataset: 954 eyes of 220 healthy and 332 glaucomatous participants. Test dataset: 144 eyes of 72 glaucomatous participants.
Main outcome measures Pointwise prediction mean error (ME), mean absolute error (MAE) and correlation of predictions with the BAE VF sensitivity.
Results The median mean deviation was -4.17 (-14.22 - 0.88) dB. Model 2 had excellent accuracy (ME 0.5, standard deviation [SD] 0.8, dB) and overall performance (MAE 2.3, SD 3.1, dB), and significantly (paired t-test) outperformed the other methods. For single VFs predicting the BAE VF, the pointwise MAE was 1.5 (SD 0.7) dB. The association between SDOCT and single VF predictions of the BAE pointwise VF sensitivities was $R^2 = 0.78$ and $R^2 = 0.88$, respectively.
Conclusions Our method outperformed standard statistical and DL approaches. Predictions of BAEs from OCT images approached the accuracy of single real VF estimates of the BAE.

Introduction

Glaucoma is the leading cause of irreversible blindness. Evaluating the progression rate of the pathology is crucial in order to assess the risk of functional impairment and to establish sound treatment strategies [1]. Clinically, optical coherence tomography (OCT) is used as a surrogate measure to evaluate retinal ganglion cell (RGC) loss by measuring retinal nerve fibre layer (RNFL) thickness around the optic nerve head (ONH), and other structural parameters, whereas standard automated perimetry (SAP) is employed to assess the status of the visual field (VF). Assessing the way in which structural and functional measures in glaucoma interact is clinically important. Visual loss is assumed to follow from, and correlate to, structural loss caused by the disease process. It would be clinically useful to know the magnitude and location of structural loss that will result in visually important functional loss. However, current clinical devices for measuring structural and functional deficits are far from accurate and have imperfect precision. Standard automated perimetry (SAP), the clinical cornerstone of functional testing in glaucoma, is subject to considerable measurement variability and is also a poor surrogate for RGC count and function, whereas optical imaging techniques provide only surrogate measures of the biological variable of real interest [1]. Despite their limitations, these techniques are currently central to the diagnosis and management of glaucoma. It would, therefore, be beneficial if structure and function measurements were directly linked in some way, allowing clinicians to corroborate damage estimates by considering the measurements in tandem and to combine measurements to gain precision in estimates of the rate of progression.

Several studies have been conducted in an attempt to quantify the structure–function relationship using clinical measurements [2-11]. Most typical approaches proceed by taking one summary value to represent function (for example, mean deviation [MD] of the visual field from SAP) and one number to represent the structural data (for example, average neuroretinal rim area or mean RNFL thickness (RNFLT)), then assessing the curvilinear (e.g., log-linear) or monotonic association between the two variables via R^2, Pearson, or Spearman coefficients. This approach has two major flaws: The use of summary data loses spatial information and may reduce power, while classical association measures and regression models assume a linear shape of the relationship. Furthermore, these analyses fail to take account of spatial associations in the data, an integral attribute of glaucomatous loss. These shortcomings provide a motivation to explore whether it may be possible to predict a visual field test by including structural data in its high-resolution form. For instance, in spectral-domain OCT (SDOCT), RNFLT estimates are yielded over an image space of several hundred pixels. The high dimensionality of this kind of data ideally should be taken into account when developing methods linking structural measures to the 50 or so individual locations in the VF. Moreover, individual locations from both structure (pixel or sector values) and function (areas of VF or individual locations) are more likely to interact as groups rather than single independent measurements. Spatial information contained in raw imaging data, such as SDOCT or scanning laser ophthalmoscopy (SLO), as well as in RNFLT measurements derived from OCT image segmentation, could be efficiently combined to guide the structure-function learning process by imposing helpful, otherwise unknown, anatomical priors. Linear methods to predict visual fields using OCT images have been proposed, but the accuracy of the results has been poor [12-14].

Meanwhile, deep learning algorithms based on Convolutional Neural Networks (CNNs) have been shown particularly efficient at extracting relevant image features from 2D and 3D images [15]. In ophthalmology, the application of deep learning led to advances in automated disease detection, such as the development of models to detect diabetic retinopathy and glaucoma using fundus images [16-19] or to transform image quality and appearance of OCT images [20-22]. Deep learning models have also been applied to SDOCT images with respect to diagnosis and segmentation tasks [23-27]. Recently, it was also shown that deep learning can provide previously unimaginable insights into images, such as predicting the sex of a person from a snapshot of their ocular fundus [28]. Even though this particular application is not clinically relevant, as sex can be readily known, it showcases that deep learning can identify links between quantities that may have been considered as disconnected. However, little has been done to apply deep learning models to predict function from structure in glaucoma. Zhu et al. [29] predicted measurements of the RNFLT derived from scanning laser polarimetry (SLP) and individual VF locations from SAP. However, they used a simple shallow mutli-layer preceptor (MLP) for the high-dimensional RNFLT estimates which might be insufficient to fully learn and characterise the required mapping function. In other work [30-33], deep learning models were applied to map structure to function. However, the modeling methods had important limitations and thus, the results provided marginal improvements over previous methods. For instance, in two studies [30, 33] the method was a simple CNN architecture, whereas in another [31] the authors used a combination of software-generated macular ganglion cell-inner plexiform layer (mGCIPL) and peripapillary retinal nerve fibre layer (pRNFL) thicknesses maps and an off-the-shelf deep learning network. In one study [32], the network was mostly focused on removing the noise from the VFs.

We propose an ensemble of two custom deep learning models to predict visual fields using RNFLT estimates from OCT alone and OCT images along with the corresponding RNFLT estimates. We train our ensemble model in one dataset and we test and evaluate its performance in an independent dataset. We built our ensemble model using a state-of-the-art custom architecture attempting to provide a clinically useful tool for mapping and charting concordance between VF measurements and RNFLT measurements in glaucoma.

Methods

Subjects

The study sample was derived from two independently acquired populations, the COMPASS and RAPID cohorts. These are the training/internal validation and test/external validation datasets, respectively.

COMPASS

444 healthy and 499 glaucoma subjects were recruited to an industry-sponsored technology assessment study at eight study sites, with 5 sites acquiring OCT images with the Spectralis. These were as follows: ASST - Santi Paolo e Carlo, Milan, Italy; Azienda Ospedaliero Universitaria Santa Maria della Misericordia di Udine, Udine, Italy; NIHR Clinical Research Facility at Moorfields Eye Hospital, London, United Kingdom; Department of Ophthalmology and Visual Sciences University of Iowa, 200 Hawkins Drive, Iowa City, Iowa; Department of Optometry & Vision Sciences, The University of Melbourne, Parkville, Australia. The study was designed to compare the clinical performance of the HFA and the Compass perimeter and it was funded by CenterVue, SpA (Padova, Italy). Only data obtained from the HFA test have been used in this research and will be described. The study was undertaken in accordance with good clinical practice guidelines and adhered to the Declaration of Helsinki. All patients gave their written informed consent to participate in the study. Ethics Committee approval was obtained (International Ethics Committee of Milan, Zone A, 22/07/2015, ref: Prot. n° 0019459) and the study was registered as a clinical trial (ISRCTN13800424). Participants were recruited consecutively and required to be aged between 18 and 90 years, to have best corrected visual acuity > 0.8 decimals (if ≤ 50 years old) or > 0.6 decimals (if >50 years old) in the study eye, refractive error between -10 Diopters (D) and +6 D, astigmatism ±2 D, absence of systemic pathologies that could affect the VF, no use of drugs interfering with the correct execution of the perimetric test and no past ocular trauma or surgery (apart from uncomplicated cataract surgery) in the tested eye. Healthy subjects were additionally required to have a normal optic nerve head in both eyes (no evidence of excavation, rim narrowing or notching, disc haemorrhages, RNFL thinning), Intraocular Pressure (IOP) less than 21 mmHg in both eyes and no other signs of ocular disease. Glaucoma subjects were additionally required to have glaucomatous optic neuropathy (GON) defined as glaucomatous changes to the ONH or retinal nerve fibre layer (RNFL) as determined by a specialist from fundus photograph or SD-OCT, independently of the VF, to be receiving anti-glaucoma therapy and to have no ocular pathologies, other than glaucoma, in the tested eyes. Eligible glaucoma patients were identified based on a clinical diagnosis of GON from the clinical registry of the glaucoma clinics in the recruiting centres. An expert clinician confirmed the diagnosis of GON using imaging data (RNFL SD-OCT or optic nerve photograph) acquired during the protocol examination (see below).

Each subject underwent an ophthalmological evaluation following a standard operating procedure. A perimetric practice test was offered to subjects naïve to perimetry. All subjects performed a perimetric test with the HFA 24-2 grid (SITA Standard) to both eyes (if both eligible).

Fundus pictures with the Compass perimeter and SD-OCT scans of the ONH and the of circumpapillary RNFL were acquired for the purpose of clinical confirmation of GON; the acquisition of OCT data was not subject to a standardised procedure. For the purpose of this study, we only included eyes with a circumpapillary RNFL scan performed with a Spectralis SD-OCT. The final selection included 954 eyes from 552 people (332 with GON).

Descriptive characteristics of the COMPASS cohort are summarised in Table 1. More details can be found elsewhere [34].

RAPID

Eighty-two clinically stable glaucoma patients under standard treatment (IOP mean 14.0 mmHg [5th to 95th percentile 8.0 to 21.0 mmHg] and VF MD -4.17 dB [5th to 95th percentile -14.22 to 0.88dB]) were recruited to a test–retest study [35]. Seventy-two participants (144 eyes) attended for up to 10 visits within a 3-month period, for a total of 1251 patient-eye visits; two VFs were obtained at one of the visits. These seventy-two participants were used in this study; this data set was taken to represent a ‘stable glaucoma’ cohort; assumptions made include that, over such a short length of time, no clinically meaningful changes in the VF or RNFL structure would occur and that the variability characteristics of the VF and RNFL measurements are similar to those seen in clinical practice over longer periods of time. The study was undertaken in accordance with good clinical practice guidelines and adhered to the Declaration of Helsinki. The study was approved by the North of Scotland National Research Ethics Service committee on 27 September 2013 (reference no.: 13/NS/0132) and NHS Permissions for Research was granted by the Joint Research Office at University College London Hospitals NHS Foundation Trust on 3 December 2013. All patients provided written informed consent before the screening investigations were carried out. Recruitment criteria were based on those for the UKGTS [36]. Patients were required to have reproducible VF loss with corresponding damage to the ONH and no other condition that could lead to VF loss, be aged > 18 years and have a visual acuity of 20/40, a refractive error within ± 8 dioptres and an IOP of < 30 mmHg. The VF MD had to be better than -16 dB in the worse eye and better than -12 dB in the better eye. VF loss was defined as a reduction in sensitivity at two or more contiguous locations with $p < 0.01$ loss or more, three or more contiguous locations with $p < 0.05$ loss or more, or a 10-dB difference across the nasal horizontal midline at two or more adjacent locations in the total deviation plot.

