

# Automated classification of normal and Stargardt disease optical coherence tomography images using deep learning

Mital Shah BM BS<sup>1,2</sup>, Ana Roomans Ledo MEng<sup>3</sup>, Jens Rittscher PhD<sup>3</sup>.

<sup>1</sup> Oxford Eye Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, UK

<sup>2</sup> Nuffield Laboratory of Ophthalmology, Nuffield Department of Clinical Neurosciences,  
University of Oxford, Oxford, UK

<sup>3</sup> Institute of Biomedical Engineering, Department of Engineering Science, University of  
Oxford, Oxford, UK

Corresponding Author: Dr Mital Shah, Oxford Eye Hospital, John Radcliffe Hospital, Oxford  
OX3 9DU, UK; Email: [mital.shah@ndcn.ox.ac.uk](mailto:mital.shah@ndcn.ox.ac.uk); Mobile: +447958 606 176

## Abstract

**Purpose:** Recent advances in deep learning has seen an increase in its application to automated image analysis in ophthalmology for conditions with a high prevalence. We wanted to identify whether deep learning could be used for the automated classification of optical coherence tomography (OCT) images from patients with Stargardt disease (STGD) using a smaller dataset than traditionally used.

**Methods:** Sixty participants with STGD and 33 participants with a normal retinal OCT were selected and a single OCT scan containing the centre of the fovea was selected as the input data. Two approaches were used: Model 1 – a pre-trained convolutional neural network (CNN); Model 2 – a new CNN architecture. Both models were evaluated on their accuracy, sensitivity, specificity and Jaccard Similarity score (JSS).

**Results:** 102 OCT scans from participants with a normal retinal OCT and 647 OCT scans from participants with STGD were selected. The highest results were achieved when both models were implemented as a binary classifier: Model 1 - accuracy 99.6%, sensitivity 99.8%, specificity 98.0%, JSS 0.990; Model 2 - accuracy 97.9%, sensitivity 97.9%, specificity 98.0%, JSS 0.976.

**Conclusion:** The deep learning classification models used in this study were able to achieve high accuracy despite using a smaller dataset than traditionally used and are effective in differentiating between normal OCT scans and those from patients with STGD. This preliminary study provides promising results for the application of deep learning to classify OCT images from patients with inherited retinal diseases.

**Key words:** machine learning, deep learning, optical coherence tomography, Stargardt disease, retinal degeneration, image analysis.

## Introduction

Retinal imaging plays an essential role in the diagnosis and clinical management of patients with retinal disease. Optical coherence tomography (OCT) is now the most commonly used retinal imaging modality in ophthalmology and the number being performed in the United Kingdom National Health Service and United States Medicare populations is increasing each year (Centers for Medicare and Medicaid Services n.d., NHS Digital n.d.). OCT has become crucial in evaluation of the retina: for diagnostic purposes, for making treatment decisions and assessing treatment response, and in monitoring patients for signs of disease progression. Both the increase in utility of OCTs and their importance in the management of patients with retinal disease emphasises the increasing time spent by clinicians in the interpretation of these images.

OCTs contain a wealth of information, however, in clinical practice the majority of this information is not fully utilised as image analysis is performed using manual techniques. This is due to a lack of the sophisticated techniques that are required for the automated analysis of these images. Machine learning techniques provide a unique opportunity for the automated analysis of retinal images. Unlike traditional programming that uses explicit instructions to perform a task, machine learning algorithms build statistical models to learn patterns within sample data (known as training data) in order to perform a specific task. Deep learning is a subfield of machine learning based on artificial neural networks, algorithms inspired by the biological neural networks of the brain. The word “deep” refers to the number of layers within the artificial neural network through which data are transformed. Traditional machine learning approaches including deep learning require large volumes of training data. However, acquiring large volumes of high-quality data that are representative of the patient cohorts seen in real-world healthcare settings is difficult. These datasets also require accompanying annotations

generated by experts which are time consuming to produce. Not only are these large volumes of data difficult to acquire but this is not practical for retinal diseases with a low prevalence.

