Quantification of Key Retinal Features in Early and Late Age-Related Macular Degeneration Using Deep Learning

BART LIEFERS, PAUL TAYLOR, ABDULRAHMAN ALSAEDI, CLARE BAILEY, KONSTANTINOS BALASKAS, NARENDRA DHINGRA, CATHERINE A. EGAN, FILIPA GOMES RODRIGUES, CRISTINA GONZÁLEZ GONZALO, TJEBO F.C. HEEREN, ANDREW LOTERY, PHILIPP L. MÜLLER, ABRAHAM OLVERA-BARRIOS, BOBBY PAUL, ROY SCHWARTZ, DARREN S. THOMAS, ALASDAIR N. WARWICK, ADNAN TUFAIL, AND CLARA I. SÁNCHEZ

• PURPOSE: We sought to develop and validate a deep learning model for segmentation of 13 features associated with neovascular and atrophic age-related macular degeneration (AMD).

• DESIGN: Development and validation of a deep-learning model for feature segmentation.

• METHODS: Data for model development were obtained from 307 optical coherence tomography volumes. Eight experienced graders manually delineated all abnormalities in 2712 B-scans. A deep neural network was trained with these data to perform voxel-level segmentation of the 13 most common abnormalities (features). For evaluation, 112 B-scans from 112 patients with a diagnosis of neovascular AMD were annotated by 4 independent observers. The main outcome measures were Dice score, intraclass correlation coefficient, and free-response receiver operating characteristic curve.

• RESULTS: On 11 of 13 features, the model obtained a mean Dice score of 0.63 ± 0.15, compared with 0.61 ± 0.17 for the observers. The mean intraclass correlation coefficient for the model was 0.66 ± 0.22, compared with 0.62 ± 0.21 for the observers. Two features were not evaluated quantitatively because of a lack of data. Free-response receiver operating characteristic analysis demonstrated that the model scored similar or higher sensitivity per false positives compared with the observers.

• CONCLUSIONS: The quality of the automatic segmentation matches that of experienced graders for most features, exceeding human performance for some features. The quantified parameters provided by the model can be used in the current clinical routine and open possibilities for further research into treatment response outside clinical trials. (Am J Ophthalmol 2021;226:1–12. © 2021 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).)

AGE-RELATED MACULAR DEGENERATION (AMD) IS A complex disease affecting elderly people that can lead to severe vision loss in the advanced (late) stages.[1] Therapy with intraocular injection of anti–vascular endothelial growth factor (anti-VEGF) is effective and reduces the risk of visual loss from the neovascular form of late AMD, but the costs and number of required injections form a significant burden on health care systems and patients.[2–4] To reduce the treatment burden, the use of personalized treatment intervals has been investigated, which necessitates monitoring of disease activity to guide treatment and follow-up interval choice.[5–7] In the longer term, what may limit visual outcomes in patients treated for neovascular AMD is either related to the neovascular process or caused by progression of the late atrophic form of AMD.

The monitoring is typically performed by optical coherence tomography (OCT), providing cross-sectional images

AJO.com Supplemental Material available at AJO.com.
Accepted for publication Dec 28, 2020.

From the A-eye Research Group (B.L., C.G.G., C.I.S.), Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, The Netherlands; Donders Institute for Brain (B.L., C.G.G., C.I.S.), Cognition and Behaviour, Radboud University Medical Center, Nijmegen, The Netherlands; Institute of Health Informatics (P.T., D.S.T.), University College London (UCL), London, UK; College of Medicine (A.A.), Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia; Moorfields Eye Hospital NHS Foundation Trust (A.A., C.A.E., F.G.R., T.F.C.H., P.L.M., A.O-B., R.S., A.N.W., A.T.) London; University Hospitals Bristol NHS Foundation Trust (C.B.), Bristol; Moorfields Ophthalmic Reading Centre (K.B.), Moorfields Eye Hospital NHS Foundation Trust, London; National Institute for Health Research Biomedical Research Centre at Moorfields Eye Hospital National Health Service (NHS) Foundation Trust (K.B. C.A.E., F.G.R.), UCL Institute of Ophthalmology, London; Mid Yorkshire Hospitals NHS Trust (N.D.), Southampton; Institute of Ophthalmology (T.F.C.H., P.L.M., A.O-B., A.T.), University College London, London, UK; Faculty of Medicine (A.L.), University of Southampton, Southampton; Department of Ophthalmology (P.L.M.), University of Bonn, Bonn, Germany; Barking, Havering and Redbridge University Hospitals NHS Trust (B.P.), Romford, United Kingdom; Institute of Cardiovascular Science (A.N.W.), University College London, London; Department of Ophthalmology (C.I.S.), Radboud University Medical Center, Nijmegen, The Netherlands.