Participants attended approximately once a week for 10 visits, with VF testing and OCT imaging carried out twice at the first visit and once at each subsequent visit. VF testing was undertaken with the Humphrey Field Analyser^{TM} (HFA) and OCT imaging was carried out using Stratus TD OCT^{TM} (Carl Zeiss Meditec Inc., Dublin, CA, USA) and Spectralis SD OCT (Heidelberg Engineering, Heidelberg, Germany) (software version 5.2.4) (protocol “Peripapillary circular scans”: 16 averaged consecutive circular B-scans; diameter of 12 degrees, 1536 A-scans). If there was more than one image or VF at a visit, and all pass quality checks, we choose one at random. The principal baseline characteristics of the RAPID test-retest cohort can be seen Table 1. More details can be found elsewhere [36].

Table 1 Principal baseline characteristics for the COMPASS and RAPID cohorts. Age is a subject variable; IOP, refractive error, and SAP MD, and RNFL thickness are eye variables. Data are provided for eligible eyes, n = number; D = dioptres; dB = decibel; mmHg = millimetres of mercury; IOP = intraocular pressure; SAP = standard automated perimetry; MD = mean deviation

Data preparation

To optimize the input into the deep learning models, all OCT images were ‘flattened’ based on a pilot estimate of the retinal pigment epithelium (RPE) position, which is the most hyper-reflective layer in the scan, and aligned to each other. If a subject's left eye VF was tested, the recorded data were mapped to a right-eye format for analysis, and, similarly, all left-eye scans were mirrored to conform to the scans of the right eye. All scans were resized to 512 x 512 pixels. Training images were augmented with random probability using channel ratio modification and Gaussian and speckle noise corruption. All OCT images used are SD-OCT peripapillary circular B-scans as per Heidelberg Engineering protocol “Peripapillary circular scans”. Segmented RNFL thickness profiles from the same images were derived using the segmentation obtained with the Heidelberg Eye Explorer software.

Learning models

What follows is a description of the principal methods and models developed. We first developed two models (a multi-input convolutional neural net [MICNN] and multi-channel variational autoencoder [MCVAE]) with the same objective: mapping a structural measurement (RNFLT values from a software generated profile with or without additional raw imaging data, i.e. the OCT image, input) to a sensitivity value profile (in decibels) for all VF locations. Both models attempt to represent the relationship between VF and OCT measures without the limiting assumptions associated with the standard linear models, concerning the linearity of the relationship between VF and OCT and the independence across spatially distributed measurements. We then sought to combine these two models into an ensemble model. Such an ensemble model would allow the prediction of a VF from RNFLT values and the OCT image by maximising the information provided by the two sub-models. We generated two ensemble models, one with an RNFLT input (model 1) and one with an RNFLT and OCT image input (model 2).

Multi-input convolutional neural net

The multi-input convolution neural net consists of two separate sub-models trained on the data. It is composed of two separate input heads, taking as input the OCT image and the corresponding RNFLT measurements respectively, and of a shared regression module.
The first head takes the OCT image as input and is composed of 6 convolutional blocks. The first 4 blocks are composed of two convolutional layers, each followed by a Leaky ReLU activation, while the last two blocks are composed of 3 convolutional layers followed by a Leaky ReLU. A batch normalisation layer follows each activation layer, and a Max Pooling operation is applied after each activation. Kernel size is kept constant at 3 and stride at 1, while the number of filters starts at 32 and is doubled at each block. The final convolutional block is followed by two linear layers: a Leaky ReLU activation, Batch Normalisation and Dropout are applied after the first linear layer; only a ReLU activation is applied after the second linear layer.
The second input head takes the RNFLT segmentation as input and is composed of 5 linear layers, with Leaky ReLU activation, Dropout and Batch Normalisation layers after each linear layer except the last, that is followed only by a ReLU activation. Both input heads output a 1x52 vector (matching the 52 VF locations to be predicted) and the two vectors are combined through summation. The resulting 1x52 vector is then passed on to a linear layer, followed by a Leaky ReLU activation, Dropout layer and a Batch Normalisation layer, and to a final regression head, composed of a Linear layer with ReLU activation.
The models are initialised with Xavier initialisation[37] and trained for 150 epochs with the Adam optimiser and a learning rate of 0.001, and using a Mean Squared Error loss.

Multi-Channel Variational Autoencoder

Variational Autoencoders (VAEs)[38] are models that couple a recognition function, or encoder, to infer a lower dimensional representation of the data, with a generative function, or decoder, which transforms the latent representation back to the original observation space. The VAE is a Bayesian model: the latent variables are inferred by estimating the associated posterior distributions. Within this setting, we jointly analyse OCT and VF by using the multi-channel VAE (MC-VAE)(https://gitlab.inria.fr/epione_ML/mcvae) [39]. This approach extends the standard VAE by assuming the existence of a latent representation common to the different data channels, e.g., VF, OCT image, and RNFLT measures, which describes their common variability. Similarly, as with classical VAEs, the latent space is estimated from the data itself through an encoding operation and is optimized to predict the different channels through a decoding operation. Being a generative model, MC-VAE also allows cross-channels imputation and prediction.

In what follows, we implemented the MC-VAE so as the encoding to the latent space and the decoding from the latent space are convolutional neural networks, with architecture similar to that of the multi-input convolutional neural net presented above. Solving the optimization problem allows the discovery of the common latent space from which the observed data in each channel are generated, along with decoding and encoding transformations allowing cross-channels prediction. We choose a 3-dimensional latent space shared by each channel; we selected the subspace generated by the most relevant latent dimensions identified by variational dropout ($p < 0.2$). More information can be found in Antelmi et al. [39].

Ensemble Technique

We adopt stacked generalization or “stacking”[40] in order to combine the predictions of our two models. Stacked generalization is an ensemble method where a new model learns how to best combine the predictions from multiple existing models. In the absence of specific domain knowledge, it is better to ensemble different models rather than intensify computational efforts into selecting and optimising a specific model type.

The motivation to ensemble our two models is that each model performs well on a different range of VF locations. Also, model stacking is less sensitive to changes in a data set and generalizes better than a single model. That is, it makes better predictions on unseen data than just a single model. Furthermore, model stacking deduces the bias in a model on a particular data set so that we later can correct for the bias in a meta-learner.

We combine our models using tree boosting, namely XGBoost [41]. XGBoost takes the outputs of our two models as input and attempts to learn how to best combine the input predictions to make a better output prediction. This final model is trained on the predictions made by the two base models. That is, data not used to train the base models are fed to the multi-input CNN and the MC-VAE, predictions are made, and these predictions, along with the expected outputs, provide the input and output pairs of the training dataset used to fit the meta-model. The outputs from the base models used as input to the meta-model are real values since we perform regression. The training dataset for the meta-model is trained via 5-fold cross-validation of the base models, where the out-of-fold predictions are used as the basis for the training dataset for the meta-model. Also note that this cross-validation was only used for the purpose of training, whereas the actual testing was performed on an independent dataset (RAPID). Note that unlike a weighted average ensemble, a stacked generalization ensemble can use the set of predictions as a context and conditionally decide to weigh the input predictions differently, resulting in better performance.

Linear and Bayesian Radial Basis Function models

Linear Model In the classic linear model, individual VF sensitivity values are predicted from a set of independent variables x_i, i.e. RNFLT values, and their corresponding weights w_i. The weights quantify the contribution made by x values to predict the y values. The largest absolute weight value indicates the x value contributing most to the prediction. Similarly, the next largest absolute weight term would indicate the second most important term and so on. To find the optimal weights w, the difference between the predicted and measured values must be minimal. Thus, this difference is optimised to predict a complete VF from a given vector of x values.

Radial Basis Functions The RBF models the relationship between y and x without the following limiting assumptions associated with the classic linear model: (a) each x value is independent of all the other x values (b) assumes that the relationship between y and x is either linear or becomes linear after some transform (typically logarithmic) (c) outlier points exert an overly strong influence and can yield a false association. The central idea of the RBF is the basis functions, each of which performs very much like a dynamic window or kernel that moves across the data, both spatially and at various stages in disease severity, identifying groups of measurements that appear to behave in a similar pattern. Moreover, the RBF learns the parameters from the data and makes predictions in multiple dimensions. The non-normalized Gaussian basis function used in Zhu et al. [29] has an activation field that has a center—that is, a particular input value at which it has a maximal output. The output tails off as the input moves away from this point. In this way, those hidden basis functions that have centers similar to the input x patterns will have stronger activation and will thus contribute more to the prediction of y. On the other hand, those basis functions with weak activation will be isolated and will not affect the prediction. More information can be found in Zhu et al. [29].

Testing (external validation)

It is essential to evaluate a modelled relationship between variables on an independent external validation dataset that doesn't involve the data used to learn that relationship, i.e. train the learning algorithm. If the validation is performed on the training dataset, then the model estimates will be overly optimistic because the model has already seen the data and knows exactly how to handle them, identify patterns and determine how to best predict the target variables. Therefore, the models were developed on the COMPASS study data alone, leaving the RAPID data as an external validation/test dataset.

The predictive performance of the tested models was evaluated with a location-by-location analysis of the predictions of VF sensitivity from each OCT scan in the RAPID dataset.
A single VF is an imprecise estimate of the true retinal sensitivity; VF testing is known to exhibit considerable test retest variability[50]. As the RAPID dataset comprises up to 10 VFs per eye obtained over a short period, we used the series to calculate a best available estimate (BAE) for the sensitivity at each VF location (selecting one VF per visit if more than one VF had been obtained on the same day). The BAE is the median of the test-retest sensitivity values for each location, assuming that the error distribution would be symmetric around the true sensitivity. This allowed us to have a BAE that was not affected by the lower bounds on the measurement (i.e. values censored at 0 dB) as opposed to the raw mean of the sensitivity values. In clinical perimetry, the BAE is as close as it is possible to get to the 'true' retinal sensitivity.
We assess prediction performance for each model for both 1-1 pairs (single OCT prediction to single VF) and for the BAE (single OCT prediction to the BAE of the VF). For the 1-1 pair comparisons, for each visit, we take the OCT and the corresponding VF. As a result, we have ~10 OCT-VF comparisons per eye. With these, we calculate the prediction error for the pairs. For the BAE, the difference between each OCT prediction and the BAE is calculated. The errors of the models to predict the BAE were taken as the principal evaluation of model performance. The prediction errors from the 1-1 pair analysis represents the error of the model plus the variability inherent in both OCT imaging and VF testing.
The overall prediction performance was summarized by the mean error (ME) and mean absolute error (MAE) for predictions at the 52 locations of the VF. Graphic representation of the predictions is stratified by VF location sensitivity level in the BAE VF. We compare our ensemble models (RNFLT-only and RNFLT + OCT image) with the classic linear model and with the Bayesian Radial Basis Function (BRBF) network [29].

