tvst

Special Issue

Application of a Deep Machine Learning Model for Automatic Measurement of EZ Width in SD-OCT Images of RP

Yi-Zhong Wang[1,2], **Daniel Galles**[1], **Martin Klein**[1], **Kirsten G. Locke**[1], **and David G. Birch**[1,2]

[1] Retina Foundation of the Southwest, Dallas, TX, USA
[2] Department of Ophthalmology, University of Texas Southwestern Medical Center at Dallas, Dallas, TX, USA

Correspondence: Yi-Zhong Wang, Retina Foundation of the Southwest, 9600 N. Central Expressway, Suite 200, Dallas, TX 75231, USA. e-mail: yiwang@retinafoundation.org

Received: August 6, 2019
Accepted: November 21, 2019
Published: March 17, 2020

Keywords: deep machine learning; retinitis pigmentosa; automatic ellipsoid zone segmentation

Citation: Wang Y-Z, Galles D, Klein M, Locke KG, Birch DG. Application of a deep machine learning model for automatic measurement of EZ width in SD-OCT images of RP. Trans Vis Sci Tech. 2020;9(2):15, https://doi.org/10.1167/tvst.9.2.15

Purpose: We applied a deep convolutional neural network model for automatic identification of ellipsoid zone (EZ) in spectral domain optical coherence tomography B-scans of retinitis pigmentosa (RP).

Methods: Midline B-scans having visible EZ from 220 patients with RP and 20 normal subjects were manually segmented for inner limiting membrane, inner nuclear layer, EZ, retinal pigment epithelium, and Bruch's membrane. A total of 2.87 million labeled image patches (33 × 33 pixels) extracted from 480 B-scans were used for training a convolutional neural network model implemented in MATLAB. B-scans from a separate group of 80 patients with RP were used for testing the model. A local connected area searching algorithm was developed to process the model output for reconstructing layer boundaries. Correlation and Bland-Altman analyses were conducted to compare EZ width measured by the model to those by manual segmentation.

Results: The accuracy of the trained model to identify inner limiting membrane, inner nuclear layer, EZ, retinal pigment epithelium, and Bruch's membrane patches in the test dataset was 98%, 89%, 91%, 94%, and 96%, respectively. The EZ width measured by the model was highly correlated with that by two graders ($r = 0.97$; $P < 0.0001$). Bland-Altman analysis revealed a mean EZ width difference of 0.30 mm (coefficient of repeatability = 0.9 mm) between the model and the graders, comparable to the mean difference of 0.34mm (coefficient of repeatability = 0.8 mm) between two graders.

Conclusions: The results demonstrated the capability of a deep machine learning-based method for automatic identification of EZ in RP, suggesting that the method can be used to quantify structural deficits in RP for detecting disease progression and for evaluating treatment effect.

Translational Relevance: A deep machine learning model has the potential to replace humans for grading spectral domain optical coherence tomography images in RP.

Introduction

Retinitis pigmentosa (RP) is a group of genetic eye disorders causing visual impairment. Patients with RP experience gradual decline in their vision and may lose all useful sight owing to retinal degeneration. With potential new and emerging treatments on the horizon for inherited retinal degenerations, especially RP,[1] it is essential to have efficient and sensitive biomarkers for detecting disease progression and for evaluating treatment effects. For patients with RP, visual acuity loss is relatively slow until at late stages of the disease, so visual field and/or full-field electroretinograms are often used to quantify visual function defects and to monitor disease progression. Advanced retinal imaging techniques, such as spectral domain optical coherence tomography (SD-OCT) have more recently become available for revealing and visualizing structural changes in the retina at various stages of disease progression.

A number of studies using SD-OCT demonstrated that the structural defects in RP mainly occur in the outer retina as the disease progresses, including a decrease in thickness of outer nuclear layer, total photoreceptor thickness, and/or photoreceptor outer segment (OS) thickness.[2–4] The visual field sensitivity loss in a transition zone between relatively healthy and relatively affected outer retinal areas is more rapid than it is elsewhere in the retina.[5–8] In this transition zone, OS thickness changes from visible to nonmeasurable. The measurement of the more healthy retina can be the width or area of the remaining ellipsoid zone (EZ) or EZ area. Hence, EZ metrics obtained from OCT scans could be potential biomarkers for detecting disease progression and as outcome measures in prospective clinical trials for RP. However, one of the main limitations is that conventional graph search–based automated OCT image layer segmentation algorithms require prior definitions of retinal structure and often incorrectly identify the EZ in the transition zone or in the region where EZ is missing, thus requiring time-consuming manual correction for accurate layer segmentation.

Recent advances in deep machine learning and convolutional neural networks (CNN)[9] have shown promising applications in ophthalmology, especially in fundus photo and OCT image processing.[10–15] A deep CNN model can learn how to identify features in images through training with a classified dataset. For instance, deep neural networks have been trained for automatic identification of diabetic retinopathy in retinal fundus photographs,[11,12,15] for automatic identification of retinal layer boundaries in OCT images of dry age-related macular degeneration,[16] and for quantification of EZ defects on OCT images of macular telangiectasia type 2.[17]

The purpose of this study was to train and test a CNN model to automatically delineate outer retinal layers in SD-OCT B-scan images obtained from patients with RP, and to evaluate the capability of the deep machine learning-based method for automatic measurements of EZ width and photoreceptor OS length in RP by comparing it with manual segmentation method (gold standard) as well as with automatic segmentation by Heidelberg Spectralis (Heidelberg Engineering, Heidelberg, Germany).

Methods

OCT Scan Images for CNN Model Training and Testing

Nine-millimeter (30°) SD-OCT high-speed (768 A-scans) and high-resolution (1536 A-scans) B-scans with an automatic real-time tracking setting of 100 were obtained using a Heidelberg Spectralis (HRA-OCT, Heidelberg Engineering). B-scan images from patients with RP over the past 10 years at the Retina Foundation of the Southwest were reviewed. From 400 patients with SD-OCT scans, 220 patients with RP were identified with EZ transition zones visible in their midline B-scan images and used to generate image datasets for training and validation of a CNN model. The other 180 patients were excluded owing to no identifiable EZ transition zone in the B-scan images (either no visible EZ band or EZ band extended beyond the scan areas). Among these 220 patients, 50 were autosomal-dominant RP (adRP), 30 autosomal-recessive RP, 20 X-linked RP (xlRP), and 120 isolated RP. In addition, midline B-scan images from 20 normal subjects were also included for CNN model training and validation. All 480 line B-scans from two eyes of 240 subjects were first automatically segmented then manually corrected by one grader using Spectralis software (ver. 1.9.10) for the following five layer boundaries: inner limiting membrane (ILM), distal (basal) INL (dINL), center of the EZ, proximal (apical) retinal pigment epithelium (pRPE), and Bruch's membrane (BM). For CNN model testing, we identified two separate groups of patients with RP who had multiple visits with SD-OCT scans and had measurable EZ in the central retina at their first visit. Group 1 included 36 patients with adRP. Group 2 included 44 patients with xlRP. The outputs of the model were compared with the gold-standard of manual segmentation for ILM, dINL, EZ, pRPE, and BM by two graders.

