Using Deep Learning and Transfer Learning to Accurately Diagnose Early-Onset Glaucoma From Macular Optical Coherence Tomography Images

RYO ASAOKA, HIROSHI MURATA, KAZUNORI HIRASAWA, YURI FUJINO, MASATO MATSUURA, ATSUYA MIKI, TAKASHI KANAMOTO, YOKO IKEDA, KAZUHIKO MORI, AIKO IWASE, NOBUYUKI SHOJI, KENJI INOUE, JUNKICHI YAMAGAMI, AND MAKOTO ARAIE

• PURPOSE: We sought to construct and evaluate a deep learning (DL) model to diagnose early glaucoma from spectral-domain optical coherence tomography (OCT) images.

• DESIGN: Artificial intelligence diagnostic tool development, evaluation, and comparison.

• METHODS: This multi-institution study included pretraining data of 4316 OCT images (RS3000) from 1371 eyes with open angle glaucoma (OAG) regardless of the stage of glaucoma and 193 normal eyes. Training data included OCT-1000/2000 images from 94 eyes of 94 patients with early OAG (mean deviation >-5.0 dB) and 84 eyes of 84 normal subjects. Testing data included OCT-1000/2000 from 114 eyes of 114 patients with early OAG (mean deviation >-5.0 dB) and 82 eyes of 82 normal subjects. A DL (convolutional neural network) classifier was trained using a pretraining dataset, followed by a second round of training using an independent training dataset. The DL model input features were the 8×8 grid macular retinal nerve fiber layer thickness and ganglion cell complex layer thickness from spectral-domain OCT. Diagnostic accuracy was investigated in the testing dataset. For comparison, diagnostic accuracy was also evaluated using the random forests and support vector machine models. The primary outcome measure was the area under the receiver operating characteristic curve (AROC).

• RESULTS: The AROC with the DL model was 93.7%. The AROC significantly decreased to between 76.6% and 78.8% without the pretraining process. Significantly smaller AROCs were obtained with random forests and support vector machine models (82.0% and 67.4%, respectively).

• CONCLUSION: A DL model for glaucoma using spectral-domain OCT offers a substantive increase in diagnostic performance. (Am J Ophthalmol 2019;198:136–145. © 2018 Elsevier Inc. All rights reserved.)

GLAUCOMA CAUSES IRREVOCABLE VISUAL FIELD (VF) damage associated with progressive degeneration of retinal ganglion cells (GCs).[1,2] Glaucomatous structural changes can now be evaluated in detail with the advent of spectral-domain optical coherence tomography (OCT). This instrument can measure the thickness of the macular GC complex (GCC) (the macular GC layer and inner plexiform layer combined) and the retinal nerve fiber layer (RNFL). Glaucomatous damage to these layers usually occurs in characteristic patterns.[3,4] Reflecting this, the diagnosis of glaucoma can be improved by analyzing the multitude of structural measurements from spectral-domain OCT in combination, using machine learning methods.[5-8] The development of deep learning (DL) offers the opportunity to further improve the accuracy of glaucoma diagnosis.[9,10] The successes of DL methods have been reported in various research fields; DL models often significantly outperform other machine learning methods.[11,12] Indeed, we recently reported that a DL model significantly outperformed other machine learning methods for detection of preperimetric glaucomatous VF change.[13] Similar improvements in diagnostic performance may be observed by applying DL to spectral-domain OCT measurements from patients with early glaucoma. This is clinically important because glaucoma is one of the leading causes of blindness worldwide,[14] and early detection is essential to maintain visual function.

Accepted for publication Oct 3, 2018.

From the Department of Ophthalmology (R.A., H.M., Y.F., M.M.), The University of Tokyo, Inouye Eye Hospital (K.I.), JR Tokyo General Hospital (J.Y.), and the Kanto Central Hospital of the Mutual Aid Association of Public School Teachers (M.A.), Tokyo, Japan; Moorfields Eye Hospital National Health Service Foundation Trust and University College London (K.H., M.M.), Institute of Ophthalmology, London, United Kingdom; Department of Ophthalmology (K.H., N.S.), School of Medicine, Kitasato University, Kanagawa; Department of Ophthalmology (A.M.), Osaka University Graduate School of Medicine, Osaka; Department of Ophthalmology (T.K.), Hiroshima Memorial Hospital, Hiroshima; Department of Ophthalmology (Y.I.), Kyoto Prefectural University of Medicine, and the Oike Ikeda Eye Clinic (Y.I.), Kyoto; and the Tajimi Iwase Eye Clinic (A.I., K.M.), Tajimi, Japan.

Inquiries to Ryo Asaoka, Department of Ophthalmology, University of Tokyo Graduate School of Medicine, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8655 Japan; e-mail: rasaoka-tky@umin.ac.jp

0002-9394/$36.00
https://doi.org/10.1016/j.ajo.2018.10.007

For DL to be successful, a large training dataset is imperative; however, only a finite amount of OCT images can be prepared in the clinical setting. Several manufacturers have developed OCT machines and there is considerable difference between measured retinal layer thicknesses from different devices.[15,16] In addition, OCT machines are often updated without exchangeability of the data, so OCT data collected using an older version of the machine cannot be combined with data from newer versions. Heterogeneous data such as these can be used to pretrain a DL model ("transfer learning") and obtain an initial representation.[9] DL can then usually can make an accurate diagnosis with only a small amount of additional training data. Transfer learning is a popular approach in DL, where a model developed for a somewhat different task is then reused as the starting point for a model on a second separate task.[17] In other words, the performance of a DL model can be vastly improved by conducting a preliminary training phase using a different large dataset. For instance, several deep convolutional neural networks (GoogLeNet[18] and AlexNet[19]) pretrained using 1.2 million everyday color images from ImageNet (http://www.image-net.org/) outperformed the equivalent deep convolutional neural networks that had not undergone the pretraining process for the purpose of classifying pulmonary tuberculosis.[20] There are numerous other examples, such as pretraining a deep convolutional neural network using the ImageNet for chest pathology identification[21] and pretraining deep convolutional neural networks (CifarNet,[22] GoogLeNet,[18] and AlexNet[19]) using everyday images (Cifar10 dataset: 60,000 everyday color images or ImageNet) for diagnosing pulmonary computerized tomography.[23] We have also recently reported that applying a VGG16 model[24] pretrained using the ImageNet is advantageous to predict visual field sensitivity from OCT-measured retinal thickness.[25]

In the current study, a DL model was initially constructed using a large pretraining dataset obtained with the RS 3000 OCT machine (Nidek Co. Ltd., Aichi, Japan). This initial model was then further trained using a smaller dataset obtained with the Topcon OCT-1000 or OCT-2000 machine (Topcon Corporation, Tokyo, Japan). The diagnostic performance of this model was then tested using an independent dataset consisting of early-onset glaucomatous eyes and normative eyes.