All experiments were performed on a NVIDIA Titan V (12GB) GPU using PyTorch and 5-fold cross-validation is used for training: 80% of the training data is used for training and the remaining 20% for validation in each fold. We present a structure–function map in a format similar to that described by Gardiner et al. [6] as well as location-by-location predictions of each subject's VF, as represented by the HFA grayscale (which was replicated for this purpose). These outputs were considered for (1) the classic linear model, (2) the BRBF model, and (3) for our ensemble model 2.

Results

Figure 1 illustrates the distributions of the error between the pairwise predicted and the measured sensitivity for each VF location in the RAPID data, stratified by VF sensitivity, for each of the two individual models that make up the ensemble model 2, MCVAE (RNFLT + OCT image) and Multi-Input CNN (RNFLT + OCT image), respectively. The different error distribution justifies the rationale for stacking the two models into an ensemble to obtain the final predictions.

Figure 2 summarizes the pairwise predictive performance of the linear model, the BRBF model [29] and our two ensemble models across the range of VF sensitivity measurements. Each error bar summarizes the predictive performance over a 1-dB range from 0 to >36 dB. Predictions at higher sensitivities (>30 dB) tend to be slightly lower than the actual values, whereas at lower sensitivities (<20 dB), the predictions tend to be higher.

Figure 1: Distributions of the error between the predicted and the measured sensitivity (single OCT/VF pairs) for each VF location in RAPID data, stratified by VF sensitivity, for the two sub-models of the final ensemble model 2 (RNFLT + OCT image inputs). Left: Multi-channel variational autoencoder. Right: Multi-input convolutional neural net.
OCT: optical coherence tomography. VF: visual field. RNFLT: retinal nerve fiber layer thickness.

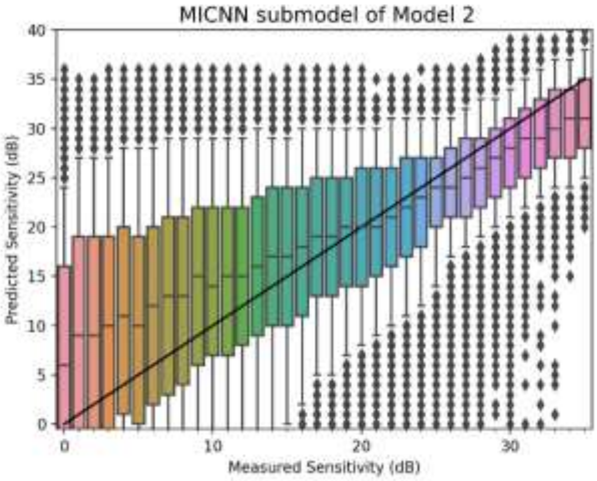
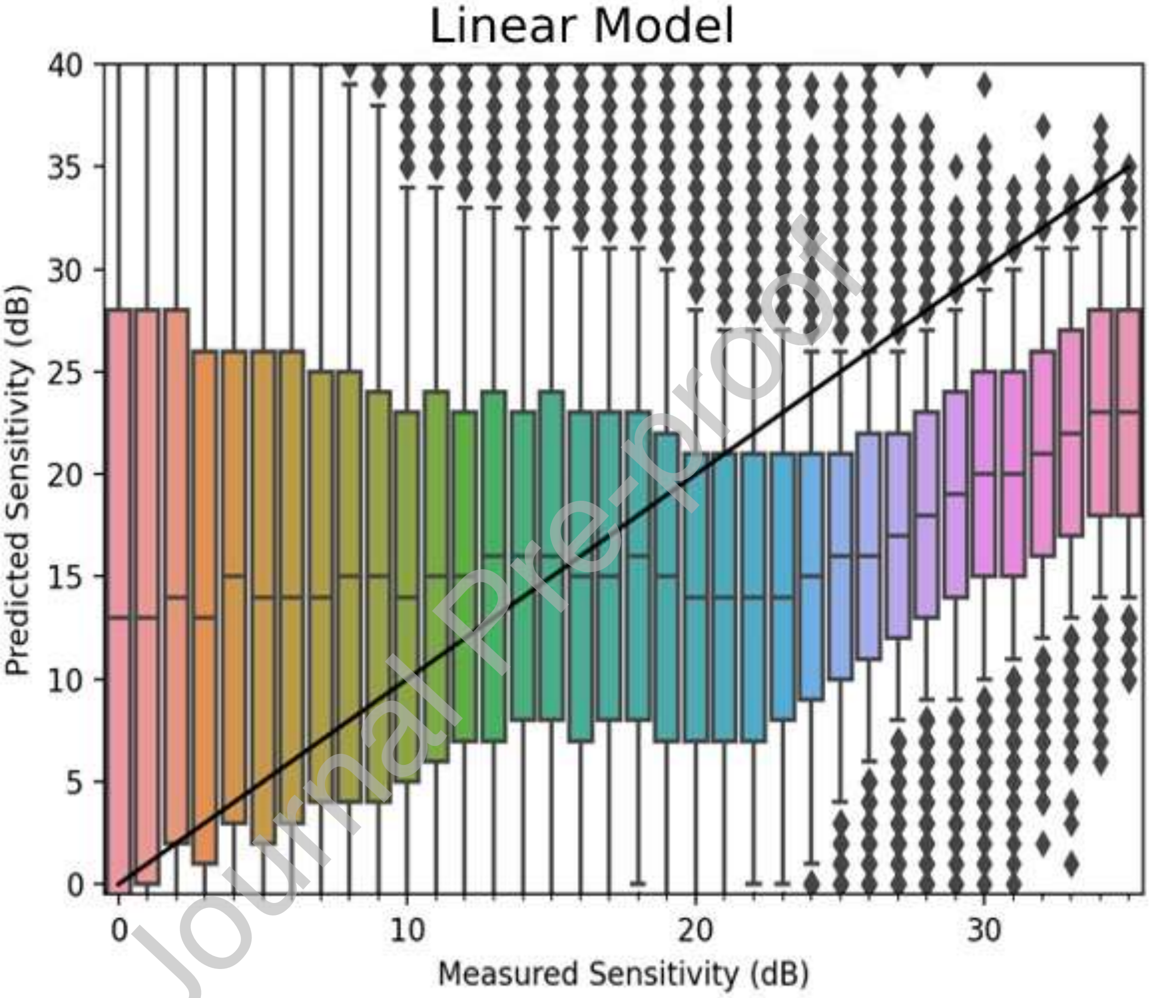
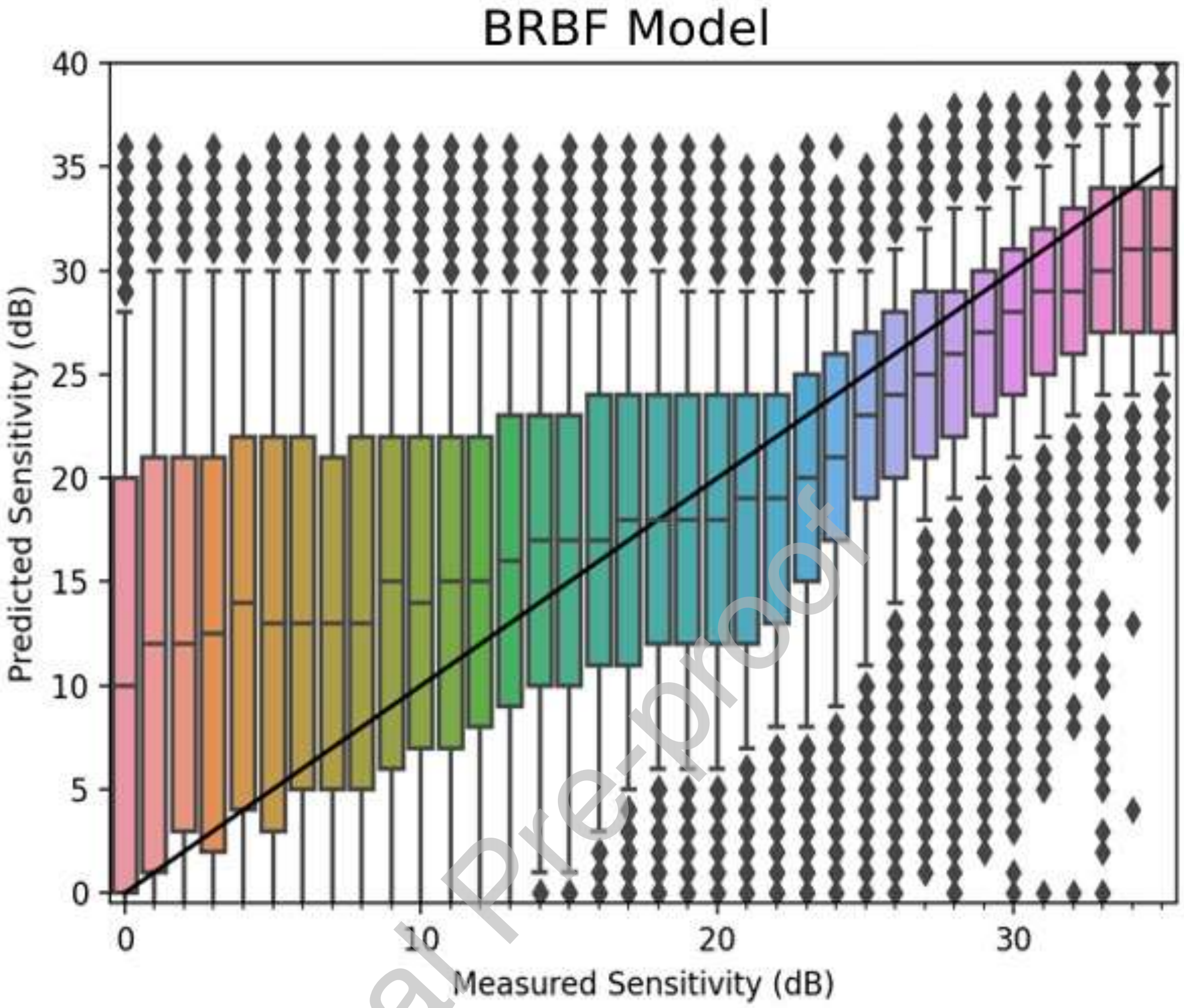
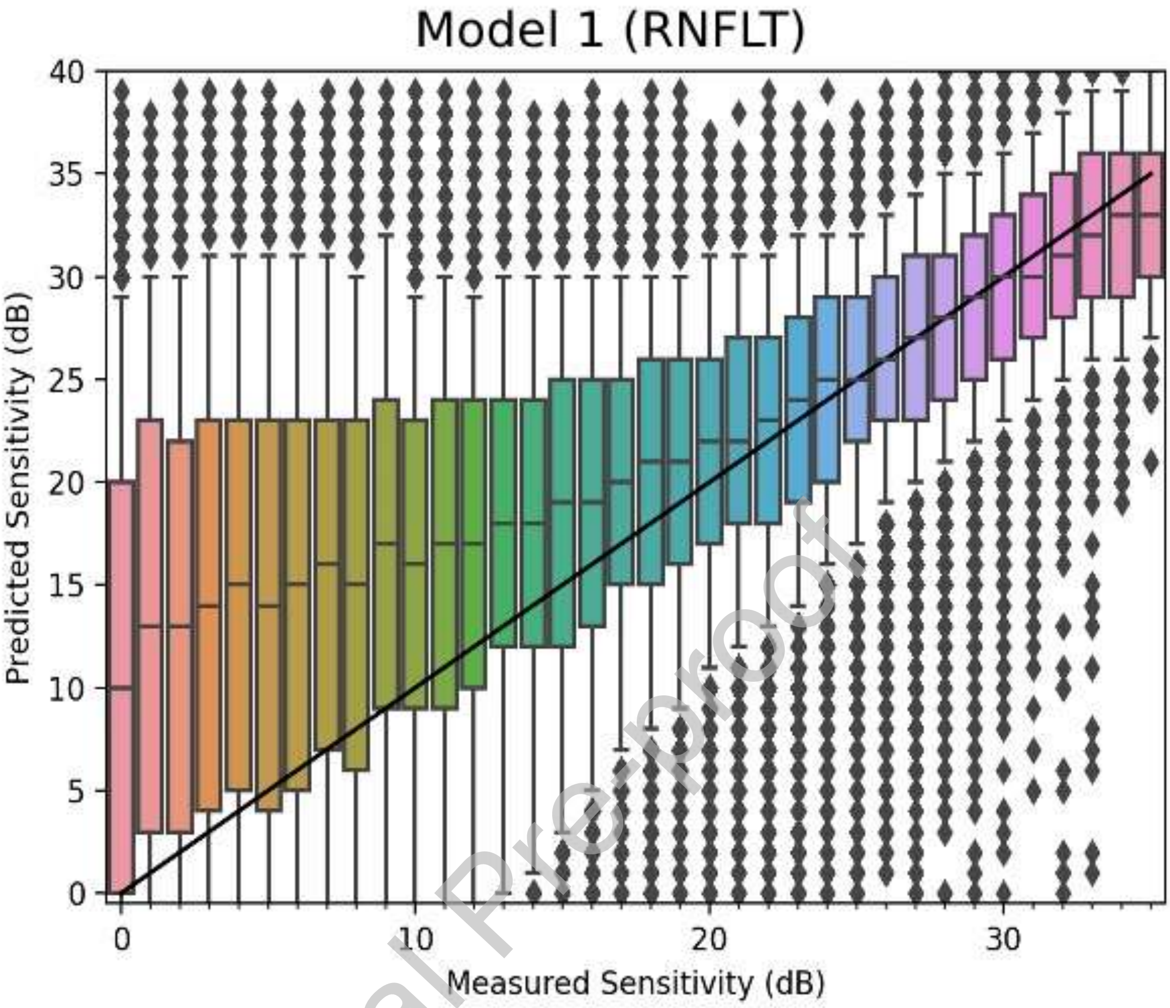