CNN Model Architecture

In this study, we adopted a well-established CNN framework developed for classifying tiny images.[16,18] This CNN model has shown promising results for automatic segmentation of retinal layer boundaries in OCT images of patients with dry age-related macular degeneration.[16] The CNN model was implemented in MATLAB using MatConvNet.[19] MatConvNet is a MATLAB toolbox implementing CNNs especially for image classification applications. Table 1 summarizes the architecture and the parameters of the model, which were the same as those used by Fang et al.,[16]

Table 1. Architecture of the CNN model and its parameters

	Type	Filter Size	Stride	Filter Number	Padding
Layer 1	Convolution	5 × 5 × 1	1 × 1	32	2
Layer 2	Max pooling	3 × 3	2 × 2	—	0
Layer 3	ReLU	—	—	—	—
Layer 4	Convolution	5 × 5 × 32	1 × 1	32	2
Layer 5	ReLU	—	—	—	—
Layer 6	Average pooling	3 × 3	2 × 2	—	0
Layer 7	Convolution	5 × 5 × 32	1 × 1	64	2
Layer 8	ReLU	—	—	—	—
Layer 9	Average pooling	3 × 3	2 × 2	—	0
Layer 10	Fully connected	4 × 4 × 64	—	64	0
Layer 11	ReLU	—	—	—	—
Layer 12	Fully connected	1 × 1 × 64	—	6	0
Layer 13	Softmax	—	—	—	—

ReLU, Rectified linear unit.

except for Layer 12 where the number of filters was 6 for six classes in our study, including ILM, dINL, EZ, pRPE, BM, and background.

As shown in Table 1, the CNN model has a total of 13 layers, including convolutional layers, pooling layers, rectified linear unit layers, fully connected layers, and a final softmax classification layer. A convolutional layer convolves the input with different spatial filters (kernels or receptive fields) to extract various features in the input. For instance, the first convolutional layer in the model has 32 filters of size 5 × 5 × 1. Each filter extracts a different feature in the input. The deeper a convolutional layer is, the higher level features it extracts. A pooling layer is to reduce the dimensions of the feature maps to ease computational burdens. The rectified linear unit layer performs a simple nonlinear transformation to accelerate the CNN training process. The final softmax layer outputs six numbers between 0 and 1, giving the probability of the input image in each of six classes.

Create Labeled Image Datasets for CNN Training and Validation

The training data for the CNN model were tiny classified image patches of 33 × 33 pixels extracted from B-scan images. The classification or labeling of each patch was determined by the class of its center pixel. The CNN model trained with such tiny image patches served as a pixel classifier[20] to determine the probabilities of each pixel in a B-scan image that belongs to one of six classes, five retinal layer boundaries, and background, based on its surrounding features.

Layer Boundary Classifications of B-Scan Images

For a CNN model to determine if a pixel in a B-scan image falls on a retinal layer boundary, the model needs to be well-trained with a large image dataset classified according to manually graded retinal layer boundaries. Figure 1 illustrates the classification of layer boundaries in a B-scan image. The Spectralis software (ver. 1.9.10) was used to automatically segment and manually correct the ILM, dINL, EZ, pRPE, and BM in each B-scan image (Fig. 1a). Manually corrected OCT scans were exported as XML files, which were then imported into MATLAB to extract B-scan images and corresponding layer segmentation data. The pixels in a B-scan image on ILM, dINL, EZ, pRPE, or BM boundaries were labelled as 1, 2, 3, 4, or 5, respectively. Any pixels in a B-scan image not on these five boundaries was labeled as 0. Figure 1c shows an example of classifications of all pixels in a B-scan image.

Preprocessing of B-Scan Images

Because OCT B-scan images vary in intensity, images were preprocessed for intensity normalization and for reducing the effect of hyperintense reflections. The method of preprocessing followed that previously reported.[16,21] Specifically, the intensity values of an original B-scan image, I_{origin}, were first linearly rescaled to the range of [0, 1], resulting in $I_{rescaled}$. Then a median filter with a kernel size of 20 × 2 was applied to the rescaled image $I_{rescaled}$. The maximum pixel intensity value of the filtered image $I_{filtered}$ was used to set a threshold. To reduce hyperintense reflections, any pixels in the rescaled image $I_{rescaled}$ that were above the threshold were set to the value of the threshold, resulting in $I_{threshold}$. Finally, the intensity values of

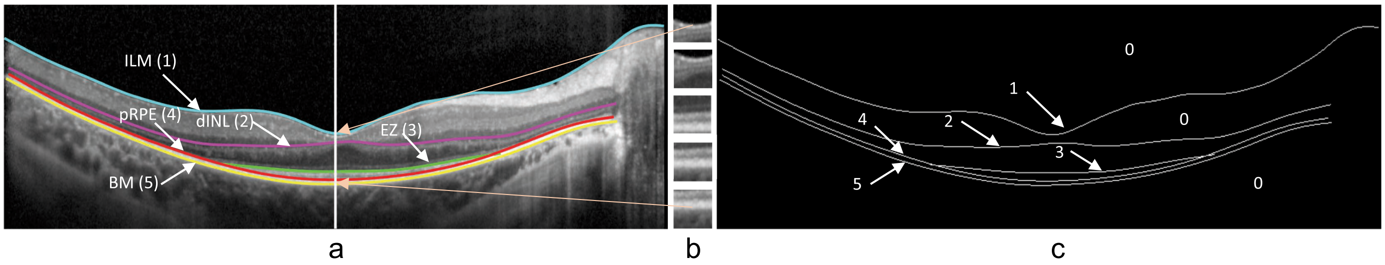


Figure 1. Classification of layer boundaries in a B-scan image. (a) B-scan image with five manually corrected layer boundaries, ILM, dINL, EZ, pRPE, and BM labeled as 1, 2, 3, 4, and 5, respectively. (b) Examples of 5 image patches, each centered on one of the five layer boundaries, extracted from an A-scan indicated by the vertical line in (a). (c) Classification of every pixel in the B-scan image.

all pixels in $I_{threshold}$ were normalized by dividing by the maximum value of $I_{threshold}$, generating the final normalized intensity image, $I_{normalized}$.

