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 [5^{th} to 95^{th} percentile 8.0 to 21.0 mmHg] and VF MD -4.17 dB [5^{th} to 95^{th} 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)

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	5^{th} to 95^{th} percentile	Median	5^{th} to 95^{th} percentile	Median	5^{th} to 95^{th} 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 / Method	Pairwise (dB)		BAE (dB)	
	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.