MICNN submodel of Model 2
Predicted Sensitivity (dB)
Measured Sensitivity (dB)
0
5
10
15
20
25
30
35
40
0
10
20
30

Figure 2 Distributions of the error between the pairwise (single OCT/single VF) predicted and the measured sensitivity for each VF location in RAPID data, stratified by VF sensitivity. Each error bar summarizes the predictive performance over a 1-dB range from 0 to >36 dB. Each box indicates the interquartile range of the prediction error (25th and 75th percentile error) with the line in the box indicating the median error. The dotted line of unity indicates perfect prediction (no error). The predictive performances of (a) the classic linear model, (b) the BRBF and (c) the ensemble method using only RNFLT (model 1) (d) the ensemble model using both RNFLT and OCT images (model 2) are shown.
VF: visual field. dB: decibel. BRBF: Bayesian radial basis function. RNFLT: retinal nerve fiber layer thickness. OCT: optical coherence tomography.




BRBF Model
Predicted Sensitivity (dB)
Measured Sensitivity (dB)
0
5
10
15
20
25
30
35
40
0
10
20
30


Model 1 (RNFLT)
Predicted Sensitivity (dB)
Measured Sensitivity (dB)
0
5
10
15
20
25
30
35
40
0
10
20
30

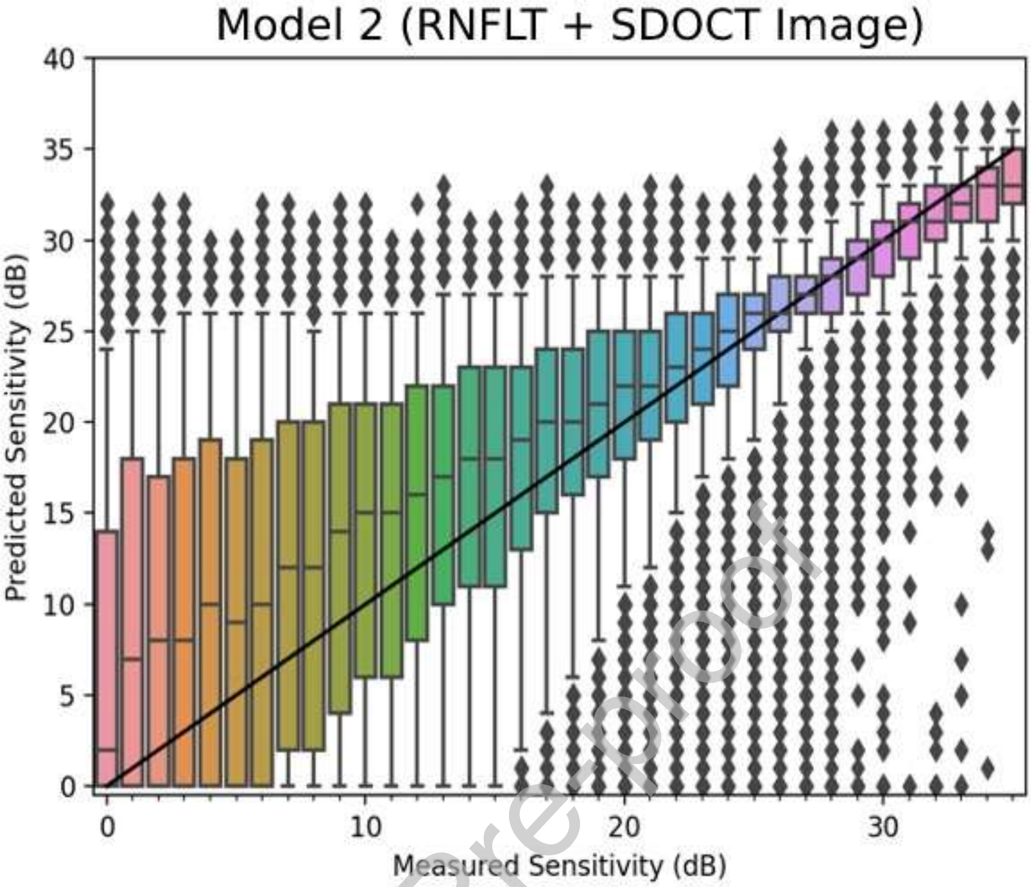
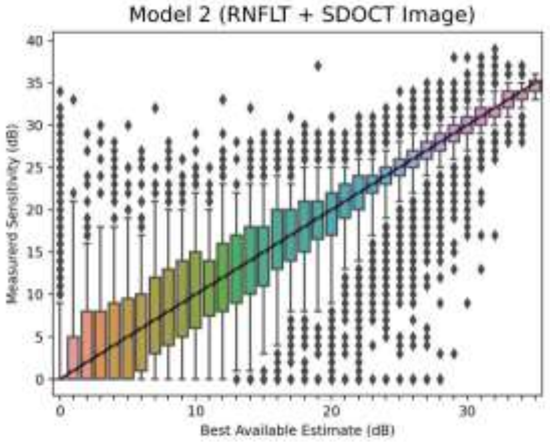
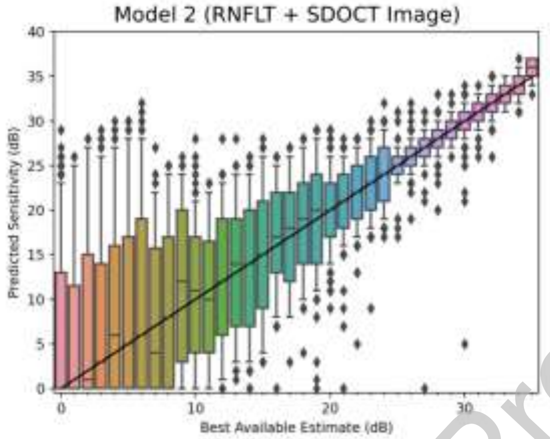


Table 2 summarises the MAE and SD of AE for pairwise and BAE comparisons for all methods evaluated. For pairwise predictions, the ensemble model 2 achieved significantly (P<0.05) better predictions than model 1 (mean absolute prediction errors 2.8 dB and 3.6 dB, respectively). Compared with the linear regression and BRBF models, our ensemble model 2 yielded a statistically significant improvement (P<0.001 paired t-test) in performance of predicting VF sensitivity in the test/external validation dataset.

Table 2: Quantification of pairwise and Best Available Estimate (BAE) pointwise prediction errors for each method.
RNFLT: retinal nerve fiber layer thickness. OCT: optical coherence tomography. MAE: Mean Absolute Error. SD: Standard Deviation of AE. dB: Decibels. BRBF: Bayesian Radial Basis Function

Figure 3: Prediction error: individual VFs and OCT scans (model 2) predicting the BAE VF. 3a: Prediction errors for individual VFs predicting the BAE (represents the VF prediction accuracy and measurement variability). 3b: Prediction errors for individual OCTs (model 2) predicting the BAE (represents the OCT prediction accuracy and measurement variability).
VF: visual field. BAE: best available estimate. OCT: optical coherence tomography.