Visual impairment caused by inherited retinal diseases (IRD) is increasing and they are now the commonest cause of blindness in England and Wales in working age adults (Liew, Michaelides & Bunce 2014) and a leading cause of visual impairment in children (Rahi & Cable 2003). A number of therapeutic strategies are being developed for IRDs with significant international interest in this field of research, evidenced by the number of ongoing treatment trials (Duncan et al. 2018, European Medicines Agency n.d., Hafler 2017, U.S. National Library of Medicine n.d.). The transition zone from an area of relatively normal to degenerating retina in patients with IRDs, where the leading disease edge is located, is important for many therapeutic approaches. OCT allows this transition zone to be studied through direct observation of *in vivo* structural changes within retinal layers.

Automated image analysis approaches have been successfully demonstrated using colour and OCT images for a number of ophthalmic diseases including diabetes and age related macular degeneration (Abràmoff et al. 2018, De Fauw et al. 2018, Lee, Baughman & Lee 2017, Li et al. 2018). One of the challenges of using machine learning approaches in a healthcare setting is to do so with a smaller volume of data than traditionally used. We describe a preliminary study to evaluate the use of deep learning for the automated classification of OCT images from a carefully annotated cohort of patients with Stargardt disease (STGD).

## Methods

This study was approved by the London Surrey Borders Research Ethics Committee (reference 17/LO/1753) and was conducted in adherence to the tenets of the Declaration of Helsinki.

### Patient selection

Patients with a clinical diagnosis of STGD were included in this study (Table 1). Patients with no evidence of retinal disease as determined by a retinal specialist were defined as controls.

### OCT selection

OCTs (Spectralis; Heidelberg Engineering, Heidelberg, Germany) of the macula from the selected participants performed between October 2008 and September 2017 were extracted. There was no evidence of other concomitant retinal pathology in the patients identified with STGD. Extracted macular OCTs containing at least one OCT scan providing a cross section of the fovea were included in this study. As retinal degeneration in STGD proceeds centrifugally from the fovea, the OCT scan containing the centre of the fovea from each macular OCT was selected as the input data.

### Deep learning models

A number of mathematical approaches (Chen et al. 2012, Kafieh, Rabbani & Kermani 2013, Tian et al. 2016) have been developed for extracting quantitative information from OCT data. Here, image noise and the presence of artefacts which are inherent to OCT imagery pose a significant challenge. Deep learning provides an effective way of learning such features directly from the image data.

While the advantages of these approaches are well known, we also have to take the limited availability of training data and the interpretability of the results into account. Most deep learning approaches only provide an end-to-end classification pipeline and only provide limited information on how a given case has been classified (output prediction). An end-to-end classification pipeline directly converts input data into an output prediction, bypassing the intermediate steps of a traditional pipeline. To address this concern we are proposing to fit a model that uses deep learning to analyse a set of predefined image regions. In this study we contrast this approach with a more naïve deep learning approach and compare these methods to evaluate the suitability of deep learning for the classification of OCT images from patients with STGD.

### *Model 1 – pre-trained model (VGG19)*

#### *Input data*

OCT scans were cropped to a size of 256 x 256 pixels with the fovea at the centre and were labelled as either normal, mild STGD or severe STGD by a retinal specialist.

A trinary classification system (normal, mild and severe STGD) was implemented to capture the spectrum of retinal degeneration which the scans exhibited, while avoiding misclassification of scans with mild disease. Scans at this stage of disease exhibit features common to normal and severely degenerated retinas, and with a binary classification of normal or diseased would result in a convolutional neural network (CNN) output probability of [0.5, 0.5]. A CNN is a class of artificial neural network that is most commonly used for image recognition and classification. The word “convolution” in “convolutional neural network”

refers to the mathematical convolution operation, which preserves the spatial relationship between image pixels by using small squares of input data to learn image features.

Data augmentation was used to increase the effective amount of training data by 10-fold and reduce overfitting of the final model. This was achieved through a combination of image translation, reflection on the y-axis and rotation in a range of  $30^\circ$ ; in such a way as to preserve clinical relevance in the set of resulting images.