Extracting Classified Image Patches from B-Scan Images for CNN Training

After classification and preprocessing, B-scan images were ready for the extraction of image patches for training and validation. Similar to the method used by Fang et al.,[16] the training dataset included both positive and negative image patches of 33 × 33 pixels. A positive patch was centered at a pixel on a layer boundary and the patch's label was the same as the center pixel (label of the boundary, 1–5). A negative patch was centered at a pixel labeled as 0. For each A-scan containing EZ, five positive patches, each centered at a pixel on one of five layer boundaries, and one randomly selected negative patch were extracted. Figure 1b showed examples of five positive training image patches obtained from an A-scan indicated by the vertical line in Figure 1a. For A-scans not containing EZ, an additional negative patch was extracted. If 6 patches (5 positive and 1 negative) were generated for each A-scan, then a B-scan with 768 A-scans could produce 4608 image patches if all A-scans contain the layer boundaries included in the model (ILM, dINL, EZ, pRPE, and BM). In this way, a total of 2.87 million classified patches were extracted from 480 line B-scan images of 240 subjects for training (80%) and validation (20%).

CNN Model Training and Testing

CNN Model Training and Validation

All classified image patches were randomly grouped into two datasets, one for training (80% of the 2.87 million patches) and the other for validation (20%). Before the training started, all filter weights were set to random numbers. The training and validation datasets were divided into batches. The default batch size was 100 image patches of size 33 × 33 pixels. After the CNN was trained for each batch, the error between the patch classifications generated by the CNN and the predefined patch labels (the ground truth) was calculated. The filter weights were updated using a stochastic gradient descent optimization algorithm and backpropagation (backward propagation of errors)[22,23] to minimize the error. The training stopped after the model was trained for 45 epochs (the full dataset was used for training 45 times). The default values of weight decay (1.0e-04) and learning rates (0.05 from 1 to 30 epochs; 0.005 from 31 to 40 epochs; and 0.0005 from 41 to 45 epochs) of the CNN model were also adopted.

CNN Model Testing

The trained CNN model was tested using a separate dataset (OCT line B-scan images from 36 patients with adRP and 44 with xlRP who had multiple visits over time with OCT scans and visual function results, so that their data can be used for testing in future studies). For each pixel in the test image that was on one of the five manually corrected layer boundaries, a patch of size 33 × 33 centered at that pixel was extracted and classified by the CNN. A total of 600,000 patches were generated from test B-scan images. The CNN classifications for test patches were compared with corresponding manually defined classes. The accuracy (or error rate) of the model to classify layer boundary patches extracted from the test dataset were calculated.

Local Connected-Area Searching (LCASA) Algorithm for Postprocessing of Classification Maps

The trained CNN is effectively a pixel classifier. Applying the trained CNN model to each pixel in a B-scan image creates classification probabilities for all pixels. Figures 3b and 3f show two examples of classification maps based on maximum probability of classes generated for B-scan images of two patients with RP. It is evident that a band of pixels could be classified as the same boundary class and there are also false positives. Hence, postprocessing of classification maps was required to reconstruct a single line for each layer boundary. For this purpose, we developed a LCASA algorithm to process classification maps to localize layer boundaries.

In the LCASA algorithm, we assumed that the largest local connected area for a class in the classification map truly belonged to that class. This assumption was based on the high accuracy (90% or higher) of the model to correctly identify the class of a pixel; thus, the largest area of connected pixels on a classification map necessarily represents the true class. With the largest connected area as a starting reference, the LCASA algorithm first eliminated smaller local connected areas in the same class that were separated vertically from the larger ones but fully overlapped horizontally with larger ones (i.e., either above or below larger ones). For smaller areas having partial overlapping horizontally with or completely separated from the larger ones, nearest neighbor distance rules were applied to determine if the smaller areas belong to the same class of the larger ones. Based on the model's accuracy to identify these classes (see Results), the order to reconstruct layer boundaries by the LCASA algorithm was from ILM, BM, pRPE, EZ, and finally to dINL. Once the search for a layer boundary class was completed, that boundary was added to the list of reference boundaries for the search of remaining boundaries. A single-pixel layer boundary was obtained by averaging vertical locations of the same class pixels.

CNN Model Evaluation

The evaluation of the effectiveness of the deep machine learning-based method (the CNN model + LCASA postprocessing algorithm) for automatic segmentation of outer retinal layer boundaries, especially the EZ band, was conducted using the test B-scan images from a separate group of 80 patients with RP (36 patients with adRP and 44 with xlRP). EZ width, EZ–pRPE thickness, dINL–pRPE (photoreceptor+) thickness, and ILM-BM (total retinal) thickness generated by the deep machine learning-based method were compared with those obtained from manual segmentation by human graders (gold standard), as well as compared with the results of automatic segmentation by Spectralis. Correlation and Bland-Altman analyses were performed.