METHODS

THIS STUDY WAS APPROVED BY THE RESEARCH ETHICS Committee of the Graduate School of Medicine and Faculty of Medicine at the University of Tokyo, School of Allied Health Sciences at the Kitasato University, Osaka University, Kyoto Prefectural University of Medicine, Inouye Eye Hospital, and JR Tokyo General Hospital. Written consent was given by patients for their information to be stored in the hospital database and used for research; otherwise, based on the regulations of the Japanese Guidelines for Epidemiologic Study 2008 issued by the Japanese Government, the study protocols did not require that each patient provide written informed consent. Instead the protocol was posted at the outpatient clinic to notify participants of the study. This study was performed according to the tenets of the Declaration of Helsinki (2013).

• **SUBJECTS:** *Pretraining dataset*. The pretraining dataset consisted of 4073 OCT images obtained from 1371 eyes of 747 patients with open angle glaucoma (OAG) and 243 OCT images from 193 healthy eyes of 113 subjects from the Japanese Archives of Multicentral Images of Glaucomatous Optical Coherence Tomography database. These images were obtained at the Department of Ophthalmology, The University of Tokyo Hospital, Inouye Eye Hospital, JR Tokyo General Hospital, Hiroshima Memorial Hospital, Osaka University Hospital, University Hospital of Kyoto Prefectural University of Medicine, and Oike Ikeda Eye Clinic between July 2009 and August 2017. All subjects underwent complete ophthalmic examinations, including biomicroscopy, gonioscopy, intraocular pressure measurement, fundoscopy, refraction, and best corrected visual acuity measurements, as well as spectral-domain OCT imaging and VF testing. A diagnosis of glaucoma was made by an ophthalmologist specializing in glaucoma. Primary open angle glaucoma (POAG) was defined as (1) the presence of typical glaucomatous changes in the optic nerve head, such as a rim notch with a rim width ≤0.1 disc diameters or a vertical cup-to-disc ratio of >0.7 and/or a RNFL defect with its edge at the optic nerve head margin greater than a major retinal vessel, diverging in an arcuate or wedge shape; and (2) gonioscopically wide open angles of grade 3 or 4 based on the Shaffer classification, regardless of the presence of glaucomatous VF change. Subjects were not excluded by age, axial length, refractive error, or disease stage for this pretraining dataset. Patients with other ocular disorders that could affect the VF who were <20 years of age and eyes with possible secondary ocular hypertension in either eye were excluded.

Inclusion criteria for the normal group were no abnormal findings except for clinically insignificant senile cataract on biomicroscopy, gonioscopy, and fundoscopy and no history of ocular diseases that could affect the results of the spectral-domain OCT examinations, such as diabetic retinopathy or age-related macular degeneration. Other inclusion criteria were age >20 years and spherical equivalent refractive error ≥6.0 D and <3.0 D. Eyes with anomalous discs including tilted discs[26] were cautiously excluded. The normal diagnosis was made regardless of the status of the VF in this pretraining dataset, in contrast to the training and testing datasets, as detailed below.

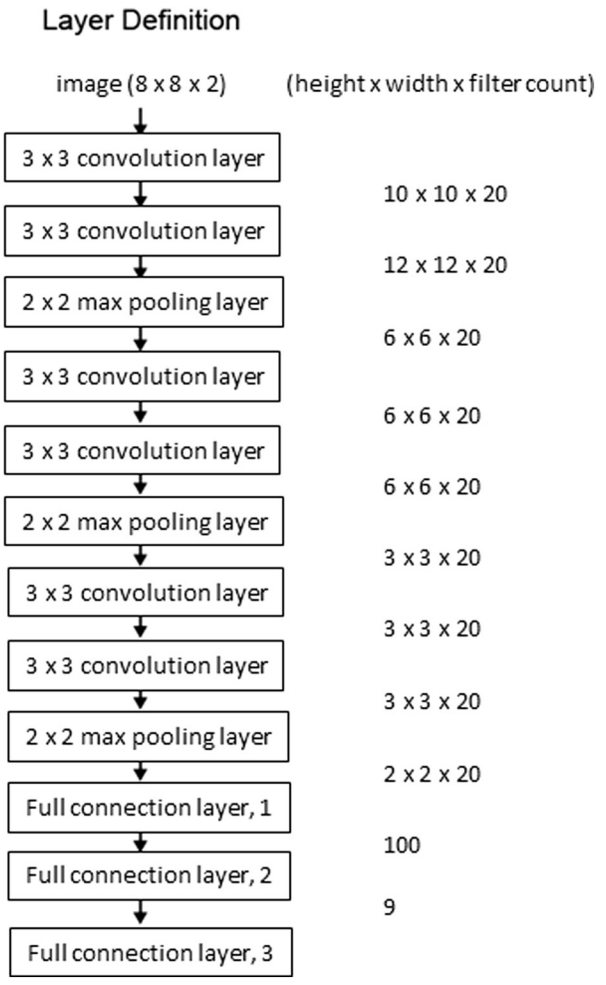

FIGURE 1. The deep learning algorithm to diagnose glaucoma using retinal nerve fiber layer and ganglion cell complex thicknesses from optical coherence tomography. This network has 6 convolutional layers and 6 maximum pooling layers in total. Each convolutional layer was followed by a batch normalization.