### *Convolutional neural network*

The VGG19 CNN architecture (Simonyan & Zisserman 2014), pre-trained on the ImageNet dataset, was used and implemented as a fixed feature extractor where only the end layers were randomly initialised and re-trained (Fig. 1). ImageNet is a large-scale database consisting of over 14 million natural images with accompanying human annotations and is used to train many image classification networks. A pre-trained image classification network is a network that has already learned to extract useful features from natural images. A pre-trained CNN can be used as a starting point to learn a new specific task and is generally faster and easier than training a CNN from scratch. Training was run on a NVIDIA GeForce GTX 1080 Ti GPU. The key consideration was to determine at what depth ( $L$ ) of the model the filters still extracted relevant features in the OCT scans, and hence the number of layers of the model that should be frozen. Under the assumption that this point would be that which would yield the highest accuracy, a quantified comparison was made between five potential freezing points using 4-fold cross-validation. Retaining all the convolutional blocks ( $L = 5$ ) provided the most accurate results. The network was trained using a batch size of 16 in 50 epochs. One epoch is when an entire training dataset has passed forwards and backwards through the neural network, updating

the internal network parameters. The batch size represents the number of training samples in one forward and backward pass. An epoch usually comprised of more than one batch.

## Model 2 – new classification model

### *Input data*

In order to circumvent the limitation of a small dataset, the data were augmented in two ways: firstly, by dividing each OCT scan into 256 x 32 pixel columns and using each column as an individual input; and then with vertical translation of the column inputs. This increased the size of the dataset by approximately 30-fold.

As each input represents 256 x 32 pixels of the OCT scan, the localised information contained within each column allows the overall shape of the retina to be captured; and the extent and distribution of retinal degeneration within each scan to be represented. A modified classification originally described by Lazow *et al.* (Lazow et al. 2011) to identify the transition zone from OCTs in patients with STGD was used to label each column into one of three classes. The criteria used for this classification system is shown in Fig. 2.

### *Convolutional neural network*

A new CNN with a simple architecture similar to LeNet (LeCun et al. 1998), requiring approximately 3 million parameters to be trained, was used (Fig. 3). Training was run on a NVIDIA GeForce GTX 1080 Ti GPU. Weights were initialised by drawing samples from a truncated normal distribution and were optimised using adaptive moment estimation. In order



to improve the stability and performance of the network and ensure it did not overfit the training data, batch normalisation was applied prior to the activation functions and regularisation was implemented in the form of Dropout, with a keep probability of 0.9. The network was trained using a batch size of 50 in 300 epochs. The same grayscale image comprised all three RGB channels.

### *OCT classification*

#### *Weighted Voting*

A weighted voting system was implemented to combine the probabilistic output of Model 2 for each input column into a single classification for the entire OCT scan such that  $\mathbf{p}_{scan} = \sum_{i=1}^n w_i \cdot \mathbf{p}_i \quad \forall \text{ columns } i$ , where  $n$  is the total number of columns. Each OCT scan was then categorised by assigning it to the class with the maximum weighted probability. This categorisation was done in two ways: as binary classification into two classes (normal and STGD); and as trinary classification into three classes (normal, mild and severe STGD). The weights  $\mathbf{w}$  were hand-designed to be location specific and to vary linearly from 0.5 at the column furthest from the fovea to 1.0 at the fovea.

#### *Profile Analysis*

When the weighted probability of the OCT scan was evenly distributed between the mild and severe STGD classes, the CNN classification of each input column within the OCT scan was analysed. At the level of the OCT scan, each stage of retinal degeneration (mild and severe) was defined by the number of columns belonging to each of the three classes and their spatial distribution; a value in the range  $[-1, 1]$  was assigned to each column based on its CNN

classification and spatial location relative to the fovea. The final value assigned to each OCT scan was created by combining the values for all columns, with its sign determining the final classification, and its distance from 0 the confidence of the assigned class.

### Evaluation metrics

Both CNN models were assessed based on their accuracy, sensitivity, specificity and Jaccard similarity score. The Jaccard similarity score is a statistic used to quantify the performance of the CNN models, with a range between 0 and 1. The higher the statistic, the more accurate the CNN model is. The classification of individual column inputs and the OCT scan with weighted voting from Model 2 were also evaluated by comparison of their Receiver Operator Characteristic (ROC) curves.