B.L. is currently affiliated with Moorfields Ophthalmic Reading Centre, Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom, and the Department of Ophthalmology, Erasmus University Medical Center, Rotterdam, The Netherlands. C.I.S. is currently affiliated with the Informatics Institute, Faculty of Science, University of Amsterdam, Amsterdam, The Netherlands, and the Department of Biomedical Engineering and Physics, Amsterdam University Medical Center, Amsterdam, The Netherlands.

Inquiries to Bart Liefers, Moorfields Ophthalmic Reading Centre, Moorfields Eye Hospital NHS Foundation Trust, 162 City Rd, London EC1V 2PD, United Kingdom; e-mail: b.liefers@nhs.net

0002-9394
https://doi.org/10.1016/j.ajo.2020.12.034

of the retina that allow identification of fluid and other pathomorphologic features, such as drusen, subretinal hyperreflective material (SHRM), or atrophy.[8–10] In addition, the extent (size and distribution) of these features can be measured from OCT, providing improved diagnostic and prognostic information.[11–13] However, manual volumetric quantification of these features is problematic because of the large amount of work required.[14] Therefore, treatment decisions in current clinical routine rely largely on estimated parameters, which are, particularly in comparing follow-up visits, prone to error, subjectivity, and bias.[14,15]

Machine learning algorithms can provide a powerful support tool in automatic quantification of pathologic features.[9] Promising results in the application of deep learning algorithms for automated classification of retinal diseases,[16,17] referral recommendation,[18] and prediction of conversion to neovascular AMD[19] have already been achieved. These models are generally developed to solve a specific task, providing global (image or patient-level) output for the input they receive. On the other hand, models for volumetric segmentation provide labels to every voxel in the input. This generates a precise morphologic signature of the disease, from which features can be assessed independently and in combination, taking into account their size and spatial distribution. This can be used to model interactions and to create standardized reports that are flexible for adaptation to different clinical guidelines or treatment protocols. Moreover, automated volumetric quantification of these parameters enables comprehensive study of structure/function correlation and the development of prediction models for treatment response and personalized treatment intervals.[20,21]

Several algorithms for the segmentation or quantification of retinal pathology in OCT have been proposed. A limitation of these algorithms is that they often only focus on a single feature, or a subset of relevant features, with segmentation of fluid receiving most attention.[14,22–25] Other pivotal work, in which segmentation of a larger set of normal and abnormal structures is addressed, does not provide quantitative validation of performance of the segmentation model.[18,19] Validation on large real-world datasets is an important prerequisite, both for the integration of these algorithms into the current clinical workflow and for their application in the study of response to anti-VEGF therapy outside of clinical trials.

In this article, we present a novel deep learning model for segmentation of 13 features commonly found in neovascular and atrophic AMD. The model is developed and validated on a large dataset of OCT volumes collected from secondary care providers in the United Kingdom (UK). The performance of the model is compared against 4 independent graders to validate the quality of the automatic segmentation regarding overlap as well as quantification.

METHODS

• **STUDY POPULATION:** Imaging data for this study were obtained from 30,337 patients at 5 centers in the UK. The institutional review board (NRES Committee London, City Road and Hampstead, London) ruled that approval was not required for this study because all data were completely anonymized before being released to research. The study adhered to the Declaration of Helsinki.

Eyes were scanned with either a Topcon 3D OCT-1000 or a Topcon 3D OCT-2000 (Topcon, Tokyo, Japan), resulting in OCT volumes with a resolution of either 512 × 128 A-scans or 256 × 256 A-scans, covering a macular area of 6 mm × 6 mm. Two independent sets were created. First, we created a set that was used for model development, in which all abnormalities in a sparse subset of B-scans per OCT volume were manually delineated. Second, we created a test set in which 4 graders independently delineated all abnormalities in a single selected B-scan per OCT volume and that was used to assess agreement between graders and the model.

OCT scans were selected by querying the electronic medical records database (Medisoft, Leeds, UK). To select the set of scans that was used for development of the segmentation model, the table with indications for treatment with anti-VEGF was used to enrich the number of abnormalities (more details on the data selection procedure can be found in the Supplemental Material). This set included scans with retinal diseases other than AMD, presenting the model with a diverse representation of features. A total of 2712 annotated B-scans from 307 OCT scans from 307 eyes were used for model development.