Training dataset. The training dataset was inherited from the training dataset in our previous study[8]; the glaucoma group consisted of 94 eyes from 94 subjects with OAG and the control group comprised 84 eyes from 84 normal subjects. Study participants were enrolled between January 2009 and March 2010 at the University of Tokyo Hospital or the Tajimi Iwase Eye Clinic.

Full details of the glaucoma group are described elsewhere,[27] but in short, a diagnosis was made when ophthalmoscopically apparent glaucomatous optic disc change was confirmed by a panel of glaucoma specialists (R.A., A.M., and N.S.), and glaucomatous VF change, defined by the Anderson–Patella criteria,[28] was present. The mean deviation (MD) value was required to be > -5 dB. Each evaluator made a diagnosis masked to each other's diagnoses. Subjects ≥20 years of age and eyes with refractive error ≥−6.0 D and <3.0 D were included. Eyes with other systemic or ocular disorders were carefully excluded. Eyes with anomalous discs including tilted discs[26] were also carefully excluded. The definition of POAG was identical to that in pretraining dataset, except for the requirement of MD > -5.0 dB.

Inclusion criteria for the normal group were inherited from our previous report[27]: no abnormal findings except for clinically insignificant senile cataract on biomicroscopy, gonioscopy, and fundoscopy, and no history of ocular diseases that could affect the results of spectral-domain OCT examinations, such as diabetic retinopathy or age-related macular degeneration. Other inclusion criteria were age >20 years, spherical equivalent refractive error ≥6.0 D and <3.0 D, and normal VF test results according to the Anderson–Patella criteria.[28] Eyes with anomalous discs including tilted discs[26] were cautiously excluded.

Testing dataset. The testing dataset was also used in our previous study.[8] This dataset consisted of 114 eyes of 114 subjects with OAG with a MD > -5 dB and 82 eyes of 82 normal subjects. Participants were enrolled prospectively after the training dataset was established, also at the Tokyo University Hospital, Kitasato University Hospital, or the Tajimi Iwase Eye Clinic. Inclusion and exclusion critera were identical to those applied in the training dataset; patients ≥20 years of age and refractive error ≥−6.0 D and <3.0 D were included, whereas those with other systemic or ocular disorders and anomalous discs were excluded. The diagnosis was made when ophthalmoscopically apparent glaucomatous optic disc change was confirmed by a panel of glaucoma specialists (R.A., A.M., and N.S.), and glaucomatous VF change, defined by the Anderson–Patella criteria,[28] was present. The criteria to normal group were completely identical to those used in the training dataset.

VF measurement. The details of VF measurements in the training and testing datasets are described elsewhere.[8] In short, VF testing was performed within 3 months of the spectral-domain OCT examination, using the Humphrey Field Analyzer with the SITA Standard strategy, the Goldmann size III target, and the 24-2 or 30-2 test program. Near refractive correction was used as necessary. All participants had previous experience in VF examinations, and unreliable VFs, defined as fixation losses >25% or false positive responses >15%, were excluded.[29]

• **SPECTRAL-DOMAIN OCT DATA ACQUISITION:** *RS 3000*. In the pretraining dataset, spectral-domain OCT data were obtained using the RS 3000. All spectral-domain OCT measurements were performed after pupil dilation with 1% tropicamide and OCT imaging was performed using the raster-scan protocol. Data obtained during apparent eye movements or influenced by involuntary blinking or saccade or those with a signal

TABLE 1. Demographics of the Training and Testing Datasets

	Training dataset			Testing dataset				
	Glaucoma	Normal	*P* Value[a]	Glaucoma	Normal	*P* Value[a]	*P* Value[b]	*P* Value[c]
Eye, right/left	50/44	65/19	.0013	56/58	53/29	.044	.66	.10
Age, y (mean ± SD)	57.9 ± 11.0	52.6 ± 15.6	.044	61.9 ± 13.9	52.6 ± 13.9	<.001	.0053	.50
Gender, male/female	35/59	47/37	.019	49/65	26/56	.15	.49	.0028
MD, dB (mean ± SD)	−2.7 ± 1.8	−0.40 ± 1.3	<.001	−2.5 ± 1.9	−0.36 ± 1.9	<.001	.37	.00020
Axial length, μm (mean ± SD)	24.5 ± 1.4	23.6 ± 0.78	<.001	24.5 ± 1.4	24.5 ± 1.4	.42	.97	<.0001
RNFL, μm (mean ± SD)	27.0 ± 16.5	36.8 ± 18.3	<.001	30.0 ± 16.9	37.1 ± 20.8	<.001	.00083	.21
GCC, μm (mean ± SD)	60.6 ± 15.2	68.6 ± 17.0	<.001	61.7 ± 15.6	69.1 ± 17.5	<.001	.87	.31

GCC = ganglion cell complex; MD = mean deviation; RNFL = retinal nerve fiber layer; SD = standard deviation.

Demographic data were compared between the training and testing datasets using the unpaired Wilcoxon test (numerical variables) or χ^2 test (categorical variables). MD was calculated from the 24-2 or 30-2 Humphrey visual field.

[a]Comparison between glaucomatous and normative eyes in the training or testing dataset.

[b]Comparison between glaucomatous eyes in the training dataset and those in the testing dataset.

[c]Comparison between normative eyes in the training dataset and those in the testing dataset.

TABLE 2. Demographics of the Pretraining Dataset

	Glaucoma	Normal	*P* Value
Eye, right/left	691/680	97/96	3.97
Age, y (mean ± SD)	58.5 ± 13.4	40.0 ± 16.7	<.001
Gender, male/female	347/400	39/74	<.001
RNFL, μm (mean ± SD)	28.8 ± 16.8	38.4 ± 18.7	<.001
GCC, μm (mean ± SD)	44.8 ± 18.0	58.2 ± 17.9	<.001
MD, dB (mean ± SD)	−5.9 ± 6.8	N/A	N/A

GCC = ganglion cell complex; MD = mean deviation; N/A = not applicable; RNFL = retinal nerve fiber layer; SD = standard deviation.