## **Results**

The data used to train and test the algorithms were composed of 102 OCT scans from participants with a normal retinal OCT and 647 OCT scans from participants with STGD, which were divided into 110 mild cases and 537 severe cases. Of these, 280 OCT scans (80 from normal, 80 from mild STGD and 120 from severe STGD) were used for training with an 80:20 split between training and validation data and 5-fold cross-validation; and the remaining 469 OCT scans (22 from normal, 30 from mild STGD and 417 from severe STGD) were used for testing. The size of this data set is sufficient to evaluate the developed algorithms and provide the necessary validation data. Performance is evaluated through comparison of the network's output to the ground truth, set by clinical diagnosis and annotation. Four metrics are used for this purpose: accuracy, sensitivity, specificity and the Jaccard Similarity score.

### Model 1

Training on top of the pre-trained VGG19 model required under 50 epochs to converge to its top accuracy of 99.8% on validation data (Fig. 4A). When the trained model was run on a separate test dataset, it output results with 99.0% accuracy, a sensitivity of 96.0% in mild STGD and 99.5% in severe STGD, specificity of 98.0%, and a mean Jaccard Similarity score of 0.958. When implemented as a binary classifier, it output results with an accuracy of 99.6%, sensitivity of 99.8%, specificity of 98.0%, and a Jaccard Similarity score of 0.990.

### Model 2

Training Model 2 required over 250 epochs to reach over 90% accuracy on validation data (Fig. 4B). Results from the trained model when run on a separate test dataset are shown in Table 2. ROC curves from binary classification at the level of individual column inputs and the OCT scan with weighted voting are shown in Fig. 5A, with an area under the ROC curve of 98.0% and 99.7%, respectively. ROC curves from trinary classification of disease severity at the level of the OCT scan after weighted voting and profile analysis are shown in Fig. 5B.

## **Discussion**

This preliminary study demonstrates the effective use of deep learning for the classification of OCT images from controls and patients with STGD with a smaller volume of data than traditionally used. Two different deep learning approaches were successfully used in this study with Model 1 providing an accuracy of 99.6% and Model 2 providing an accuracy of 85.3% at distinguishing STGD from normal OCT images, however, when aggregating weighted probabilities across an entire OCT scan, the accuracy of Model 2 improved to 98.0%. When

distinguishing the severity of STGD from OCT scans, Model 1 provided an accuracy of 99.3% and Model 2 provided an accuracy of 76.8%, which improved to 92.3% at the level of the OCT scan.

One of the limitations of using a small dataset includes a limited ability of the classification model to differentiate OCT scans with a milder disease phenotype. This may be because STGD only affects the entire macula in very severe disease and therefore OCT scans will contain features common to normal and severely degenerated retinas. In order to overcome this limitation the data for Model 2 were augmented by dividing each OCT scan into 256 x 32 pixel columns and using each column as an individual input. This approach enables the classification model output to provide information on a smaller retinal area, rather than the entire OCT scan length, which will represent a wider disease phenotype spectrum. The sensitivity of Model 2 to a mild disease phenotype was 19.3%, with a significant proportion of mild phenotypes being wrongly classified as severe. By combining the CNN output probabilities of the column inputs that make up an entire OCT scan with weighted voting and profile analysis, the sensitivity of Model 2 to mild disease phenotypes improved to 60.9%.

There has been a recent increase in the application of machine learning and deep learning in ophthalmology. Deep learning has been used for the automated detection of a number of ophthalmic diseases including diabetic retinopathy (Abràmoff et al. 2018, Lee et al. 2017), age related macular degeneration (Grassmann et al. 2018, Lee, Baughman & Lee 2017) and glaucoma (Asaoka et al. 2019, Li et al. 2018). Most of these studies report sensitivities and specificities of greater than 90% while using datasets consisting of thousands of images. This preliminary study uses two different approaches to classify OCT images in patients with STGD. Model 1, a naïve deep learning approach; and Model 2, that uses deep learning to