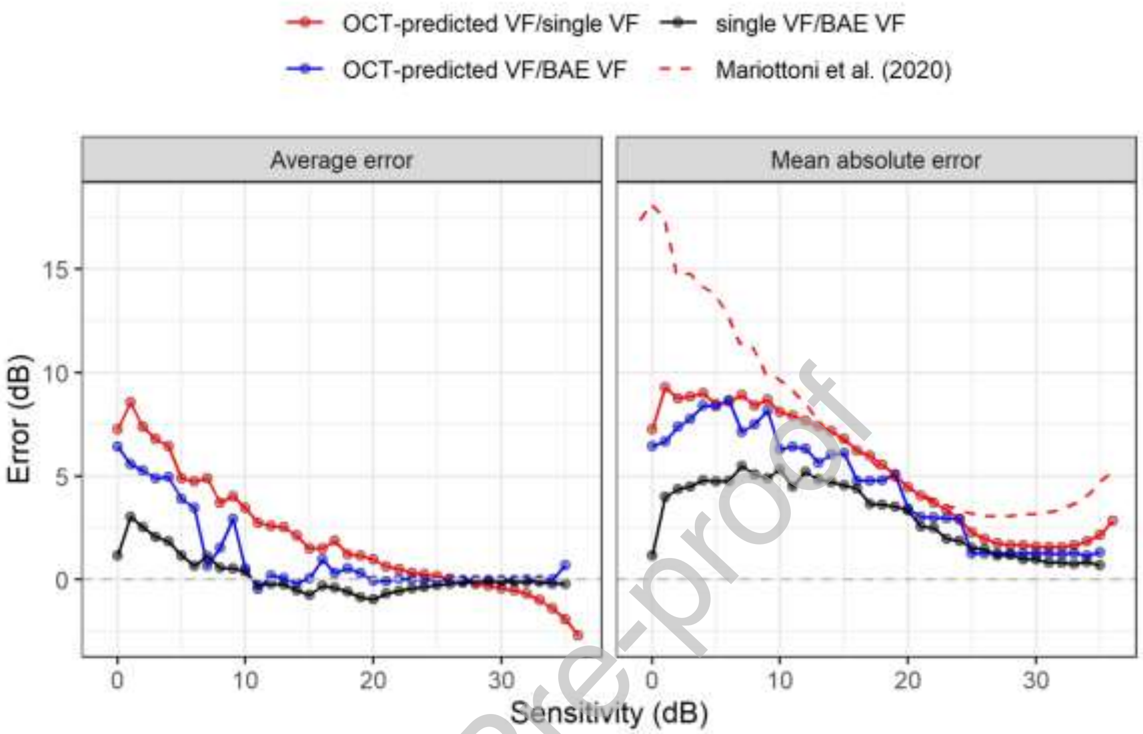




To set the OCT prediction errors in the context of the measurement variability inherent in VF testing, we plot model 2 single (real) VF pointwise sensitivity against the BAE VF (Figure 3a) and pointwise OCT sensitivity predictions against the BAE VF (Figure 3b). On average, the OCT predictions are highly accurate (the median prediction is very close to the BAE sensitivity). The average mean error (ME) per eye between the OCT-predicted VF and the BAE VF was 0.5 (SD 0.8) dB and the ME stratified by pointwise BAE VF sensitivity is shown in Figure 4a. The average MAE between the OCT-predicted VF and the BAE VF per eye was 2.3 dB (SD 1.2). The MAE stratified by pointwise BAE VF sensitivity is shown in Figure 4b. The MAE for single VFs predicting the BAE VF per eye is 1.5 dB (SD, 0.7 dB). The association between OCT-predicted and BAE pointwise VF sensitivity was R2 = 0.78, compared to R2 = 0.88 for single VFs and the BAE. Thus, the precision of the VF predictions from OCT scans compares favourably with the prediction from single real VF measurements.

For predictions of the VF summary measure ‘mean sensitivity’, the MAE for the prediction of the BAE of mean sensitivity was 0.64 dB for ensemble model 2 (ME 0.45 dB), compared to 0.67 dB for single VF predictions of the BAE (ME -0.10 dB).

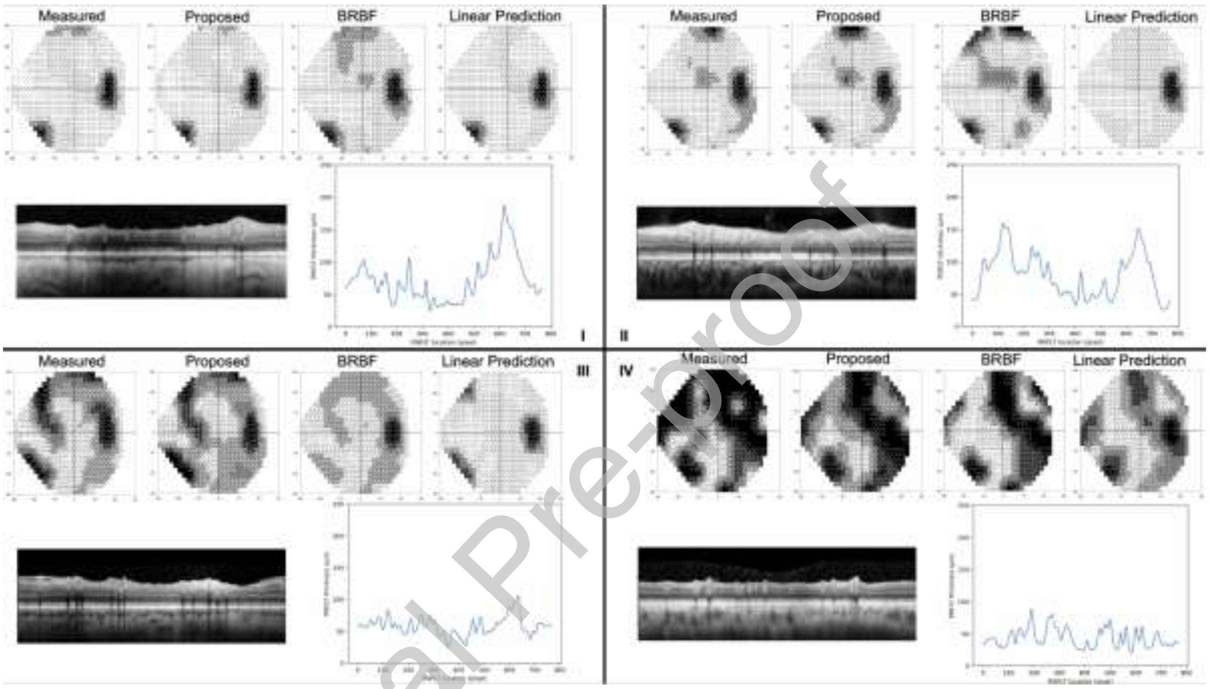
Figure 4: Mean error (ME) and mean absolute (MAE) error for single OCT/VF pair (model 2) and for single OCT/BAE VF (model 2); the MAE for single OCT/VF pair (model 2), for single OCT/BAE VF (model 2) and the MAE for single VF to BAE VF. Single OCT/VF pair MAE from Mariottoni et al.[33] is reported for comparison. VF: visual field. BAE: best available estimate. OCT: optical coherence tomography.



The negative bias at higher sensitivities and positive bias at lower sensitivities seen for single OCT/single VF pair prediction (Figure 2d and Figure 4a) was almost eliminated for the single OCT/BAE VF prediction (Figure 3b). As the sensitivity values are ranked, the smaller bias for the BAE likely represents a reduction in regression-to-the-mean obtained by averaging ~10 VFs for each eye. The residual positive bias in the ME below about 10dB (Figure 4a) results from the censoring of VF sensitivity at 0dB (the median error is very close to the line of equivalence [Figure 3] for both real VF and OCT predictions of the BAE.

Figure 5 gives some case examples of the predictions. In all cases, the linear model underestimates the defect severity of the VF, when compared with the true (paired) single measured VF. In Figure 5I, Figure 5III, Figure 5IV, the linear model matches the overall average sensitivity of the VF but fails to capture the spatial location of this loss (Figures 5I – 5IV). The BRBF model provides better estimates compared to the linear model, better predicting the damaged VF and partially capturing the spatial location of the loss. In each case, the proposed ensemble method (model 2) better estimates the true VF, with spatial features of the measured defects generally retained. The proposed ensemble method (model 2) not only predicts the damaged VF and captures the spatial location of the loss but it also manages to predict the advanced defect severity (Figure 5III, Figure 5IV).

Figure 5: Model predictions for four cases from the RAPID dataset. For each case (I–IV), the top row shows, from left to right, VF grayscales for the measured VF, the VFs predicted from the proposed ensemble method (model 2), the BRBF and the classic linear regression, respectively. The row of graphics (below) shows the corresponding OCT image and 768-pixel segmentation RNFLT profile (blue line) used to predict the VFs.
VF: visual field. BRBF: Bayesian radial basis function. OCT: optical coherence tomography. RNFLT: retinal nerve fiber layer thickness



Discussion

The main objective of this study was to develop a state-of-the-art deep learning architecture to predict 24-4 VF threshold values at each location of the VF from OCT imaging. Although the application of artificial neural networks (ANNs) to both functional and structural measurements in glaucoma is not novel [42-48], most of these studies have used a conventional shallow multi-layer perceptron (MLP) which presents important limitations. The main disadvantage of MLPs is that the number of total parameters can grow to be very great because it is fully connected; each perceptron is connected with every other perceptron. This is inefficient because there is redundancy in such high dimensions, resulting in slow convergence during training. Another disadvantage is that MLPs disregard spatial information. While there are many reasons for that disadvantage, one of them is because their dense connections do not allow them to scale easily and do not provide a translationequivariant data representation. This means that if there is a signal in one part of the image to which they needed to be sensitive, they would need to re-learn how to be sensitive to it if that signal moved around. This reduces the capacity of the network, and so training becomes hard. CNNs solved the signal-translation problem, because they convolve each input signal with a detector, i.e. kernel, and thus are sensitive to the same feature regardless of its location in the image. Hence, MLP ANNs are less suitable for the mapping of points in different measurement spaces, which requires a detailed understanding of the hidden layer output and other manipulation within the network.

The proposed model allows the unsupervised stratification of the latent space by disease status, providing evidence for a clinically meaningful interpretation of the latent space. This relationship indicates that both RNFLT values and OCT images are correlated with the VF measurements (see appendix for discussion).

The range and distribution of differences between the measured VF sensitivity values and those predicted by the various models, stratified by sensitivity level, for individual OCT-VF pairs is shown in Figure 2. The best predictions are clearly obtained by our ensemble model 2 (Figure 2d). Although the MAE in predictions from single OCT to single VF in model 2 is reduced compared to the linear and the BRBF models, the standard deviations of the absolute prediction errors of our model are still relatively high (3.7 dB), although lower than those reported in previous studies. There is a general similarity between the prediction limits (Fig. 2d) and VF test–retest limits (Fig. 7a of Artes et al. [50]), with predictions at the normal end of the range tending to be more precise and with a small negative bias (slightly lower than the actual VF measurements) and less precise at the damaged/low sensitivity end, with a positive bias (predictions tending to be a bit higher than actual VF measurements). This bias is likely a regression-to-the-mean effect, because the sensitivity values have been ranked. The floor effect, which is associated with glaucoma severity,[49] may be an additional cause of the small overestimation at the lower end of the VF sensitivity. However, when the predictions of model 2 were compared with the BAE VF, the regression-to-the-mean effect is largely removed and the median prediction was very close to the ‘true’ value across the range of sensitivity measurements.

To give context to the VF predictions from OCT, we plot the errors for single VFs predicting the BAE for the same eye (Figure 3a). This essentially reflects the test-retest noise. The OCT-based VF predictions from our model resembles this noise profile. This similarity suggests that, on average, a VF predicted by our model has measurement noise only slightly greater than that found in a newly measured field. This finding is not as exciting as it may first appear, because it is well established that the measurement noise in VFs is already very high, hindering clinical diagnosis of glaucomatous defects and monitoring progression.