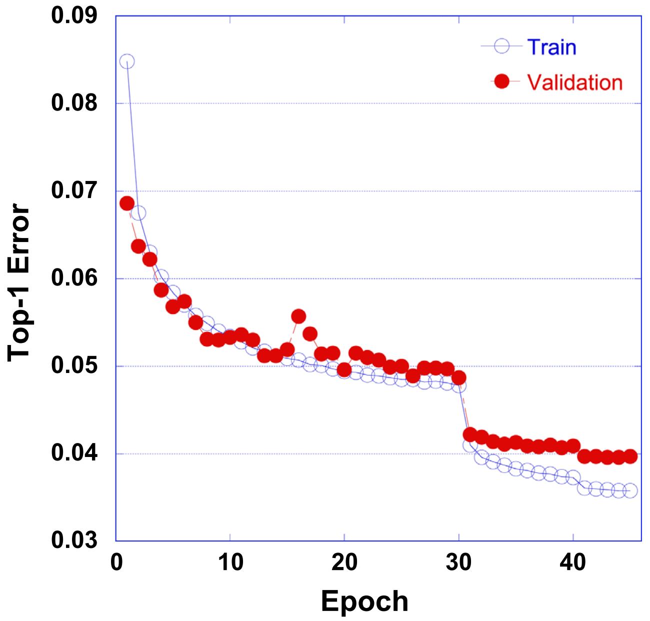


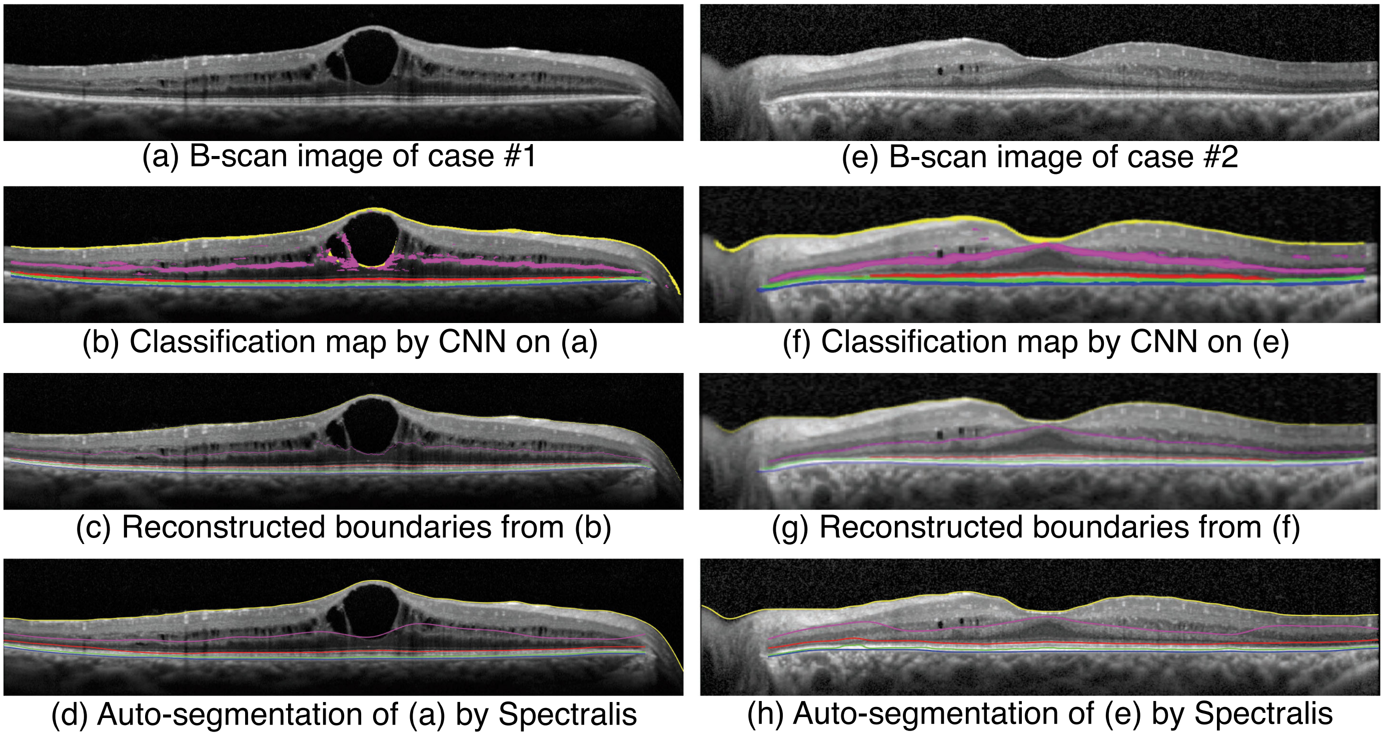
Figure 2. Top-1 error as a function of training epochs. The reduction of top-1 error at the 31st epoch was related to the change of the learning rate from 0.05 to 0.005. Training stopped after 45 epochs.

Results

Accuracy of the Trained CNN Model to Classify Image Patches on Layer Boundaries

Figure 2 plots the top-1 error (the rate that the class having the highest probability determined by the model is different from the target class defined by manual segmentation) as a function of training epochs. After the completion of the training at 45 epochs, the overall accuracy (1 minus top-1 error) of the CNN model to correctly identify the classes of image patches in the validation set was 96%.

To access the accuracy of the trained CNN model to classify pixels on individual layer boundary, the model was used to classify all pixels of 160 test B-scan images from a separate group of 80 patients with RP to obtain the classification maps for all classes (ILM, dINL, EZ, pRPE, BM and background). Figures 3a and 3e show two examples of B-scan images, and Figures 3b and 3f show the B-scan images overlapping with their corresponding classification maps based on maximum probabilities of five layer boundaries generated by the model. The model-determined classes of the pixels were compared with those of manual segmentation of line boundaries at the same locations, and the accuracies of the model to correctly identify ILM, dINL, EZ, pRPE, and BM patches extracted from the test B-scan images were 98%, 89%, 91%, 94%, and 96%, respectively, as shown in Figure 4 at 100% of 240 patients.


(a) B-scan image of case #1

(e) B-scan image of case #2

(b) Classification map by CNN on (a)

(f) Classification map by CNN on (e)

(c) Reconstructed boundaries from (b)

(g) Reconstructed boundaries from (f)

(d) Auto-segmentation of (a) by Spectralis

(h) Auto-segmentation of (e) by Spectralis

Figure 3. Examples of midline B-scans (a) and (e) from two patients with RP. (b) and (f) B-scan images with classification maps of five layer boundaries based on maximum probabilities (*yellow*, ILM; *magenta*, dINL; *red*, EZ; *green*, pRPE; *blue*, BM) from the output of the CNN model. (c) and (g) B-scan images with reconstructed single-pixel layer boundaries after applying the LCASA algorithm to postprocess the classification maps. (d) and (h) B-scan images with the results of automatic segmentation by Spectralis.

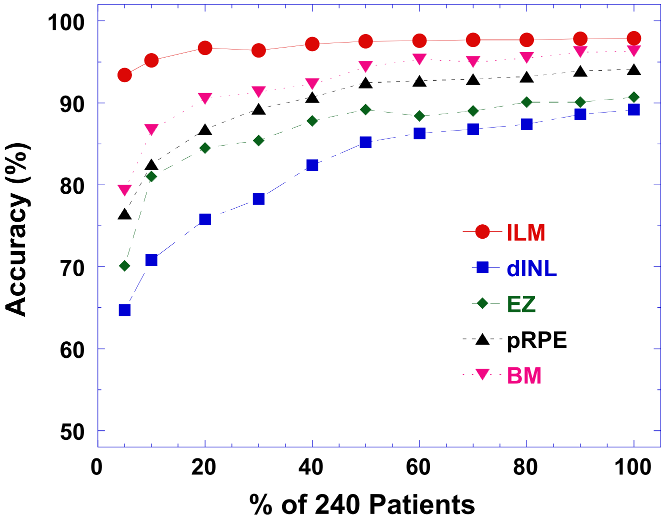


Figure 4. The accuracy of the model to identify patches on each of five layer boundaries in the test B-scan images as a function of percent number of patients in the training set whose B-scan images were used to train the CNN model.