294 analyse a set of predefined image regions, both achieve similarly high sensitivities and  
295 specificities with binary classification. To date, there has only been one study reporting the use  
296 of deep learning for the classification of OCT images in patients with IRDs (Fujinami-  
297 Yokokawa et al. 2019); the study by Fujinami-Yokokawa *et al.* used a commercially available  
298 deep learning platform, which utilises the Inception-v3 CNN (Szegedy et al. 2016), to classify  
299 OCT images from patients with three different IRDs and reported a mean overall test accuracy  
300 of 90.9% (Fujinami-Yokokawa et al. 2019). Training and testing data were composed of 178  
301 OCT macular images from three different devices. There are significant differences in OCT  
302 image characteristics between different device types (De Fauw et al. 2018); it is therefore  
303 possible that the high reported accuracy is related to learning the differences in the image  
304 characteristics between device types rather than disease features. Fujinami-Yokokawa *et al.* do  
305 not report on how these differences in the input data were accounted for during the training and  
306 testing of their model. In this study, OCT images from one device were used for training and  
307 testing the classification models. While this precludes classification due to differences in image  
308 characteristics from different device types, it does limit the generalisability of the classification  
309 models. The objective of a CNN is to produce a generalisable model that performs well on new  
310 data on which the model will make predictions. Overfitting of the model occurs when it  
311 correspond too closely to the data used for training and fails to reliably make predictions on  
312 new data. In the study by Fujinami-Yokokawa *et al.* a mean training accuracy of 96.9% is  
313 reported, with two of their four experiments attaining a training accuracy of 100% (Fujinami-  
314 Yokokawa et al. 2019), however, test accuracy showed significant variation between the four  
315 classes, suggesting overfitting of their model. In this study, Model 1 achieved an accuracy of  
316 99.8% and over 99.0% on validation and test data, respectively, suggesting that the model is  
317 not overfitting. Model 2 achieved an accuracy of 90.0% on validation data, and 85.2% and  
318 76.8% on test data as a binary and trinary classifier, respectively. This suggests that overfitting

of Model 2 as a trinary classifier may have occurred and also could have been impacted by the relative underrepresentation OCT images with a mild disease phenotype. However, with weighted voting and profile analysis, Model 2 achieved a top accuracy of 97.9% and 92.2% as a binary and trinary classifier, respectively.

OCT is essential in evaluation of the retina in patients with retinal disease. The increasing utility of OCTs yields a great deal of information that clinicians are required to comprehensively analyse within a short space of time. The deep learning approaches used in this preliminary study could be utilised in the development of programmes to assist clinicians with the diagnosis and treatment of macular diseases, similar to computer-aided detection systems that have been developed to assist radiologists in the interpretation of medical images. They may also be applied to retinal diseases with a lower prevalence where it is not possible to obtain the large datasets with accompanying annotations that are traditionally used; or to identify the transition zone in patients with STGD. By using inputs that represent smaller retinal areas, similar to those used in this study, it may be possible to automatically identify the leading disease edge where a transition from normal to degenerated retina occurs. This is important as it will help to inform patient selection and optimal retinal locations to target for therapeutic interventions such as gene therapy. By incorporating temporal data into the model, this approach may allow the leading disease front to be automatically identified and monitored over time. Observation of the sequence of intra-retinal structural changes that occur in the transition zone over time may be helpful in providing information: to patients on disease progression, on the efficacy of treatment outcomes from interventional clinical trials, and on disease mechanisms.

The deep learning classification models from this study were trained and tested on OCT images from a single centre and device type, using limited training data that were selected to meet the study inclusion criteria; therefore their generalisability to OCT images from other centres is unknown. Once an independent validation dataset becomes available, we will be able to test the models to evaluate their generalisability. Participants with a normal retinal OCT were not age-matched to those with STGD. Retinal layer thickness measured on OCT has been shown to vary with age (Nieves-Moreno et al. 2018) and therefore may have impacted OCT classification using the deep learning classification models. Although Model 2 provided an accuracy of 92.2% for distinguishing the severity of STGD from OCT scans using weighted voting and profile analysis, the highest sensitivity achieved for distinguishing mild disease was 60.9%. Further work is required to refine the severity grading using this approach. In the future more modern CNN architectures, such as the Inception network (Szegedy et al. 2016) with its improved speed and accuracy, will be used to improve the results. Future work including genetic data with sufficient patient numbers to represent the significant phenotypic and genetic heterogeneity of STGD may help with phenotype-genotype correlation.

In this preliminary study we demonstrate that it is possible to use deep learning classification models to differentiate between normal OCT images and those from patients with STGD, and distinguish the severity of STGD from OCT images; using a smaller dataset than traditionally used. The use of more modern CNN architectures and independent datasets for evaluation will improve results; and prospective studies will be crucial for clinical translation. We hypothesise that the general principle demonstrated in this study can be applied not only to other IRD, but also to other retinal diseases with a low prevalence. This preliminary study has therefore laid the foundation for further work to develop automated image analysis techniques using deep

367 learning for the interpretation of OCTs and other retinal imaging data in retinal diseases with  
368 a low prevalence.