Nevertheless, this finding illustrates that the range and scale of the average predictive performance of our model is much better than most modern approaches and the classic linear model, which completely fails to predict the full range of VF values (Fig. 2a). For model 2, the MAE, across all sensitivity levels, for a single OCT predicting the BAE VF was only 2.3 (SD 3.1) dB. This compares favourably with a single VF predicting the BAE VF: MAE 1.5 (SD 0.7) dB. The predictability of the BAE VF, both by single VFs and OCT-predicted VFs, varies with the VF sensitivity itself. Whereas, on average, the predictions are accurate across the range (Figures 3b and 4a), the size of prediction errors increases as VF sensitivity decreases (Figure 4b); the effect is slightly greater with OCT-predicted VFs than with single real VFs predicting the BAE. The R^2 value for the association of OCT-predicted VF sensitivity values with the BAE VF values was 0.78; the association of single VF values with BAE VF values, the R^2 0.88. The prediction errors for the VF mean sensitivity are even lower: the MAE was 0.64 (ME 0.45 dB) for model 2 predictions of the BAE, compared to 0.67 (ME -0.10 dB) for single VF predictions of the BAE. Thus, it appears that an OCT-predicted VF is almost as accurate a representation of the 'true' VF (BAE) as a real single VF test result. This has clear implications for clinical practice and clinical trials, were taking an OCT in addition to a VF in one visit may improve the precision of estimates of rates of VF progression. It also implies that assessment of concordance between VF and OCT results will be less error prone.

The improvement of model 2 (Fig. 2d) over model 1 (Fig. 2c), obtained with the addition of the OCT image to the RNFLT profile, indicates that additional information can be extracted from OCT images besides the RNFLT. This might include RNFL reflectivity[53, 54], choroidal features[55] and the location of the major vessels, which is associated with the RNFLT profile and bundle geometry [56, 57]. This might have important clinical implications. The MAE was reduced by approximately 22% and the improvement was observed both at the higher and lower sensitivities. This would obviously lead to better detection of which portions of the VF are expected to be healthy or damaged. Moreover, the subtle features present in the OCT B-scan but not captured by the simple RNFLT might help customise predictions for individual eyes. For example, the location of the blood vessels within the scan are known to affect the RNFL bundle trajectories and the corresponding structure-function mapping[57]. Exploring the different aspects contributing to better predictions will be the focus of future work.

Our method outperforms other methods described in the literature. In a recent study, Christopher et al. [49] used a deep learning method to predict glaucomatous visual fields from Spectralis SD-OCT ONH images. The authors used various inputs (RNFLT maps, RNFL en-face images, and SLO images) and the predictions for each input were compared. The main limitation of this study is that it used only one type of input each time and predicted only visual field global indices, including the mean deviation (MD), pattern standard deviation (PSD), and mean sectoral pattern deviation. The best MAEs, between the predicted and real (single VF) values, were 2.5 dB for MD and 1.5 dB for PSD; this compares to the MAE of 0.64 (ME 0.45 dB) for our ensemble model 2 predictions of the BAE mean sensitivity. The improvement in the results from our model, compared to those of Christopher et al., probably underestimates the difference because the magnitude of prediction error is related to the VF sensitivity (Figures 3 and 4). Christopher et al. did not stratify prediction errors by VF sensitivity level. Although not stated directly, their test data set probably had an average MD of around -2.3 dB, whereas our external validation data set had an average MD of -4.2 dB. Park et al. [31] introduced a deep learning method and the inputs were macular ganglion cell-inner plexiform layer (mGCIPL) and peripapillary retinal nerve fibre layer (pRNFL) thickness maps acquired from Cirrus SD-OCT images. The authors achieved root mean square error (RMSE) of 4.79 dB for pointwise predictions. This compares with an MAE 2.8 dB for single OCT/VF pairs and 2.3 dB for single OCT/BAE predictions with our model. Park et al. identified that glaucoma severity was related to the prediction errors, but did not stratify their prediction errors by severity. The average MD in their external validation data set was about -4.5 dB, so it is reasonable to compare their average prediction error (RMSE 4.8 dB) to ours (MAE 2.8 dB). Zhu et al.[29], using the BRBF framework for 'single scan to single VF' predictions, achieved a MAE of 2.9 dB, which was better than both the classical linear regression model (4.9 dB) and that reported by Park et al.[31] (4.79 dB). The main limitation of this study is that the test dataset (Blue Mountains Eye Study data) largely consisted of healthy subjects (230 healthy and 76 glaucomatous subjects). As expected, the prediction error was worse in glaucoma patients than healthy subjects; the large proportion of healthy subjects in their study likely reduced the prediction error. The distribution and magnitude of errors of the BRBF model in OCT is similar to that reported for SLP (Fig. 2b, Fig. A2), underlining the superior performance of our ensemble models. Moreover, the BRBF model assumes that the variability in the VF measurements is largely Gaussian, which is not optimal, given that it is often skewed and heavily tailed. Mariottoni et al. [33] developed a deep learning-based structure-function map using RNFL thickness profiles from SDOCT images and VF measurements. The authors achieved an average pointwise MAE of 4.25 dB in their test dataset which had an average MD of -4.5 dB. Appropriately, they plotted the MAE by VF sensitivity level. We included their results, stratified by sensitivity level, for comparison in Figure 4; the prediction performance is similar to our model 2 between about 13 and 23 dB sensitivity, but notably less good above and below that range. Hashimoto et al. predicted the VF in the central 10° from SDOCT images using the thickness of the retinal nerve fibre layer, the ganglion cell layer + inner plexiform layer and the outer segment + retinal pigment epithelium [30]. They used the thickness of the three macular layers as input to a CNN achieving a MAE for individual locations of 5.47 dB for an average MD of -10.4 dB. They did not stratify their prediction errors by VF sensitivity, making a comparison with our results difficult. However, the MAE for our model 2 was worse than 5.5 dB only at locations with sensitivity below about 18 dB (for the single OCT/single VF pair predictions).

All the referenced studies report prediction errors for single OCT/VF pairs, which includes errors arising not only from the predictions, but from the variability inherent in VF testing. This makes it difficult to interpret the true prediction accuracy. We include predictions of the BAE VF, which should largely remove the VF variability element. Figure 4 can be used to infer the underlying prediction accuracy by comparing the single OCT/VF pair prediction error curve with that of the OCT/BAE VF curve in the data from our study.

Our methodology overcomes many of the limitations discussed. First, the entire visual field is predicted, using both the RNFLT segmentation and peripapillary retinal SD-OCT images simultaneously. Second, the deep learning architecture is purposely designed for the task as opposed to the off-the-self tools used in previous studies. Third, we employ a sound probabilistic ensemble prediction based on our sub-models to obtain a final prediction estimate derived through cross-validation on the training data. Fourth, the model does not rely on specific assumptions, i.e. linearity, with respect to the variability in the VF measurements. Finally, our training dataset consists of a sound ratio of healthy and glaucoma subjects, whereas the test/external validation dataset is a test-retest study with clinically stable glaucoma patients, for which the VF prediction is more valuable because we were able to generate a BAE VF.

One goal is for our model to provide a relevant clinical tool that indicates concordance between the VF and the chosen surrogate measure for structural loss. For instance, when a VF and a structural measure are available, a chart mapped in VF space could be provided indicating areas where the measurements are in concordance (within a certain range of accuracy and precision) and where they are not[52]. This chart could provide clinically useful information about the diagnosis or the reliability of the individual measurements. Another goal is to facilitate structure/function integration, by translating the structural measures into VF space, to improve the precision of estimates of rates of glaucoma progression. This now seems feasible, especially as the median prediction errors are close to zero across the range of VF sensitivity levels, and needs to be tested in longitudinal data.

It is an imperative that any new statistical method should be developed and tested on more than one dataset[58]. In our study, we had access to two large, independent datasets and the inclusion criteria for glaucoma were generally consistent. However, as the purpose of this study was not to determine diagnostic performance, the precise range of glaucoma damage was less important. In fact, the range of glaucoma severity in the data can be viewed as an advantage in the study design. Moreover, testing on different datasets, where realistic estimates of measurement precision have been performed (from test–retest measurement, i.e. RAPID study), is the biggest advantage of our study design.
The method is not limited to one type of input of structural measurement or imaging device. It was shown to handle input of the RNFLT profile (768 values) as well as the SD-OCT image. The same approach could be used on neuroretinal rim area values from scanning laser ophthalmoscopy technology or any other surrogate measure of glaucomatous structural loss.

In conclusion, we have introduced a methodology for translating functional and structural measurements used in the clinical evaluation of glaucoma into the same domain – predicting the VF from OCT images. Evidence from a dataset independent of that used to derive the model indicates that our method has advantages over standard statistical and deep learning approaches for modeling these relationships. Estimates of functional deficits from structural measures yielded from this method are better than those derived from previous approaches and approach the accuracy of single VF tests.

Acknowledgements

a. Funding/Support: The research was supported by the EPSRC (CDT in Medical Imaging, EP/L016478/1), the International Glaucoma Association, Santen Pharmaceutical Co., Ltd., the National Institute for Health Research (NIHR) Biomedical Research Centre based at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology. Marco Lorenzi is also supported by the French government, through the 3IA Cote d'Azur Investments in the Future project (ANR-19-P3IA-0002) managed by the National Research Agency.

b. Financial Disclosures: Georgios Lazaridis: No financial disclosures. Giovanni Montesano: Consultant for CenterVue. Saman Sadeghi Afgeh: No financial disclosures. Jibran Mohamed-Noriega: No financial disclosures. Sebastien Ourselin: No financial disclosures. Marco Lorenzi: No financial disclosures. David F. Garway-

Heath: Consultant for Aerie, Alcon, Allergan, Bausch & Lomb, Pfizer, Quark, Quethera, Santen, Santhera.