No background (negative) patches were included in the testing image patch dataset for accuracy evaluation.

To assess the impact of number of OCT scan images used for training on the performance of the CNN model, the same CNN model was trained with image patches extracted from randomly selected subsets of 240 subjects in our study (220 patients with RP and 20 normal subjects). Figure 4 shows the accuracy of the model to identify ILM, dINL, EZ, pRPE, and BM patches in the same 160 test B-scan images as a function of percent number of patients in the training set whose B-scan images were used to train the CNN model. It is evident that, for layer boundaries with less variability or more consistent surrounding features, such as the ILM and BM, a smaller number of patients or OCT images may be needed to train the CNN model to achieve a predetermined accuracy (e.g., 95%), whereas more patients or OCT images are required to train the model to achieve the same accuracy for layer boundaries with more variabilities or less consistent surrounding features (such as the EZ and dINL).

Postprocessing of classification maps was then conducted by using the LCASA algorithm to reconstruct single-pixel layer boundaries. Figures 3c and 3g show the results of automatic segmentation of ILM, dINL, EZ, pRPE, and BM by the CNN model after the LCASA algorithm was applied to the classification maps in Figures 3b and 3f. The reconstructed single-pixel layer boundaries were used for EZ width and retinal thickness measurements.

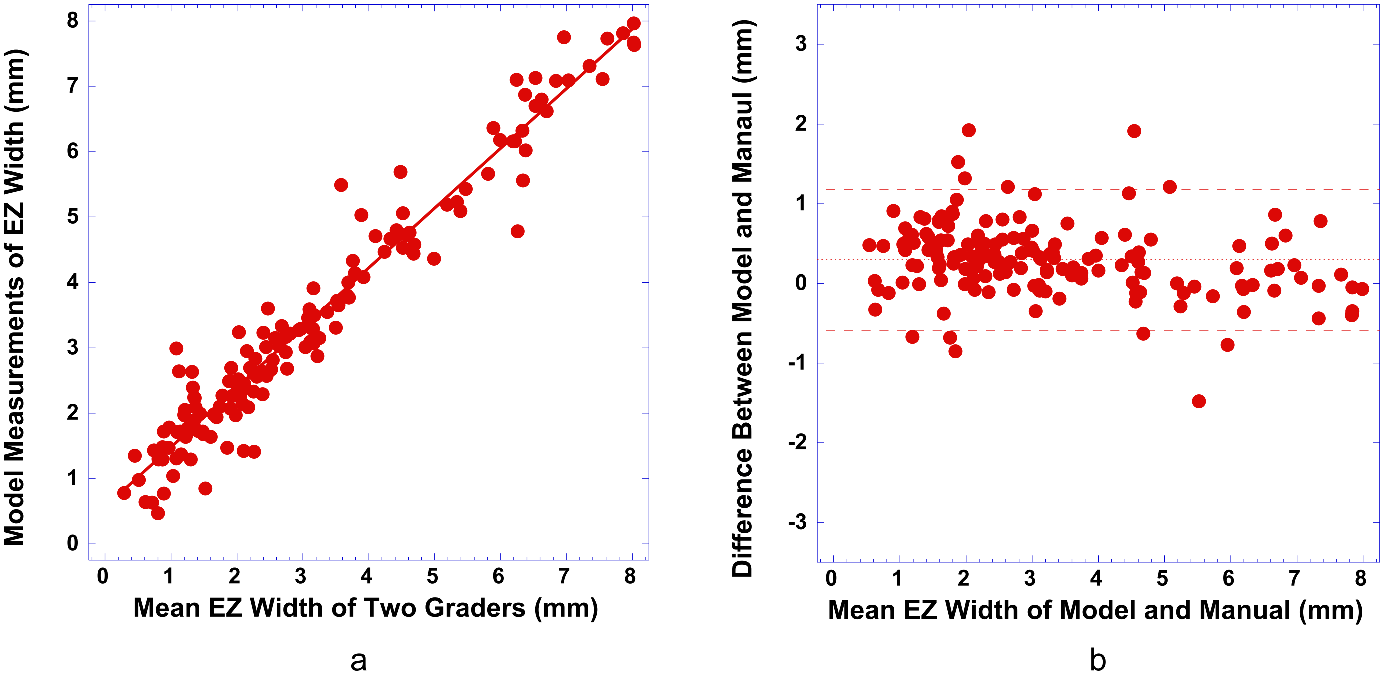


Figure 5. (a) EZ width measured by the deep machine learning-based method is plotted against the average EZ width measured by manual segmentation of two graders within central 8 mm of the test B-scan images ($n = 160$). (b) Bland-Altman plot of the difference between two measurements in (a) versus their mean. *Dotted line* shows the mean difference. *Dashed lines* show ±95% limits of differences.

In comparison, Figures 3d and 3h show the results from automatic segmentation with the Spectralis software.

EZ Width Measurement

EZ width in millimeters was obtained by first counting number of pixels that represented EZ, then the width in pixels was converted to mm using the scanning scale (mm/pixel) along the B-scan axis. The mean ± SD of EZ widths of 80 patients in the testing dataset obtained by grader 1, grader 2, the CNN model, and Spectralis were 3.26 ± 2.00 mm, 2.92 ± 2.00 mm, 3.38 ± 1.87 mm, and 7.51 ± 0.71 mm, respectively. Figure 5a plots the EZ width measured by the deep machine learning-based method (the CNN model + the LCASA postprocessing algorithm) versus the average EZ width measured by manual segmentation of two graders within central 8 mm of B-scans. The correlation between the model-measured EZ width and that by two graders was 0.97 ($P < 0.0001$). Bland-Altman analysis (Fig. 5b) revealed a mean difference of 0.30 mm with a coefficient of repeatability of 0.9 mm between the model and the average measurements of EZ width by two graders, which was comparable with the mean difference of 0.34 mm (coefficient of repeatability = 0.8) between the two graders.

In comparison, there was no significant correlation between EZ width measured with automatic segmentation by Spectralis and that by manual segmentation (r = 0.147; $P = 0.0644$). Bland-Altman analysis showed a mean difference of 4.42 mm with coefficient of repeatability of 3.9 mm between EZ width measurements by these two segmentation methods.

Retinal Layer Thickness Measurements

Retinal thickness was measured by first counting number of pixels between two layer boundaries, then the thickness in pixels was converted to millimeters using the scanning scale (mm/pixel) along the A-scan axis. Table 2 summarizes the results of comparing retinal thickness measurements obtained by manual grading with those by the CNN model and Spectralis automatic segmentation software for central line B-scan widths of 1, 3, 6, and 8 mm. Figure 6 plots the retinal layer thicknesses measured by the CNN model vs those by the average of two graders (top row of Fig. 6) and those measured by the Spectralis software versus two graders (bottom row of Fig. 6) over the central 8 mm of B-scans.