Demographic data were compared between the glaucoma and normal datasets using the unpaired Wilcoxon test (numerical variables) or χ^2 test (categorical variables). In the glaucoma group, average and SD values of MD were calculated from the 24-2 or 30-2 Humphrey visual field corresponding to the initial visual field. Visual field testing was not performed in all eyes in the normal group; visual field results were not included as criteria for defining the glaucoma and normal groups in this pretraining dataset.

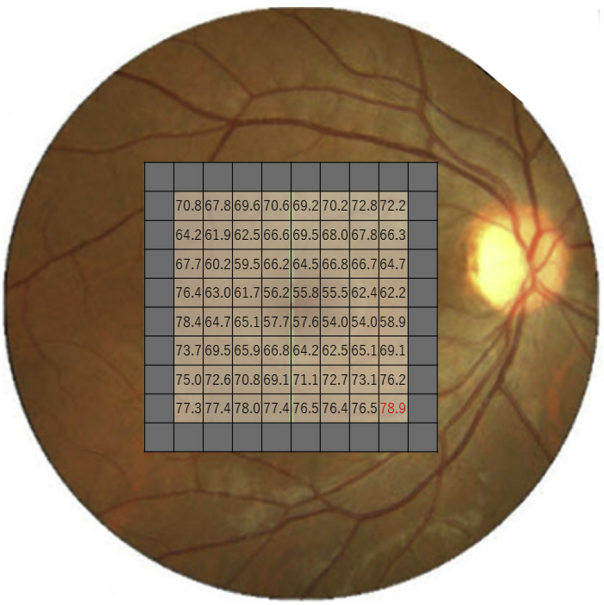

FIGURE 2. Receiver operating characteristic curves in the testing dataset, obtained with the 8 × 8 grid of retinal nerve fiber layer thickness values. Using the testing dataset, the areas under the receiver operating characteristic curve obtained with the raw values of the 8 × 8 retinal nerve fiber layer thicknesses varied between 54.0 (95% confidence interval 45.9-62.0%) and 78.9 (95% confidence interval 72.6-85.1%).

strength index <7 were excluded, as recommended by the manufacturer.

In the macula, RNFL and GCC thicknesses were exported from an imaging area (9 × 9 mm) centered on the fovea (512 × 128, 1024 × 64, 256 × 256, 256 × 128, or 512 × 64 pixels). Then a 6.0- × 6.0-mm area centered on the fovea was clipped so that the measured area matched that of the OCT-1000/2000 and was converted to a 10 × 10 grid by taking the average value in each square. To avoid the possibility that the analysis area extended outside of the 6.0- × 6.0-mm image, the outermost units were excluded, resulting in an inner 8 × 8 grid, following our previous paper.[8] Left eyes were mirror-imaged to a right eye orientation.

OCT-1000/2000. OCT-1000/2000 data in the training and testing datasets are described elsewhere.[8] These data were obtained using either the 3-dimensional (3D) OCT-1000 or 3D OCT-2000 (training dataset) or only

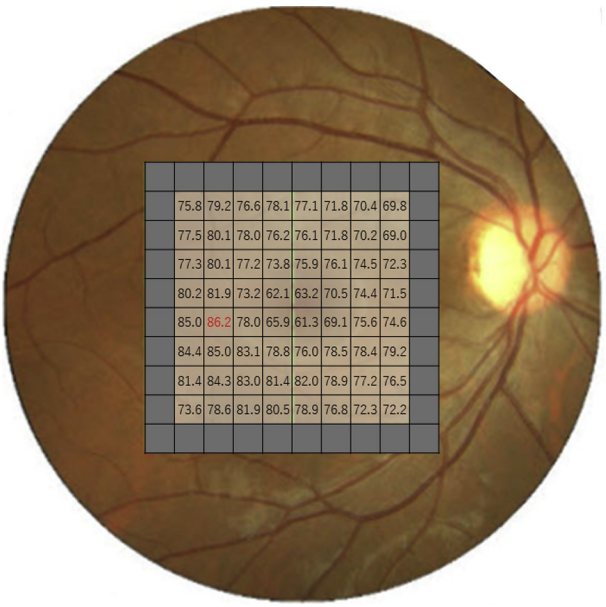

FIGURE 3. Receiver operating characteristic curves in the testing dataset, obtained with the 8 × 8 grid of ganglion cell complex thickness values. Using the testing dataset, the areas under the receiver operating characteristic curve obtained with the raw values of the 8 × 8 ganglion cell complex thicknesses varied between 61.3 (95% confidence interval 53.4-69.2%) and 86.2 (95% confidence interval 81.3-91.2%).

the 3D OCT-2000 (testing dataset), which are completely interchangeable. All measurements were performed after pupil dilation with1 % tropicamide and imaging was performed using the raster-scan protocol. Data obtained during apparent eye movements and influenced by involuntary blinking or saccade or those with quality factor <30% were excluded, as recommended by the manufacturer. In the macular area, a square area (6.0 × 6.0 mm) centered on the fovea was divided into 10 × 10 grids with an equal size for both RNFL and GCC. Outermost units were again excluded, which resulted in an inner 8 x 8 grid. Left eyes were mirror-imaged to a right eye orientation.

• **STATISTICAL ANALYSIS:** Demographic data were compared between the training and testing datasets using the unpaired Wilcoxon test (numerical variables) or χ^2 (categorical variables).