References

1. Anderson R. The psychophysics of glaucoma: Improving the structure/function relationship. Prog Retin Eye Res. 2006;25(1):79-97. doi:10.1016/j.preteyeres.2005.06.001
2. Wollstein G, Schuman JS, Price LL, Aydin A, Beaton SA, Stark PC, et al. Optical coherence tomography (OCT) macular and peripapillary retinal nerve fiber layer measurements and automated visual fields. Am J Ophthalmol. 2004;138(2): 218-225. doi:10.1016/j.ajo.2004.03.019
3. Sato S, Hirooka K, Baba T, Tenkumo K, Nitta E, Shiraga F. Correlation Between the Ganglion Cell-Inner Plexiform Layer Thickness Measured With Cirrus HD-OCT and Macular Visual Field Sensitivity Measured With Microperimetry. Invest Opthalmol Vis Sci. 2013;54(4):3046. doi:10.1167/iovs.12-11173
4. Raza AS, Cho J, Moraes CGV de, Wang M, Zhang X, Kardon RH, et al. Retinal Ganglion Cell Layer Thickness and Local Visual Field Sensitivity in Glaucoma. Arch Ophthalmol. 2011;129(12):1529. doi:10.1001/archophthalmol.2011.352
5. 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. doi:10.1016/s0161-6420(00)00284-0
6. Gardiner SK, Johnson CA, Cioffi GA. Evaluation of the structure-function relationship in glaucoma. Invest Opthalmol Vis Sci. 2005;46(10):3712. doi:10.1167/iovs.05-0266
7. Lee J, Morales E, Sharifipour F, Amini N, Yu F, Afifi AA, et al. The relationship between central visual field sensitivity and macular ganglion cell/inner plexiform layer thickness in glaucoma. Br J Ophthalmol. 2017;101(8):1052-1058. doi:10.1136/bjophthalmol-2016-309208
8. Brigatti L, Caprioli J. Correlation of visual field with scanning confocal laser optic disc measurements in glaucoma. Arch Ophthalmol. 1995;113(9):1191. doi:10.1001/archopht.1995.01100090117032
9. Weinreb RN, Shakiba S, Sample PA. Association between quantitative nerve fiber layer measurement and visual field loss in glaucoma. Am J Ophthalmol. 1995;120(6):732-738. doi:10.1016/s0002-9394(14)72726-6

10. Iester M, Mikelberg FS, Courtright P, Drance SM. Correlation between the visual field indices and Heidelberg retina tomograph parameters. J Glaucoma. 1997;6(2):78???82. doi:10.1097/00061198-199704000-00002
11. Teesalu P, Vihanninjoki ,K Airaksinen P, Tuulonen, A Läärä E. Correlation of blue-on-yellow visual fields with scanning confocal laser optic disc measurements. Invest Ophthalmol Vis Sci. 1997;38(12):2452-2459
12. Guo Z, Kwon YH, Lee K, Wang K, Wahle A, Alward WLM, et al. Optical Coherence Tomography Analysis Based Prediction of Humphrey 24–2 Visual Field Thresholds in Patients With Glaucoma. Invest Opthalmol Vis Sci. 2017;58(10):3975. doi:10.1167/iovs.17-21832
13. Bogunovic H, Kwon YH, Rashid A, Lee K, Critser DB, Garvin MK, et al. Relationships of retinal structure and Humphrey 24–2 visual field thresholds in patients with glaucoma. Invest Ophthalmol Vis Sci. 2014;56(1):259-271. doi:10.1167/iovs.14-15885
14. 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. doi:10.1364/boe.2.001734
15. LeCun Y, Bengio Y , Hinton, G. Nature. 2015;521(7553):436-444. doi:10.1038/nature14539.
16. Christopher M, Belghith A, Bowd C et al. Performance of Deep Learning Architectures and Transfer Learning for Detecting Glaucomatous Optic Neuropathy in Fundus Photographs. Sci Rep. 2018;8(1). doi:10.1038/s41598-018-35044-9
17. Abràmoff M, Lou Y, Erginay A et al. Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning. Invest Opthalmol Vis Sci. 2016;57(13):5200. doi:10.1167/iovs.16-19964
18. Shibata N, Tanito M, Mitsuhashi K et al. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. Sci Rep. 2018;8(1). doi:10.1038/s41598-018-33013-w
19. Li Z, He Y, Keel S, Meng W, Chang R, He M. Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs. Ophthalmology. 2018;125(8):1199-1206. doi:10.1016/j.ophtha.2018.01.023
20. Lazaridis G, Lorenzi M, Ourselin S, Garway-Heath DF. Enhancing OCT Signal by Fusion of GANs: Improving Statistical Power of Glaucoma Clinical Trials. MICCAI. 2019;11764:1–9. https://doi.org/10.1007/978-3-030-32239-7_1
21. Lazaridis G, Mohamed-Noriega J, Aguilar-Munoa S, Suzuki K, Nomoto H, Garway-Heath D. Imaging Outcomes in Clinical Trials of Treatments for Glaucoma. Ophthalmology. 2020. doi:10.1016/j.ophtha.2020.11.027

22. Lazaridis G, Lorenzi M, Ourselin S, Garway-Heath D. Improving statistical power of glaucoma clinical trials using an ensemble of cyclical generative adversarial networks. Med Image Anal. 2021;68:101906. doi:10.1016/j.media.2020.101906
23. De Fauw J, Ledsam J, Romera-Paredes B et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med. 2018;24(9):1342-1350. doi:10.1038/s41591-018-0107-6.
24. Kermany D, Goldbaum M, Cai W et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell. 2018;172(5):1122-1131.e9. doi:10.1016/j.cell.2018.02.010
25. Muhammad H, Fuchs T, 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
26. Devalla S, Chin K, Mari J et al. A Deep Learning Approach to Digitally Stain Optical Coherence Tomography Images of the Optic Nerve Head. Invest Opthalmol Vis Sci. 2018;59(1):63. doi:10.1167/iovs.17-22617
27. Lazaridis G, Xu M, Afgeh SS, Montesano G, Garway-Heath D. (2020) Bio-inspired Attentive Segmentation of Retinal OCT Imaging. OMIA 2020:12069. https://doi.org/10.1007/978-3-030-63419-3_1
28. Poplin R, Varadarajan A, Blumer K et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng. 2018;2(3):158-164. doi:10.1038/s41551-018-0195-0
29. Zhu H, Crabb D, Schlottmann P et al. Predicting Visual Function from the Measurements of Retinal Nerve Fiber Layer Structure. Invest Opthalmol Vis Sci. 2010;51(11):5657. doi:10.1167/iovs.10-5239
30. Hashimoto Y, Asaoka R, Kiwaki T, Sugiura H, Asano S, Murata H, Fujino Y, Matsuura M, Miki A, Mori K, Ikeda Y. Deep learning model to predict visual field in central 10° from optical coherence tomography measurement in glaucoma. Br J Ophthalmol. 2021 Apr 1;105(4):507-13.
31. Park K, Kim J, Lee J. A deep learning approach to predict visual field using optical coherence tomography. PLoS One. 2020;15(7):e0234902. doi:10.1371/journal.pone.0234902.
32. Asaoka R, Murata H, Matsuura M, Fujino Y, Yanagisawa M, Yamashita T. Improving the Structure–Function Relationship in Glaucomatous Visual Fields by Using a Deep Learning–Based Noise Reduction Approach. Ophthalmo Glaucoma. 2020;3(3):210-217. doi:10.1016/j.ogla.2020.01.001

33. Mariottoni EB, Datta S, Dov D, Jammal AA, Berchuck SI, Tavares IM, Carin L, Medeiros FA. Artificial Intelligence Mapping of Structure to Function in Glaucoma. Trans. Vis. Sci. Tech. 2020;9(2):19. doi: https://doi.org/10.1167/tvst.9.2.19
34. Montesano G, Bryan S, Crabb D et al. A Comparison between the Compass Fundus Perimeter and the Humphrey Field Analyzer. Ophthalmology. 2019;126(2):242-251. doi:10.1016/j.ophtha.2018.08.010
35. Garway-Heath DF, Zhu H, Cheng Q et al. Combining optical coherence tomography with visual field data to rapidly detect disease progression in glaucoma: a diagnostic accuracy study. Health Technol Assess 2018;22(4):1-106. doi:10.3310/hta22040
36. Garway-Heath DF, Crabb DP, Bunce C et al. Latanoprost for open-angle glaucoma (UKGTS): a randomised, multicentre, placebo-controlled trial. The Lancet. 2015;385(9975):1295-1304. doi:10.1016/s0140-6736(14)62111-5
37. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings. 2010:249-256
38. Kingma DP, Welling M. Auto-Encoding Variational Bayes. arXiv [statML]. Published online 2013. http://arxiv.org/abs/1312.6114
39. Antelmi L, Ayache N, Robert P, Lorenzi M. Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data. PMLR. Vol 97. 2019:302-311.
40. Wolpert DH. Stacked generalization. Neural Netw. 1992;5(2):241-259.
41. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proc. 22nd ACM SIGKDD on KDD. ACM Press; 2016;785–794. https://doi.org/10.1145/2939672.2939785
42. Bowd C, Chan K, Zangwill LM, et al. Comparing neural networks and linear discriminant functions for glaucoma detection using confocal scanning laser ophthalmoscopy of the optic disc. Invest Ophthalmol Vis Sci. 2002;43(11):3444-3454..
43. Goldbaum MH, Sample PA, White H, et al. Interpretation of automated perimetry for glaucoma by neural network. Invest Ophthalmol Vis Sci. 1994;35(9):3362-3373.
44. Bengtsson B, Bizios D, Heijl A. Effects of Input Data on the Performance of a Neural Network in Distinguishing Normal and Glaucomatous Visual Fields. Investi Opthalmol Vis Sci. 2005;46(10):3730-3736. doi:10.1167/iovs.05-0175
45. Brigatti L, Hoffman D, Caprioli J. Neural networks to identify glaucoma with structural and functional measurements. Am J Ophthalmol. 1996;121(5):511-521. doi:10.1016/s0002-9394(14)75425-x

46. Uchida H, Brigatti L, Caprioli J. Detection of structural damage from glaucoma with confocal laser image analysis. Invest Ophthalmol Vis Sci. 1996;37(12):2393–2401.
47. Brigatti L, Nouri-Mahdavi K, Weitzman M, Caprioli J. Automatic detection of glaucomatous visual field progression with neural networks. Arch Ophthalmol. 1997;115(6):725-728.
48. Spenceley S, Henson D, Bull D. Visual field analysis using artificial neural networks. *Ophthalmic Physiol Opt*. 1994;14(3):239-248. doi:10.1111/j.1475-1313.1994.tb00004.x
49. 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. doi:10.1016/j.ophtha.2019.09.036
50. 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;43:2654–2659.
51. Yanagisawa M, Tomidokoro A, Saito H et al. Atypical retardation pattern in measurements of scanning laser polarimetry and its relating factors. Eye. 2008;23(9):1796-1801. doi:10.1038/eye.2008.365
52. Zhu H, Crabb DP, Fredette MJ, Anderson DR, Garway-Heath DF. Quantifying Discordance Between Structure and Function Measurements in the Clinical Assessment of Glaucoma. Arch Ophthalmol. 2011;129(9):1167. doi:10.1001/archophthalmol.2011.112
53. Gardiner S, Demirel S, Reynaud J, Fortune B. Changes in Retinal Nerve Fiber Layer Reflectance Intensity as a Predictor of Functional Progression in Glaucoma. Invest Opthalmol Vis Sci. 2016;57(3):1221. doi:10.1167/iovs.15-18788
54. van der Schoot J, Vermeer K, de Boer J, Lemij H. The Effect of Glaucoma on the Optical Attenuation Coefficient of the Retinal Nerve Fiber Layer in Spectral Domain Optical Coherence Tomography Images. Invest Opthalmol Vis Sci. 2012;53(4):2424. doi:10.1167/iovs.11-8436
55. Maul EA, Friedman DS, Chang DS, et al. Choroidal thickness measured by spectral domain optical coherence tomography: factors affecting thickness in glaucoma patients. Ophthalmology. 2011;118(8):1571-1579. doi: 10.1016/j.ophtha.2011.01.016
56. Qiu K, Schiefer J, Nevalainen J, Schiefer U, Jansonius N. Influence of the Retinal Blood Vessel Topography on the Variability of the Retinal Nerve Fiber Bundle Trajectories in the Human Retina. Invest OpthalmolVis Sci. 2015;56(11):6320. doi:10.1167/iovs.15-17450

57. Lamparter J, Russell R, Zhu H et al. The Influence of Intersubject Variability in Ocular Anatomical Variables on the Mapping of Retinal Locations to the Retinal Nerve Fiber Layer and Optic Nerve Head. Invest Opthalmol Vis Sci. 2013;54(9):6074. doi:10.1167/iovs.13-11902

58. Altman DG, Royston P. What do we mean by validating a prognostic model. Stat Med. 2000;19(4):453–473.

Appendix

Figure 1 shows the encoding of the test set in the latent space given by MC-VAE. We limit the visualization to the 2D subspace generated by the three dimensions. The subspace shows that the two different channels (RNFLT values and OCT images) are correlated with the VF channel (channel 0).