It is evident that the agreement between the deep machine learning-based method and the average of two graders was comparable to that between two graders. In addition, when compared with manual grading (gold standard), the trained CNN model outperformed the Spectralis' automatic segmentation software for measuring the thickness of the photoreceptor+ and the

Table 2. Comparison of retinal layer thicknesses measured by three different methods: the deep machine learning-based method (CNN), automatic segmentation by Spectralis software, and manual segmentation (gold standard)

	CNN vs Average of 2 Graders			Spectralis vs Average of 2 Graders			Grader 1 vs Grader 2			Mean ± SD			
Retinal Layer Thickness (μm)	R^2	Mean Difference	STD Difference	R^2	Mean Difference	STD Difference	R^2	Mean Difference	STD Difference	Grader 1	Grader 2	CNN	Spectralis
Central 1 mm													
EZ-pRPE	0.842	0.9	3.5	0.146	10.5	8.6	0.815	4.8	4.0	33.1 ± 9.3	28.5 ± 8.9	31.6 ± 8.1	41.2 ± 5.9
dINL-pRPE	0.893	2.1	11.7	0.559	13.5	24.9	0.916	1.5	10.5	161.9 ± 36.2	160.5 ± 34.1	159.2 ± 35.6	174.7 ± 35.4
ILM-BM	0.998	0.4	2.6	0.996	2.8	4.2	0.998	1.0	2.9	269.2 ± 61.7	268.3 ± 61.6	268.4 ± 61.5	272.0 ± 61.2
Central 3 mm													
EZ-pRPE	0.829	1.0	2.6	0.015	14.6	7.6	0.750	3.2	3.3	26.8 ± 6.6	23.8 ± 6.3	26.3 ± 6.0	39.9 ± 5.1
dINL-pRPE	0.962	1.7	6.1	0.179	30.7	30.8	0.948	2.1	7.1	130.2 ± 31.2	128.2 ± 30.1	130.9 ± 31	159.9 ± 26.6
ILM-BM	0.998	0.0	1.9	0.978	4.3	6.9	0.998	1.4	2.2	290.8 ± 45.5	289.4 ± 45.7	290.1 ± 45.2	294.6 ± 44.3
Central 6 mm													
EZ-pRPE	0.773	1.1	2.5	0.002	16.5	6.7	0.694	2.5	3.1	24.9 ± 5.5	22.5 ± 5.3	24.8 ± 4.7	40.2 ± 4.5
dINL-pRPE	0.962	2.9	5.1	0.033	51.8	29.3	0.926	2.2	7.2	97.4 ± 26.6	95.2 ± 26.0	99.2 ± 25.8	148.1 ± 18.9
Total retina	0.998	0.3	1.8	0.909	6.6	10.8	0.995	1.6	2.4	272.2 ± 35.5	270.6 ± 35.7	271.7 ± 35.2	278.2 ± 34.7
Central 8 mm													
EZ-pRPE	0.750	1.2	2.5	0.000	17.2	6.8	0.683	2.4	3.0	24.6 ± 5.3	22.3 ± 5.1	24.6 ± 4.4	40.7 ± 4.6
dINL-pRPE	0.942	2.7	5.6	0.012	58.6	27.3	0.917	2.0	6.9	87.3 ± 23.8	85.3 ± 23.2	89.0 ± 22.4	144.9 ± 17.2
ILM-BM	0.992	0.7	2.9	0.794	11.2	14.9	0.991	1.3	3.0	263.1 ± 32.1	261.7 ± 32.3	263.1 ± 31.9	273.7 ± 31.3

STD, standard deviation.