DL methods are similar to artificial neural networks, which process information via interconnected "neurons"; however, DL classifiers have many more hidden layers compared with artificial neural networks.[30,31] In the current study, a convolutional neural network DL method was constructed using 8 × 8 grid RNFL (in the first channel) and GCC data (in the second channel). As shown in Figure 1, the convolutional neural network to diagnose glaucoma consisted of 6 convolutional layers, 3 maximum pooling layers and 3 fully connected layers. Activation between layers was conducted using the rectified linear unit[32] function, and 20 filters with size 3, stride 1 were applied in the convolutional layers. Zero padding (width 2) was added only in the initial 2 convolutional layers. Drop out (rate = 0.60) was applied to the final convolutional layer. Similar to other machine learning methods, DL can suffer from overfitting, and therefore we applied L1 regulation[33] with a regulation rate of 0.00001 to mitigate this issue. Parameter values were decided using the training dataset. The DL model was pretrained using the pretraining dataset then further trained (only hyperparameters were updated) using the training dataset. Both pretraining and training datasets were taught to classify glaucoma/normal eyes from the OCT measurement, and therefore this DL model is a supervised model. Diagnostic performance was evaluated using the area under the receiver operating characteristic curve (AROC) in the testing dataset. For comparison, the DL algorithm was also trained using (1) the training dataset only (ie, no pretraining); (2) the pretraining dataset only (ie, no further training); and (3) the pretraining and training dataset simultaneously (ie, no transfer learning).

To better understand the usefulness of DL over machine learning methods, a random forests (RF) model with 10,000 decision trees and a support vector machine (SVM) model with the radial basis function were constructed. These models were trained using: (1) both the pretraining and training datasets simultaneously; (2) only the pretraining dataset; and (3) only the training dataset.

All statistical analyses were carried out using the programming language Python (version 3.6.1; Python Software Foundation; Wilmington, DE, USA). The DL model was constructed using the package Chainer. Comparison of AROCs was carried out using the DeLong method.[34] The Wilcoxon test (for numerical data) and χ^2 test (for categorical data) were used to compare a variable between 2 groups. The Holm method[35,36] was used to correct *P* values for the problem of multiple testing.

RESULTS

DEMOGRAPHICS OF THE TRAINING AND TESTING DATASETS are shown in Table 1. The average RNFL and GCC thicknesses were significantly smaller in the glaucoma group compared with the normal group in both datasets ($P < .001$, unpaired Wilcoxon test). Axial length was significantly smaller in the glaucoma group compared with the normal group in the training dataset ($P < .001$) but not in the testing dataset. Table 2 shows the demographics of the pretraining dataset.

Using the raw values of the 8 × 8 RNFL thickness values, the AROCs obtained in the testing dataset varied between 54.0% (95% confidence interval [CI]

TABLE 3. Results of Receiver Operating Characteristic Analysis

Parameters	AROC (%)	95% CI	*P* Value	Optimum Discrimination at Sensitivity (%)/Specificity (%)	Sensitivity at 80% Specificity (%)	Sensitivity at 90% Specificity (%)
DL						
Transform model	93.7	90.6-96.8	—	82.5/93.9	83.3	86.6
Both training datasets	78.2	71.9-84.6	<.001	71.1/79.3	68.5	34.3
Training dataset	76.6	69.7-83.4	<.001	76.3/70.7	51.6	32.5
Pretraining dataset	78.8	72.7-85.0	<.001	85.1/69.5	58.1	29.1
Random forest						
Both training datasets	82.0	76.2-87.8	<.001	64.0/89.0	67.5	62.3
Training dataset	86.2	81.0-91.3	<.001	66.7/97.6	76.3	71.9
Pretraining dataset	83.1	77.4-88.8	<.001	64.0/92.7	73.7	64.2
Support vector machine						
Both training datasets	67.4	60.0-74.9	<.001	61.4/65.9	39.5	35.1
Training dataset	75.9	69.3-82.5	<.001	50.0/96.3	58.8	52.6
Pretraining dataset	63.1	55.670.5	<.001	39.5/91.5	45.1	39.5

AROC = area under the receiver operating characteristic curve; CI = confidence interval; DL = deep learning.

Model AROC values were compared against those of the DL transform model using the DeLong method.[34] The Holm method[35,36] was used to correct *P* values for the problem of multiple testing. The optimum discrimination was calculated using the Youden method.[37]

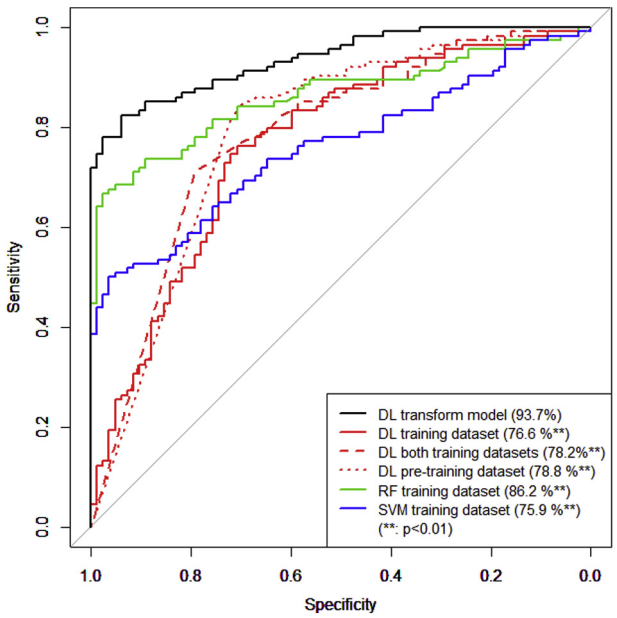

FIGURE 4. Receiver operating characteristic curves in the testing dataset, obtained with the deep learning (DL), random forest (RF), and support vector machine (SVM) methods. With DL, the receiver operating characteristic curve was calculated using the (1) pretraining and training (transform model); (2) training dataset only; (3) pretraining and training datasets in 1 process (both training datasets); and (4) pretraining dataset only. *P* values were obtained by comparing the areas under the receiver operating characteristic curve with the DL and transform models.