Data in the test set were restricted to patients presenting with neovascular AMD. Inclusion criteria were eyes receiving 2 mg aflibercept or 0.3 mg to 0.5 mg ranibizumab for surgical indications associated with neovascular AMD for patients ≥50 years of age and a best-corrected visual acuity of 6/12 to 6/96 or 25 to 73 Early Treatment Diabetic Retinopathy Study (ETDRS) letters inclusively within 30 days of baseline (first injection), following guideline NG82 of the UK National Institute for Clinical Excellence. We excluded eyes that 1) received unlicensed Avastin before baseline; 2) were ever affected by clinically significant macular edema before baseline; 3) underwent cataract surgery or phacoemulsification within 90 days of baseline; or 4) had ever received macular laser therapy or panretinal photocoagulation before baseline. This query resulted in a set of 1798 patients eligible for inclusion in the test set. The final test set comprised a random subset of 112 OCT scans, from 112 eyes, independent of the development data. Fifty-six scans were obtained from treatment-naive eyes, just before the first anti-VEGF injection. The other 56 were acquired approximately 3 months after the first injection. This provided a varied set of representations of different features, as the largest changes in retinal morphology occur in the early phase of treatment.[26]

- **MANUAL ANNOTATIONS:** Manual delineation of features was performed at the Moorfields Reading Centre by 8 experienced graders (medical retinal fellowship–trained) using a custom annotation platform. The graders had access to both the OCT volume and a corresponding color fundus image, and could zoom in for accurate delineation, using either the mouse or a touch device. Graders were instructed to delineate all abnormalities in a sparse set of 5 of the 128 or 256 B-scans per OCT volume (including the foveal B-scan), or they could indicate, per B-scan, that it did not contain any abnormalities. The annotation platform provided them with default labels for the most common abnormalities and allowed them to add new labels if they encountered rare abnormalities. This was done to ensure that all abnormalities were delineated. A document containing instructions and examples of all abnormalities of interest was circulated and discussed with the graders before the start of the data collection process.

Four graders (a subset of the graders that contributed to the data for model development) independently annotated the scans in the test set using the same annotation platform. To ensure that everyone graded the same B-scan, a single B-scan per OCT volume was selected and highlighted in the viewer. This B-scan was selected from a random normal distribution, centered on the central B-scan and with a sigma of 0.75 mm (16 B-scans for volumes with 128 B-scans). The higher probability for central slices was chosen to increase the chances of finding features of interest.

For evaluation of the model's performance, the delineations of the graders were combined to create a consensus reference. Manual delineation of most of the included features is susceptible to the graders' interpretation, and discrepancies related to inexact outlining of lesion boundaries can be considerable, especially for small lesions with indistinct borders. By combining the delineations of multiple graders, we could obtain a more reliable reference standard. In addition, this enabled the distinction between voxels that are genuinely misclassified vs voxels for which the interpretation is ambiguous.

The delineations of 3 of the 4 graders were used to create the reference, using the delineations of the fourth grader to obtain an estimate of human performance. The set of the 3 reference graders was rotated, giving 4 estimates of performance, both for the model and for manual grading. The reference consisted of those voxels where ≥2 of 3 reference graders annotated the respective feature, whereas voxels that were annotated by only 1 of the 3 graders were ignored.

- **MODEL DEVELOPMENT:** Thirteen of the most commonly annotated abnormalities were selected to be included in the development of the automated segmentation model, henceforth referred to as features. Included features were: intraretinal fluid (IRF), subretinal fluid (SRF), pigment epithelial detachment (PED), SHRM, fibrosis, drusen and drusenoid pigment epithelial detachments (PEDs; Drusen), epiretinal membrane (ERM), outer plexiform layer descent (OPL descent), ellipsoid loss, retinal pigment epithelium loss or attenuation (RPE loss), hypertransmission (HTR), hyperreflective dots and exudates (HRD), and subretinal drusenoid deposits–reticular pseudodrusen (SDD-RPD). Some of the features (such as fibrous and serous PED) were combined, and features occurring in <81 B-scans in the development data were not modeled. Table 1 summarizes the number of occurrences of each feature in the training and test sets.