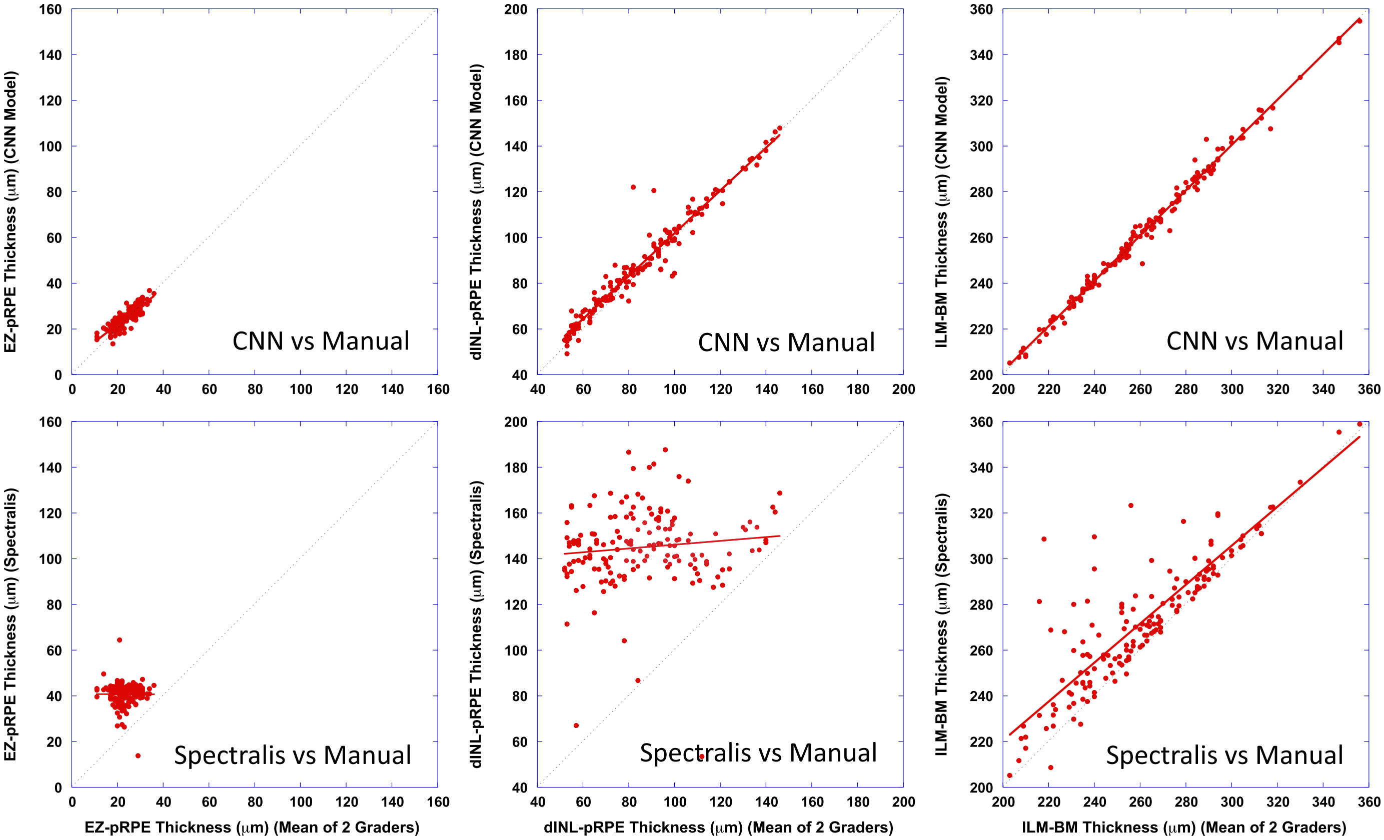


Figure 6. *Top row*: Retinal layer thicknesses measured by the deep machine learning-based method versus those by the average of two graders. *Bottom row*: Retinal layer thickness measured by the Spectralis automatic segmentation software vs those by the average of two graders. *From left to right*: Photoreceptor OS (EZ-pRPE) length; photoreceptor+ (dINL-pRPE) thickness; total retinal (ILM-BM) thickness. The width of line B-scan examined was central 8 mm.

length of photoreceptor OS (EZ-pRPE thickness). The Spectralis software performed comparably to the CNN model only for the total retinal thickness measurement within the central 3 mm of the retina (Table 2).

Discussion

The results of this study demonstrated the capability of a deep CNN model-based method for automatic segmentation of outer retinal layers in SD-OCT scan images obtained from patients with RP. The CNN model performed similarly to human graders when measuring EZ width and retinal thickness, suggesting that well-trained CNN models may be used to quantify structural deficits for detecting disease progression and for evaluating treatment effects in future clinical trials for RP.

The training of a deep CNN typically requires a large dataset. One of the questions is how many patients or OCT scan images do we need to successfully train the CNN model used in our study? To understand the impact of number of patients on the performance of the CNN model, we trained the same CNN model with data from randomly selected subsets of patients in our study. Our results (Fig. 4) suggested that a smaller number of patients or OCT images may be needed for layer boundaries with more consistent surrounding features, such as ILM and BM, whereas more patients or OCT images may be required to train the model to achieve the same accuracy for layer boundaries with less consistent surrounding features (such as the EZ and dINL). With equal number of training patches for every layer boundary, the model will have lower accuracies for classifying dINL and EZ than ILM and BM.

Lower accuracy to identify the class of image patches for a layer boundary may lead to less accurate segmentation of that boundary, and hence lead to less accurate measurements of layer thickness by the model, which may explain, at least in part, the lower correlation between the CNN model and manual graders for EZ-pRPE and photoreceptor+ thickness measurements when compared with that for the total retinal thickness (Table 2). Another possible explanation for lower correlation, especially for EZ-pRPE, is that the number of pixels representing EZ-pRPE thickness is too small owing to the A-scan resolution limit relative to the layer thickness so that one pixel difference in thickness results in a larger percent change (greater variability) of thickness.

To improve the model's accuracy to identify the pixels on EZ and dINL, more training image patches extracted from EZ and dINL boundaries from additional patients with RP can be added. One of the limitations of this study was that the current model was trained with image patches extracted from line B-scan images. Much more training data will be available with volume scans. The performance of the line B-scan CNN model on volume scan images can be examined to help determine the types of data patches needed to train and test the CNN model for automatic segmentation of outer retinal layers in volume scans obtained from patients with RP. To further increase the size of training datasets, multiple volume scans from the same patients can be included. Repeated scans introduce variations in image representation, such as rotations, image intensity, and/or quality. Therefore, using multiple scans from the same patients is equivalent to data augmentation methods[9] often used in deep machine learning to increase the size of datasets, and will help CNN training and testing.

In this study, the CNN model we adopted is similar to the one used by Fang et al.[16] for automatic segmentation of retinal layer boundaries in OCT images of dry age-related macular degeneration. Default parameters of the model were used. It has been pointed out that that some parameters of this CNN model such as patch and filter size are empirically selected and may not be optimal.[16] The effects of different parameters on the model performance remain to be evaluated to determine the optimal parameters to further improve the accuracy of the model to identify image patches of layer boundaries.

Often, CNNs are trained and tested with relatively high-quality image date sets. However, in real-life applications such as in this project with OCT scans, image quality varies. For instance, in RP, cystoid macular edema (CME) might impact the EZ signal in OCT scans. It has been shown that CNNs are susceptible to image blur and noise.[24] In this study, both high-resolution and lower quality high-speed images were used to train the CNN. As shown in Figure 3, the trained CNN model can perform well to identify the CME boundary. If the training image dataset includes OCT images with weak boundary signals that can be classified by experience graders, the trained CNN model should be able to detect weak layer boundary signals in the test images to identify EZ zone in the presence of CME, even if EZ signals may be weak.

Lower accuracy of the CNN model to identify the class of image patches for a layer boundary may also lead to the increase of false positives for the pixels not on the layer boundary, as is the case with the dINL illustrated in Figures 3b and 3f. The postprocessing of classification maps can help eliminate false positives of the model classification, as shown in Figures 3c and 3g, hence improve the model's performance. Different from the graph-search method used by Fang et al.[16,25] to find layer boundaries from classification probability maps, we developed a LCASA algorithm for the postprocessing of classification maps. Judging by the results of this study, the LCASA algorithm is promising for automatic segmentation of retinal layer boundaries in SD-OCT scan images. The current LCASA algorithm only deals with maximum class probability for each pixel. As shown in Figure 3b, some pixels near the CME were classified as ILM (yellow pixels) based on maximum probability. Although these false-positive pixels were removed by the LCASA algorithm, the next-class probability of these removed pixels could be their true class (i.e., dINL). Hence, further refinement of the LCASA algorithm is needed to improve its performance, especially for the cases where local connected areas are eliminated and the class of next-level probability in these local areas can be added for additional search. The method to combine the LCASA algorithm with the graph search algorithm can also be explored to determine if a more effective postprocessing algorithm can be developed.

For supervised CNN learning, as is the case in this study, accurate classification of the training dataset is crucial. In this study, OCT scan images manually segmented by one experienced grader (grader 1) were used as the gold standard to classify training image patches. Because there is intergrader variability in manual segmentation, OCT images segmented by additional graders may be needed to train the CNN. To further evaluate the performance of the CNN model trained with the data from one grader's classification, we examined the automatic EZ width measurements by the CNN model with those by two individual graders (grader 1 and grader 2) for the test B-scan images. As shown in Table 3, although the CNN model had closer agreement of the mean difference with grader 1, the repeatability coefficient (standard deviation) was comparable.