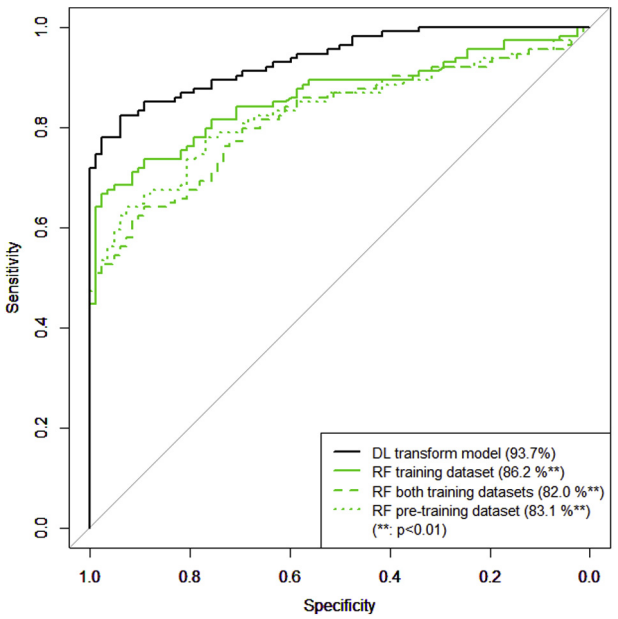

FIGURE 5. Receiver operating characteristic curves in the testing dataset, obtained with the deep learning (DL) and random forest (RF) methods. With DL, the receiver operating characteristic curve was calculated using (1) pretraining and training (transform model); (2) training dataset only; (3) pretraining and training datasets in 1 process (both training datasets); and (4) pretraining dataset only. *P* values were obtained by comparing the areas under the receiver operating characteristic curve with the DL and transform models.

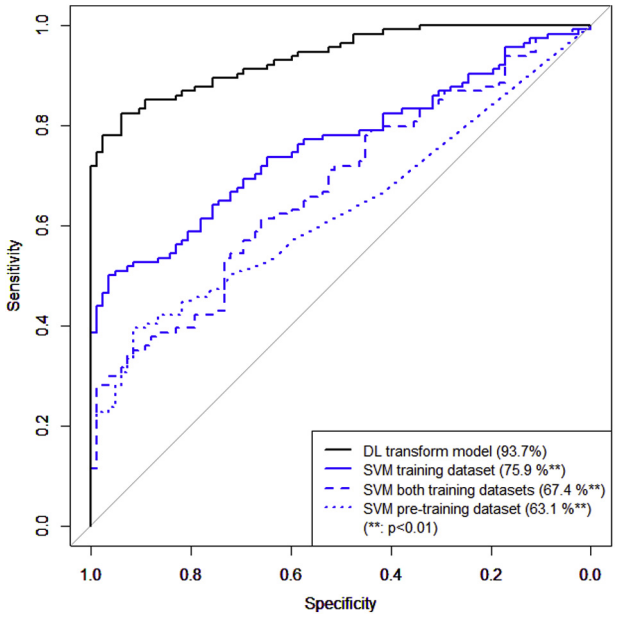

FIGURE 6. Receiver operating characteristic curves in the testing dataset, obtained with the deep learning (DL) and support vector machine (SVM) methods. With DL, the receiver operating characteristic curve was calculated using (1) pretraining and training (transform model); (2) training dataset only; (3) pretraining and training datasets in 1 process (both training datasets); and (4) pretraining dataset only. *P* values were obtained by comparing the areas under the receiver operating characteristic curve with the DL and transform models.

45.9-62.0%) and 78.9% (95% CI 72.6-85.1%), as shown in Figure 2. Similarly, using the testing dataset the AROCs obtained with the raw values of the 8 × 8 GCC thicknesses varied between 61.3% (95% CI 53.4-69.2%) and 86.2% (95% CI 81.3-91.2%), as shown in Figure 3.

The DL transfer model (ie, the model that was pretrained and then further trained) achieved an AROC of 93.7% (95% CI 90.6-96.8%), which was significantly larger than the AROC values (from 63.1-to 86.2%) of all other models ($P < .001$, using the DeLong method and the Holm method for multiple comparisons); see Table 3 and Figures 4, 5, and 6. The optimum discrimination of the model achieved a sensitivity of 82.5% and specificity of 93.9% using the Youden method.[37]

DISCUSSION

IN THE CURRENT STUDY, A DL CLASSIFIER WAS DEVELOPED to diagnose early glaucoma using OCT measurements. The classifier was first pretrained using a relatively large OCT dataset obtained with the RS3000 OCT, then further trained with a relatively small OCT dataset obtained with a different OCT instrument (OCT 1000/2000, 178 pairs of RNFL and GCC images). As a result, the AROC value of this classifier was 93.7% in discriminating early glaucoma eyes from normal eyes in an independent testing dataset. This AROC was significantly larger than that obtained from DL classifiers constructed using different training methods: (1) a DL model that was not pretrained (AROC 78.8%); (2) a DL classifier that was not further trained (AROC 76.6%); and (3) a DL model trained using the pretraining and training datasets simultaneously, in 1 training process (AROC 78.2%). In addition, the AROC value of the DL transfer model was significantly larger than those associated with other machine learning methods (RF and SVM).

Early detection of glaucoma is essential, but diagnosing early-onset glaucoma is a much more challenging task than diagnosing advanced glaucoma.[38-40] Tan and associates reported an AROC of 90% with mean GCC thickness over the macular area in discriminating early (average MD −4.6 dB) glaucoma eyes from normal eyes.[41] Kim and associates reported AROCs between 82.6% and 89.5% with total and superior and inferior GCC thickness in glaucoma eyes with an average MD of −8.49 dB.[42] In the current study, the average MD of glaucomatous eyes in the testing dataset was −2.5 dB, and the AROCs obtained were between 54.0% and 86.2% for GCC or RNFL thickness.