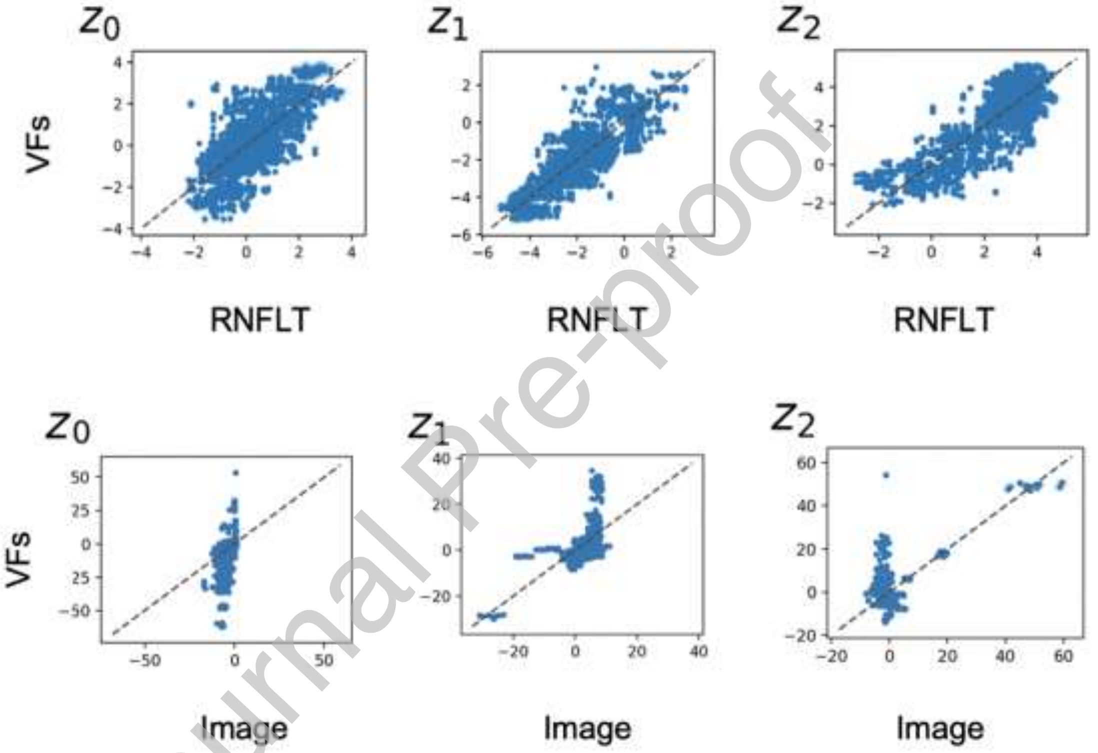


Figure A1 Projection of the RAPID subjects (test data) in the sparse latent subspace inferred from the first three least dropped out dimensions. Top: Y-axis: VFs, X-axis: RNFLT channel. Bottom: Y-axis: VFs, X-axis: OCT image channel.

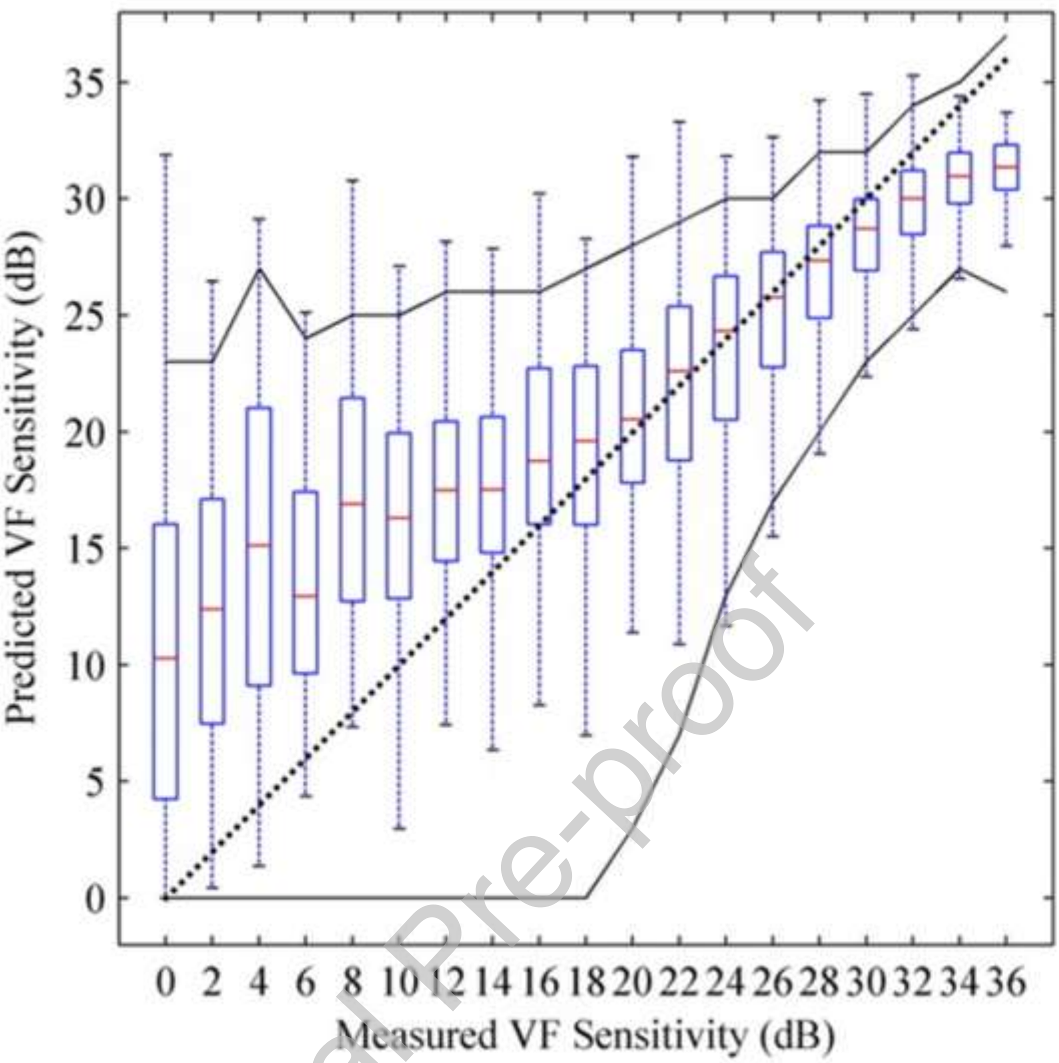


Figure A2 Distributions of the error between the predicted and the measured sensitivity for each VF location in 306 eyes from the BMES data, stratified by VF sensitivity. Each error bar summarizes the predictive performance over a 2-dB range from 0 to >36 dB. Thin vertical lines: 90% prediction limits (5th and 95th percentile of error), each box indicates the interquartile range of the prediction error (25th and 75th percentile error) with the line in the box indicating the median error. The dotted line of unity indicates perfect prediction (no error). The predictive performances of the BRBF model is shown. (solid lines) Previously published (5th and 95th percentiles) test–retest limits for VF data derived from the point-wise differences between two VFs tested over a short period. Reproduced with permission from [29].

Table 2 Principal baseline characteristics for the COMPASS and RAPID cohorts. Age is a subject variable; IOP, refractive error, and SAP MD, and RNFL thickness are eye variables. Data are provided for eligible eyes, n = number; D = dioptres; dB = decibel; mmHg = millimetres of mercury; IOP = intraocular pressure; SAP = standard automated perimetry; MD = mean deviation

	Training dataset				Test dataset	
	Healthy, n = 421 eyes		Glaucoma, n = 533 eyes		Glaucoma, n = 144 eyes	
	Median	5th to 95th percentile	Median	5th to 95th percentile	Median	5th to 95th percentile
Age (years)	46.5	29.7 – 63.0	70.8	61.8 - 77.3	70.3	50 – 85.6
IOP (mmHg)	15	13 - 16	14	13 - 16	14	8 – 21
Refractive Error (D)	-0.12	-1.75 - 0	-0.12	-1 - 0.62	-0.13	-7.48 – 2.95
RNFL thickness (μ)	99.2	92.0 - 105.4	70.4	56.8 - 81.4	69	45.1 – 95.6
SAP MD (dB)	-0.92	-1.84 - -0.15	-5.26	-11.22 - -2.01	-4.17	-14.22 – 0.88

Table 2: Quantification of pairwise and Best Available Estimate (BAE) pointwise prediction errors for each method.
RNFLT: retinal nerve fiber layer thickness. OCT: optical coherence tomography. MAE: Mean Absolute Error. SD: Standard Deviation of AE. dB: Decibels. BRBF: Bayesian Radial Basis Function

Error	**Pairwise (dB)**		**BAE (dB)**	
Method	**MAE**	**SD**	**MAE**	**SD**
Linear	5.5	6.4	5.1	6.1
BRBF	3.9	4.7	3.4	4.4
Model 1 (RNFLT only)	3.6	4.6	3.0	3.9
Model 2 (RNFLT + OCT image)	2.8	3.7	2.3	3.1

A custom deep learning architecture to predict VF from SDOCT was designed and validated. The method was developed on a training dataset and tested in an independent test-retest dataset; ~10 VFs per eye were used to provide a ‘best available estimate’ VF, thus removing noise originating from the VF which would otherwise have contributed to prediction error. Predictions from SDOCT images approached the accuracy of single real VF estimates of the ‘best available estimate’retinal sensitivity.