TABLE 1. Included Features, Occurrence in Training, and Test Set[a]

Feature	Training B-Scans (N = 2712), n (%)	Test B-Scans (N = 112), Mean ± SD
Ellipsoid loss	930 (34.29)	84.8 ± 10.9
IRF	639 (23.56)	31.0 ± 5.8
PED	549 (20.24)	54.8 ± 4.4
HRD	525 (19.36)	39.2 ± 12.6
SRF	406 (14.97)	35.8 ± 3.3
ERM	373 (13.75)	25.5 ± 9.2
HTR	289 (10.66)	32.8 ± 12.2
SDD-RPD	284 (10.47)	28.0 ± 16.4
Drusen	265 (9.77)	48.8 ± 11.8
RPE loss	249 (9.18)	31.2 ± 18.3
OPL descent	150 (5.53)	14.2 ± 15.0
SHRM	148 (5.46)	21.2 ± 10.6
Fibrosis	81 (2.99)	5.0 ± 3.2

ERM = epiretinal membrane; HRD = hyperreflective dot; HTR = hypertransmission; IRF = intraretinal fluid; OPL = outer plexiform layer; PED = pigment epithelial detachment; RPE = retinal pigment epithelium; SD = standard deviation; SDD-RPD = subretinal drusenoid deposits–reticular pseudodrusen; SHRM = subretinal hyperreflective material; SRF = subretinal fluid.

[a] For the test set, the numbers refer to the means ± SDs of number of B-scans containing the feature as annotated by the 4 graders.

The model was implemented as a convolutional neural network, with a network architecture inspired by the variation of U-Net proposed by de Fauw and associates.[18] It can be characterized by an encoder–decoder structure with shortcut-connections. The encoder converts the high-resolution input into a low-resolution abstract representation. The original resolution is reconstructed in the decoder path. The model operates mainly on 2-dimensional B-scans, but for every B-scan contextual information from 9 adjacent B-scans is included in the deeper layers.

The network architecture is fully convolutional, which means it can be applied to images of any