For the accurate diagnosis of glaucoma it is beneficial to analyze the multiple structural parameters from spectral-domain OCT comprehensively, using methods such as a logistic regression model,[5] a SVM classifier,[6] and a decision tree classifier.[7] By building a RF classifier with the same training and testing datasets used in the current study, we recently reported that an accurate diagnosis of glaucoma can be made.[8] The usefulness of the RF method in diagnosis/prediction[43,44] is well-recognized, in particular when there are interactions between different explanatory variable.[45-47] The successes of DL methods for classification have also been reported in various research fields, including computer vision[48-50] and natural language processing.[51,52] DL classifiers often significantly outperform other machine learning methods.[11,12] Indeed, the DL transfer model achieved an AROC value of 93.7%, which is almost identical to that obtained with a RF model using both GCC and RNFL thickness values, but also circumpapillary RNFL (cpRNFL) thickness, as shown in our previous study.[8] The usefulness of DL methods over the RF model was also suggested in our previous study, where we reported a significantly better diagnostic performance of the DL model for early detection of glaucomatous VF change (preperimetric glaucomatous VFs) compared with the RF method.[13]

Despite its great potential, the application of DL to OCT data is unusual. One reason may be the requirement of a very large dataset because DL usually requires a larger training dataset to obtain an accurate diagnosis. Unfortunately, OCT data from different instruments are not

interchangeable, so it is difficult to collect such a large dataset, even when OCT data are collected across multiple institutes. Even data collected from OCT machines made by the same manufacturer may not be interchangeable if the device firmware has been updated. In the current study, transfer learning was used to build an initial DL model with a large pretraining dataset and then further train this model with different OCT data. The usefulness of this DL pretraining process, using disparate data, has been reported in many research fields.[53] Very large datasets are often used, such as 1.3 million images in the LSVRC14 dataset, 80 million images in CIFAR-10, and >14 million images in the ImageNet dataset.[54] The pretraining dataset used in the current study was much smaller (4315 images) compared with these previous reports; nonetheless, a very high AROC value was obtained.

The SVM model is another machine learning method. In Burgansky-Eliash and associates,[6] an SVM built with multiple OCT parameters resulted in a very high AROC (98.1%) to discriminate normal and glaucoma eyes. We investigated its usefulness; however, the model achieved a much smaller AROC value compared with the DL and RF methods. It should be also noted that the AROC values with DL were even smaller than the those associated with RF and SVM, when DL was not pretrained. SVM and RF cannot use this pretraining technique followed by the transforming technique. Also, it is not recommended to train DL using multiple OCT data, as suggested by the significantly smaller AROC values with DL trained using both pretraining and training datasets in 1 process compared with that of transform learning. This is because retinal layer thicknesses measured with different OCT devices are very different and using a mixture of such data is not appropriate.

In a recent review by Hood,[55] it was suggested that a single OCT scan has the potential to replace the VF measurement in future; however, it was not recommended at this point that the glaucoma specialists should do away with VF test. The current results suggest that DL can accurately diagnose early-onset glaucoma from a large OCT dataset, even if it is measured with a considerably different OCT machine. However, further efforts are necessary to build a model to predict a patient's VF from OCT data and not only to diagnose glaucoma.

One limitation of the current study is the lack of cpRNFL data. Unfortunately, this measure was not available in all of the pretraining dataset. In our previous report,[8] using RNFL, GCC, and cpRNFL thickness measurements was important for a RF trained with the same training dataset. Indeed, this RF model could discriminate glaucomatous and normal eyes in the testing dataset with an AROC value of 93.0%, which is similar to that of the DL transfer model (93.7%). Therefore, the DL classifier might be improved by including cpRNFL measurements. In addition, the results were obtained using a homogeneous patient population (ie, only Japanese patients). Therefore, a further study is needed, preparing OCT data from multiple ethnicities, to generalize the current results to patients of other ethnicities.

In conclusion, we have constructed a DL model with high sensitivity and specificity to diagnose glaucoma. As a result, this model has a high diagnostic ability and may be useful to support clinicians identify early-onset glaucoma.

FUNDING/SUPPORT: THIS STUDY WAS SUPPORTED IN PART BY JAPAN SCIENCE AND TECHNOLOGY AGENCY CORE RESEARCH for Evolutional Science and Technology grant JPMJCR1304 and Ministry of Education, Culture, Sports, Science and Technology of Japan grant 17K11418. The following authors have no financial disclosures: Ryo Asaoka, Hiroshi Murata, Kazunori Hirasawa, Yuri Fujino, Masato Matsuura, Atsuya Miki, Takashi Kanamoto, Yoko Ikeda, Kazuhiko Mori, Aiko Iwase, Nobuyuki Shoji, Kenji Inoue, Junkichi Yamagami, and Makoto Araie. All authors attest that they meet the current ICMJE criteria for authorship.

REFERENCES

1. Weinreb RN, Khaw PT. Primary open-angle glaucoma. *Lancet* 2004;363(9422):1711–1720.
2. Fechtner RD, Weinreb RN. Mechanisms of optic nerve damage in primary open angle glaucoma. *Surv Ophthalmol* 1994;39(1):23–42.
3. Shields MB. Textbook of Glaucoma. Baltimore, MD: William & Wilkins; 1997.
4. Zimmerman TJ, Kooner KS. Clinical Pathways in Glaucoma. New York, NY: Thieme; 2001.
5. Mwanza JC, Warren JL, Budenz DL, Ganglion Cell Analysis Study Group. Combining spectral domain optical coherence tomography structural parameters for the diagnosis of glaucoma with early visual field loss. *Invest Ophthalmol Vis Sci* 2013;54(13):8393–8400.
6. Burgansky-Eliash Z, Wollstein G, Chu T, et al. Optical coherence tomography machine learning classifiers for glaucoma detection: a preliminary study. *Invest Ophthalmol Vis Sci* 2005;46(11):4147–4152.
7. Baskaran M, Ong EL, Li JL, et al. Classification algorithms enhance the discrimination of glaucoma from normal eyes using high-definition optical coherence tomography. *Invest Ophthalmol Vis Sci* 2012;53(4):2314–2320.
8. Asaoka R, Hirasawa K, Iwase A, et al. Validating the usefulness of the "random forests" classifier to diagnose early glaucoma with optical coherence tomography. *Am J Ophthalmol* 2017;174:95–103.
9. Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. *Neural Comput* 2006;18(7):1527–1554.
10. Boureau Y, Cun Y. Sparse feature learning for deep belief networks. *Adv Neural Inform Process Syst* 2008;1–8.

11. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* 2006;313(5786):504–507.
12. Eickholt J, Cheng J. DNdisorder: predicting protein disorder using boosting and deep networks. *BMC Bioinformatics* 2013;14:88.
13. Asaoka R, Murata H, Iwase A, Araie M. Detecting preperimetric glaucoma with standard automated perimetry using a deep learning classifier. *Ophthalmology* 2016;123(9):1974–1980.
14. Quigley HA, Broman AT. The number of people with glaucoma worldwide in 2010 and 2020. *Br J Ophthalmol* 2006;90(3):262–267.
15. Leite MT, Rao HL, Zangwill LM, Weinreb RN, Medeiros FA. Comparison of the diagnostic accuracies of the Spectralis, Cirrus, and RTVue optical coherence tomography devices in glaucoma. *Ophthalmology* 2011;118(7):1334–1339.
16. Lisboa R, Paranhos A Jr, Weinreb RN, Zangwill LM, Leite MT, Medeiros FA. Comparison of different spectral domain OCT scanning protocols for diagnosing preperimetric glaucoma. *Invest Ophthalmol Vis Sci* 2013;54(5):3417–3425.
17. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? *Adv Neural Inform Process Syst* 2014;27:3320–3328.
18. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems* 2012:1097-1105.
19. Szegedy C, Liu W, Jia Y. Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2015:1-9.
20. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 2017;284(2):574–582.
21. Bar Y, Diamant I, Wolf L, Lieberman S, Konen E, Greenspan H. Chest pathology detection using deep learning with non-medical training. *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on IEEE* 2015:294-297.
22. Roth HR, Lu L, Liu J, et al. Improving computer-aided detection using convolutional neural networks and random view aggregation. *IEEE Trans Med Imaging* 2016;35(5):1170–1181.
23. Shin HC, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging* 2016;35(5):1285–1298.
24. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:14091556* 2014.
25. Sugiura H, Kiwaki T, Yousefi S, Murata H, Asaoka R, Yamanishi K. Estimating glaucomatous visual sensitivity from retinal thickness by using pattern-based regularization and visualization. *24th ACM SIGKDD International Conference*; 2018.
26. Apple DJ, Rabb MF, Walsh PM. Congenital anomalies of the optic disc. *Surv Ophthalmol* 1982;27(1):3–41.
27. Yoshida T, Iwase A, Hirasawa H, et al. Discriminating between glaucoma and normal eyes using optical coherence tomography and the 'random forests' classifier. *PLoS One* 2014;9(8):e106117.
28. Anderson DR, Patella VM. Automated Static Perimetry. 2nd ed. St. Louis, MO: Mosby; 1999.
29. Bengtsson B, Heijl A. False-negative responses in glaucoma perimetry: indicators of patient performance or test reliability? *Invest Ophthalmol Vis Sci* 2000;41(8):2201–2204.
30. Vincent P, Larochelle H, Bengio Y, Manzagol PA. Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland; 2008.
31. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, PMLR 2010;249-256.
32. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS-11)* 2011;15:315-323.
33. Scardapane S, Comminiello D, Hussain A, Uncini A. Group sparse regularization for deep neural networks. Available at: https://arxivorg/abs/16070.0485. Accessed November 17, 2018.
34. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44(3):837–845.
35. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat* 1979;6:65–70.
36. Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *Am J Public Health* 1996;86(5):726–728.
37. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950;3(1):32–35.
38. Schulze A, Lamparter J, Pfeiffer N, Berisha F, Schmidtmann I, Hoffmann EM. Diagnostic ability of retinal ganglion cell complex, retinal nerve fiber layer, and optic nerve head measurements by Fourier-domain optical coherence tomography. *Graefes Arch Clin Exp Ophthalmol* 2011;249(7):1039–1045.
39. Rao HL, Zangwill LM, Weinreb RN, Sample PA, Alencar LM, Medeiros FA. Comparison of different spectral domain optical coherence tomography scanning areas for glaucoma diagnosis. *Ophthalmology* 2010;117(9):1692–1699.
40. Moreno PA, Konno B, Lima VC, et al. Spectral-domain optical coherence tomography for early glaucoma assessment: analysis of macular ganglion cell complex versus peripapillary retinal nerve fiber layer. *Can J Ophthalmol* 2011;46(6):543–547.
41. Tan O, Chopra V, Lu AT, et al. Detection of macular ganglion cell loss in glaucoma by Fourier-domain optical coherence tomography. *Ophthalmology* 2009;116(12):2305–2314.
42. Kim NR, Lee ES, Seong GJ, Kim JH, An HG, Kim CY. Structure-function relationship and diagnostic value of macular ganglion cell complex measurement using Fourier-domain OCT in glaucoma. *Invest Ophthalmol Vis Sci* 2010;51(9):4646–4651.
43. Breiman L. Random forests. *Machine Learning* 2001;45:5–32.
44. Breiman L, Cutler A. Random forests. San Diego, CA: Salford Systems; 2004.
45. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet* 2004;5:32.

46. Segal MR, Cummings MP, Hubbard AE. Relating amino acid sequence to phenotype: analysis of peptide-binding data. *Biometrics* 2001;57(2):632–642.
47. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. *BMC Bioinformatics* 2008;9:307.
48. Kavukcuoglu K, Sermanet P. Learning convolutional feature hierarchies for visual recognition. *Adv Neural Inform Process Syst* 2010;1–9.
49. Taylor G, Fergus R, LeCun Y, Bregler C. Convolutional learning of spatio-temporal features. *Computer VisionECCV* 2010;140–153.
50. Lee H, Grosse R, Ranganath R, Ng A. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *International Conference on Machine Learning (ICML)* 2009:1-8.
51. Sochard R, Lin CCY, Ng AY, Manning CD. Parsing natural scenes and natural language. *International Conference on Machine Learning (ICML)* 2011.
52. Collobert R. Deep learning for efficient discriminative parsing. *International Conference on Artificial Intelligence and Statistics* 2011.
53. Goodfellow I, Bengio Y, Courville A. Deep Learning (Adaptive Computation and Machine Learning series). Cambridge, MA: The MIT Press; 2016.
54. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vision* 2015;115(3):211–252.
55. Hood DC. Improving our understanding, and detection, of glaucomatous damage: An approach based upon optical coherence tomography (OCT). *Prog Retin Eye Res* 2017;57:46–75.