Furthermore, it is interesting to note that the repeatability coefficient between the CNN model and the average of two graders was closer to that between grader 1 and grader 2 than to those between the

Table 3. Comparison of EZ width measured by the deep machine learning-based method with the manual segmentation of individual graders

	CNN vs Average of 2 Graders			CNN vs Grader 1			CNN vs Grader 2			Grader 2 vs Grader 1		
Central 8 mm EZ Width (mm)	R^2	Mean Difference	STD Difference	R^2	Mean Difference	STD Difference	R^2	Mean Difference	STD Difference	R^2	Mean Difference	STD Difference
EZ width	0.95	0.30	0.46	0.94	0.13	0.49	0.93	0.46	0.52	0.96	−0.34	0.41

STD, standard deviation.

CNN model and individual graders, suggesting that the CNN model may act more like an average grader than individual grader. With SD-OCT image segmentation data from multiple graders, we could potentially establish a well-trained CNN model, together with a postprocessing algorithm, to replace humans for automatic grading of SD-OCT images from patients with RP. Nevertheless, further work is needed to evaluate the sensitivity of the CNN model-based approach to detect disease changes over time with longitudinal data.

Acknowledgments

Supported in part by EY09076 and FFB Individual Investigator Research Award.

Disclosure: **Y.-Z. Wang,** None; **D. Galles,** None; **M. Klein,** None; **K.G. Locke,** None; **D.G. Birch,** None

References

1. Smith J, Ward D, Michaelides M, Moore AT, Simpson S. New and emerging technologies for the treatment of inherited retinal diseases: a horizon scanning review. *Eye (Lond)*. 2015;29:1131–1140.
2. Aleman TS, Cideciyan AV, Sumaroka A, et al. Retinal laminar architecture in human retinitis pigmentosa caused by Rhodopsin gene mutations. *Invest Ophthalmol Vis Sci.* 2008;49:1580–1590.
3. Hood DC, Lin CE, Lazow MA, Locke KG, Zhang X, Birch DG. Thickness of receptor and post-receptor retinal layers in patients with retinitis pigmentosa measured with frequency-domain optical coherence tomography. *Investigative Ophthalmology & Visual Science* 2009;50:2328–2336.
4. Witkin AJ, Ko TH, Fujimoto JG, et al. Ultra-high resolution optical coherence tomography assessment of photoreceptors in retinitis pigmentosa and related diseases. *Am J Ophthalmol.* 2006;142:945–952.
5. Birch DG, Locke KG, Felius J, et al. Rates of decline in regions of the visual field defined by frequency-domain optical coherence tomography in patients with RPGR-mediated X-linked retinitis pigmentosa. *Ophthalmology.* 2015;122:833–839.
6. Birch DG, Locke KG, Wen Y, Locke KI, Hoffman DR, Hood DC. Spectral-domain optical coherence tomography measures of outer segment layer progression in patients with X-linked retinitis pigmentosa. *JAMA Ophthalmol.* 2013;131:1143–1150.
7. Hariri AH, Zhang HY, Ho A, et al. Quantification of ellipsoid zone changes in retinitis pigmentosa using en face spectral domain-optical coherence tomography. *JAMA Ophthalmol.* 2016;134:628–635.
8. Ramachandran R, Zhou L, Locke KG, Birch DG, Hood DC. A comparison of methods for tracking progression in X-linked retinitis pigmentosa using frequency domain OCT. *Transl Vis Sci Technol.* 2013;2:5.
9. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Proceedings of the Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press; 2012:1097–1105.
10. Aslam TM, Zaki HR, Mahmood S, Ahmad NA, Thorell MR, Balaskas K. Use of a neural net to model the impact of optical coherence tomography abnormalities on vision in age-related macular degeneration. *Am J Ophthalmol.* 2017; [Epub ahead of print].
11. Gargeya R, Leng T. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology.* 2017;124:962–969.
12. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA: the Journal of the American Medical Association* 2016;316:2402–2410.
13. Karri SP, Chakraborty D, Chatterjee J. Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration. *Biomed Opt Express.* 2017;8:579–592.

14. Lee CS, Baughman DM, Lee AY. Deep learning is effective for classifying normal versus age-related macular degeneration OCT images. *Ophthalmol Retina.* 2017;1:322–327.
15. Ramachandran N, Hong SC, Sime MJ, Wilson GA. Diabetic retinopathy screening using deep neural network. *Clin Exp Ophthalmol.* 2018;46:412–416.
16. Fang L, Cunefare D, Wang C, Guymer RH, Li S, Farsiu S. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. *Biomed Opt Express.* 2017;8:2732–2744.
17. Loo J, Fang L, Cunefare D, Jaffe GJ, Farsiu S. Deep longitudinal transfer learning-based automatic segmentation of photoreceptor ellipsoid zone defects on optical coherence tomography images of macular telangiectasia type 2. *Biomed Opt Express.* 2018;9:2681–2698.
18. Krizhevsky A, Hinton G. *Learning Multiple Layers of Features from Tiny Images*. *Technical Report*. Toronto, Ontario, Canada: University of Toronto; 2009.
19. Vedaldi A, Lenc K. MatConvNet: Convolutional neural networks for MATLAB. *Proceedings of the ACM International Conference on Multimedia*. Brisbane, Australia: ACM; 2015:689–692.
20. Ciresan DC, Gambardella LM, Giusti A, Schmidhuber J. Deep neural networks segment neuronal membranes in electron microscopy images. *NIPS*. 2012:2852–2860.
21. Lang A, Carass A, Hauser M, et al. Retinal layer segmentation of macular OCT images using boundary classification. *Biomed Opt Express.* 2013;4:1133–1152.
22. Bouvrie J. *Notes on Convolutional Neural Networks*. Cambridge, MA: Massachusetts Institute of Technology; 2006.
23. LeCun Y, Boser B, Denker JS, et al. Backpropagation applied to handwritten zip code recognition. *Neural Computation* 1989;1:541–551.
24. Dodge S, Karam L. Understanding how image quality affects deep neural networks; 2016. arXiv:1604.04004v2 [cs.CV].
25. Chiu SJ, Toth CA, Bowes Rickman C, Izatt JA, Farsiu S. Automatic segmentation of closed-contour features in ophthalmic images using graph theory and dynamic programming. *Biomed Opt Express.* 2012;3:1127–1140.