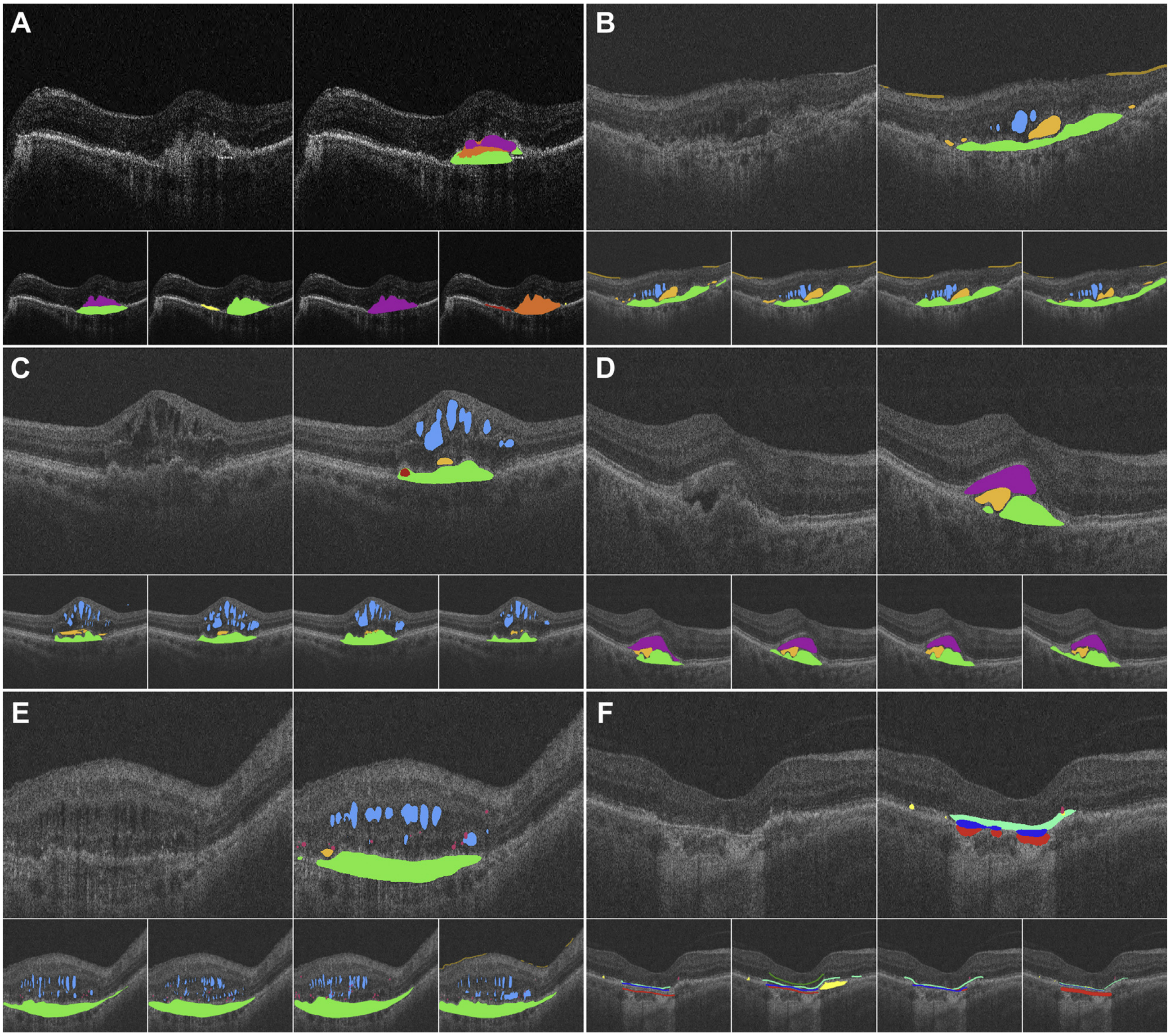

FIGURE 1. Comparison of model and grader output. Selection of B-scans from the test set, with the original B-scan (A), the output of the model (B), and the delineations of the 4 independent graders (E). For each subfigure, a subset of the features is shown to avoid clutter. (A). Showing fibrosis (orange), pigment epithelial detachment (PED; green), subretinal hyperreflective material (pink), drusen (red), and subretinal drusenoid deposits–reticular pseudodrusen (yellow). The graders give different interpretations to the lesion, while the model highlights multiple possibilities. (B). Showing intraretinal fluid (blue), subretinal fluid (SRF; orange), PED (green), and epiretinal membrane (brown). (C). Showing intraretinal fluid (blue), SRF (orange), PED (green), and drusen (red). (D). Showing SRF (orange), subretinal hyperreflective material (pink), and PED (green). (E). Showing intraretinal fluid (blue), SRF (orange), PED (green), hyperreflective dots (pink), and epiretinal membrane (brown). F. Showing subretinal drusenoid deposits–reticular pseudodrusen (yellow), hyperreflective dots (pink), outer plexiform layer descent (green), ellipsoid loss (teal), retinal pigment epithelium loss (blue), and hypertransmission (red).

horizontal or vertical size (rounded to a multiple of 256 voxels), and any number of B-scans (with a minimum of 9 B-scans). During training, patches of 9 × 512 × 512 voxels are used, requiring only the central B-scan to be fully annotated. At test time, the volumes are padded to 134 × 768 × 1024 voxels, and the model provides an output for each included feature for every voxel. Ambiguity of voxels that could represent multiple colocated features is resolved by using a sigmoid nonlinearity in the output layer. This allows the model to predict output probabilities that are not necessarily mutually exclusive: each input voxel can have a high probability for multiple features. This property can be used at inference time to discern voxels for which the output class is