369



370 **Financial Support:** Supported by OHSRC part of Oxford Hospitals Charity and the Thames  
371 Valley & South Midlands Clinical Research Network.

372

373 **Conflict of Interest:** no conflicting relationship exists for any author.

374

375 **Acknowledgments:** The authors would like to thank Professor Susan Downes for her  
376 assistance in identifying patients for this study.

377

378 **References**

- 379 Abramoff MD, Lavin PT, Birch M, Shah N & Folk JC (2018): Pivotal trial of an autonomous  
380 AI-based diagnostic system for detection of diabetic retinopathy in primary care offices.  
381 npj Digit Med **1**: 39.
- 382 Asaoka R, Murata H, Hirasawa K, et al. (2019): Using Deep Learning and Transfer Learning  
383 to Accurately Diagnose Early-Onset Glaucoma From Macular Optical Coherence  
384 Tomography Images. Am J Ophthalmol **198**: 136–145.
- 385 Centers for Medicare and Medicaid Services (n.d.). Medicare Provider Utilization and  
386 Payment Data. Medicare Natl Aggreg table.
- 387 Chen X, Niemeijer M, Zhang L, Lee K, Abramoff MD & Sonka M (2012): Three-  
388 dimensional segmentation of fluid-associated abnormalities in retinal OCT: Probability  
389 constrained graph-search-graph-cut. IEEE Trans Med Imaging **31**: 1521–1531.
- 390 De Fauw J, Ledsam JR, Romera-Paredes B, et al. (2018): Clinically applicable deep learning  
391 for diagnosis and referral in retinal disease. Nat Med **24**: 1342–1350.
- 392 Duncan JL, Pierce EA, Laster AM, et al. (2018): Inherited Retinal Degenerations: Current  
393 Landscape and Knowledge Gaps. Transl Vis Sci Technol **7**: 6.
- 394 European Medicines Agency (n.d.). European Union Clinical Trials Register. Eur Union Clin  
395 Trials Regist.
- 396 Fujinami-Yokokawa Y, Pontikos N, Yang L, et al. (2019): Prediction of Causative Genes in  
397 Inherited Retinal Disorders from Spectral-Domain Optical Coherence Tomography  
398 Utilizing Deep Learning Techniques. J Ophthalmol **2019**: 1–7.
- 399 Grassmann F, Mengelkamp J, Brandl C, et al. (2018): A Deep Learning Algorithm for  
400 Prediction of Age-Related Eye Disease Study Severity Scale for Age-Related Macular  
401 Degeneration from Color Fundus Photography. Ophthalmology **125**: 1410–1420.
- 402 Hafler BP (2017): Clinical progress in inherited retinal degenerations: Gene therapy clinical

403 trials and advances in genetic sequencing. *Retina* **37**: 417–423.

404 Kafieh R, Rabbani H & Kermani S (2013): A review of algorithms for segmentation of  
 405 optical coherence tomography from retina. *J Med Signals Sens* **3**: 45–60.

406 Lazow MA, Hood DC, Ramachandran R, Burke TR, Wang YZ, Greenstein VC & Birch DG  
 407 (2011): Transition zones between healthy and diseased retina in choroideremia (CHM)  
 408 and stargardt disease (STGD) as compared to retinitis pigmentosa (RP). *Investig*  
 409 *Ophthalmol Vis Sci* **52**: 9581–9590.

410 LeCun Y, Bottou L, Bengio Y & Haffner P (1998): Gradient-based learning applied to  
 411 document recognition. *Proc IEEE* **86**: 2278–2323.

412 Lee CS, Baughman DM & Lee AY (2017): Deep Learning Is Effective for Classifying  
 413 Normal versus Age-Related Macular Degeneration OCT Images. *Ophthalmol Retin* **2**:  
 414 322–327.

415 Lee CS, Tying AJ, Deruyter NP, Wu Y, Rokem A & Lee AY (2017): Deep-learning based,  
 416 automated segmentation of macular edema in optical coherence tomography. *Biomed*  
 417 *Opt Express* **8**: 3440.

418 Li Z, He Y, Keel S, Meng W, Chang RT & He M (2018): Efficacy of a Deep Learning  
 419 System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus  
 420 Photographs. *Ophthalmology* **125**: 1199–1206.

421 Liew G, Michaelides M & Bunce C (2014): A comparison of the causes of blindness  
 422 certifications in England and Wales in working age adults (16-64 years), 1999-2000  
 423 with 2009-2010. *BMJ Open* **4**: e004015.

424 NHS Digital (n.d.). Hospital Episode Statistics for England, Outpatient statistics. *Hosp Epis*  
 425 *Stat England, Outpatient Stat*.

426 Nieves-Moreno M, Martínez-de-la-Casa JM, Morales-Fernández L, Sánchez-Jean R, Sáenz-  
 427 Francés F & García-Feijóo J (2018): Impacts of age and sex on retinal layer thicknesses

428 measured by spectral domain optical coherence tomography with Spectralis. PLoS One  
429 **13**: e0194169.

430 Rahi JS & Cable N (2003): Severe visual impairment and blindness in children in the UK.  
431 Lancet **362**: 1359–1365.

432 Simonyan K & Zisserman A (2014): Very Deep Convolutional Networks for Large-Scale  
433 Image Recognition.

434 Szegedy C, Ioffe S, Vanhoucke V & Alemi A (2016): Inception-v4, Inception-ResNet and  
435 the Impact of Residual Connections on Learning. CoRR **abs/1602.0**:

436 Szegedy C, Vanhoucke V, Ioffe S, Shlens J & Wojna Z (2016): Rethinking the Inception  
437 Architecture for Computer Vision. Proc IEEE Comput Soc Conf Comput Vis Pattern  
438 Recognit (Vol. 2016-Decem). 2818–2826.

439 Tian J, Varga B, Tatrai E, Fanni P, Somfai GM, Smiddy WE & Debuc DC (2016):  
440 Performance evaluation of automated segmentation software on optical coherence  
441 tomography volume data. J Biophotonics **9**: 478–489.

442 U.S. National Library of Medicine (n.d.). ClinicalTrials.gov. ClinicalTrials.gov.  
443  
444

## **Figure legends**

### *Figure 1*

Figure describing the architecture of the pre-trained VGG19 convolutional neural network used for Model 1. Each coloured block represents a different layer of the convolutional neural network. INPUT – input image consisting of 3 channels (red, green and blue), CONV – convolutional layer, L – layer depth, P – probability, RELU – rectified linear unit. Class I – normal, Class II – mild Stargardt disease, Class III – severe Stargardt disease.

### *Figure 2*

The criteria used (adapted from Lazow *et al.* (Lazow et al. 2011)) for classification of the optical coherence tomography (OCT) input data for Model 2 into three groups based on the extent of retinal degeneration. Input data was derived by dividing each OCT scan into 256 x 32 pixel columns and using each column as an individual input. I – normal, II – mild Stargardt disease, III – severe Stargardt disease, BM – bruchs membrane, ONL – outer nuclear layer, OS – outer segment layer (ellipsoid zone), RPE – retinal pigment epithelium.

### *Figure 3*

Figure describing the architecture of the convolutional neural network used for Model 2. Each coloured block represents a different layer of the convolutional neural network. INPUT – input image consisting of 3 channels (red, green and blue), CONV – convolutional layer, P – probability, RELU – rectified linear unit. Trinary classification: Class I – normal, Class II – mild Stargardt disease, Class III – severe Stargardt disease.

### *Figure 4*

470 Evolution of accuracy (the proportion of correct predictions) on validation data during  
471 training for (A) Model 1 and (B) Model 2.

472

473 *Figure 5*

474 Receiver Operator Characteristic (ROC) curves for Model 2. (A) Two levels of binary  
475 classification to distinguish between normal and Stargardt disease. Classification at the level  
476 of the column was performed by considering each 256 x 32 pixel column input individually.  
477 Classification at the level of the OCT scan was performed by combining the probabilities of  
478 all the column inputs within a single OCT scan after weighted voting to find the class with  
479 the highest probability. (B) Trinary classification of disease severity to distinguish between  
480 normal, mild and severe Stargardt disease at the level of the OCT scan after weighted voting  
481 and profile analysis. AUC – area under the ROC curve.