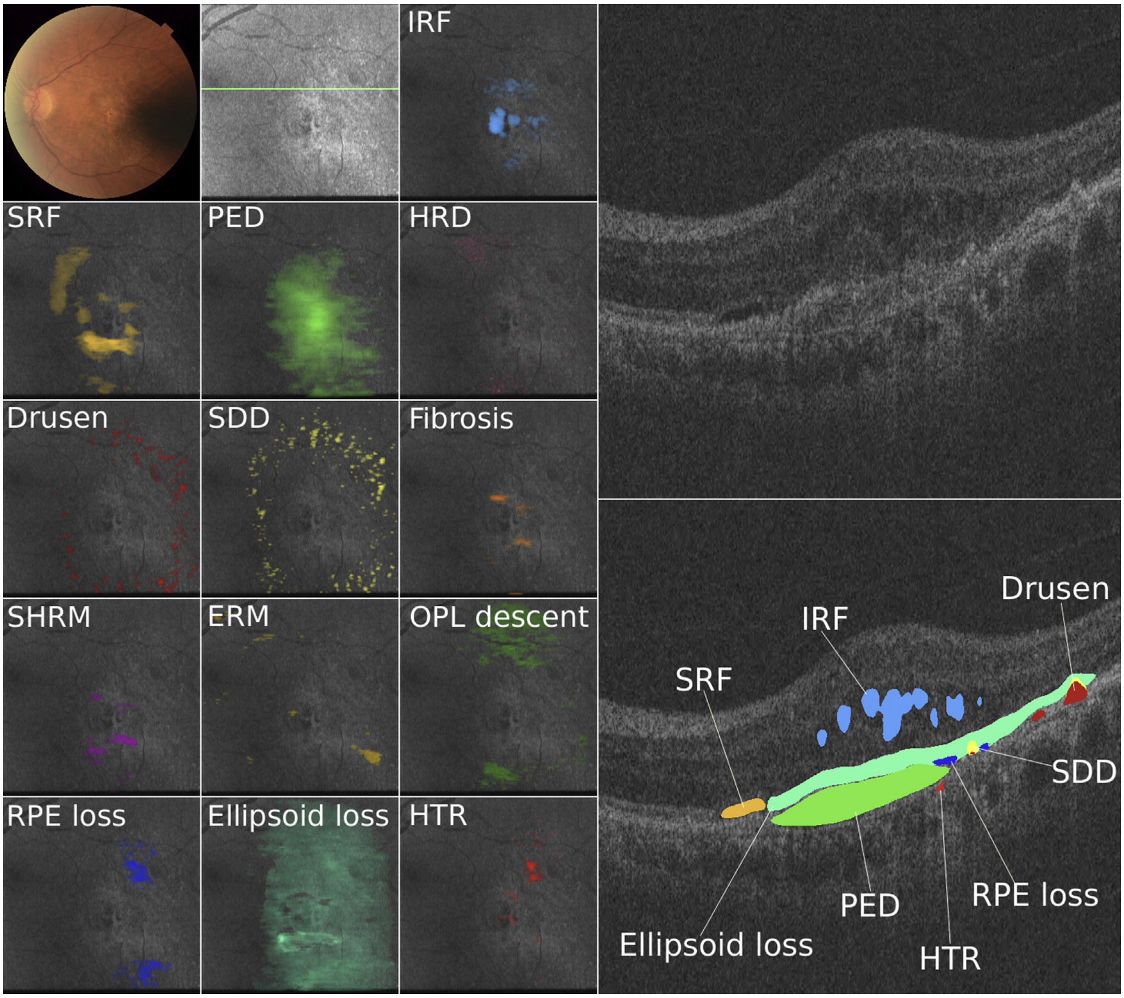

FIGURE 2. Example of the model output on a full optical coherence tomography (OCT) volume. The images on the left represent the color fundus image and en face projection of the optical coherence tomography volume, followed by 13 overlays representing the 13 segmented features. The brightness of the colors represents the number of segmented voxels per A-scan. (A) A single B-scan (indicated with a green line on the enface optical coherence tomography image) with overlays for all features. ERM = epiretinal membrane; HRD = hyperreflective dot; HTR = hypertransmission; IRF = intraretinal fluid; OPL = outer plexiform layer; PED = pigment epithelial detachment; RPE = retinal pigment epithelium; SDD = subretinal drusenoid deposit; SHRM = subretinal hyperreflective material; SRF = subretinal fluid.

ambiguous (eg, large drusen vs PED, or SHRM vs fibrosis).

The procedure for tuning the parameters of the model (training) consisted of 2 stages. The goal of the first stage was to get a model that is sensitive to each of the output features, by presenting it with a large variety of examples, using data augmentation. During the second stage, the focus was on false positive reduction and fine-tuning towards the expected presentation of data at test time.

During both stages, atrophy-related classes (OPL-descent, Ellipsoid loss, RPE-loss, and HTR) were treated separately. Because the axial (vertical) extent of the feature in the B-scan is not properly defined, and only the horizontal extent of the lesion is of interest, these were annotated as a line, roughly at the location where the feature was observed. During training, voxels in a region of 300 voxels above and below the annotated line were ignored (ie, no loss was calculated for those voxels). More implementation details can be found in the Supplemental Material.

The development data were split into 5 folds, of which 4 folds were used for training the model and 1 for monitoring its performance. By rotating the folds, 5 different models were obtained. An ensemble of the 5 models constituted the final segmentation model that was applied to the test set for performance evaluation. The ensemble was created by averaging the output of the individual models, after they were calibrated on their respective validation folds to resolve differences in sensitivity between the features and models.

TABLE 2. Dice and ICC Scores on the Test Set for Model and Observer

Feature (n)	Dice		ICC	
	Model	Observer	Model	Observer
Ellipsoid loss (930)	0.768 ± 0.005	0.714 ± 0.080	0.638 ± 0.029	0.444 ± 0.100
IRF (639)	0.637 ± 0.022	0.596 ± 0.048	0.873 ± 0.005	0.728 ± 0.124
PED (549)	0.838 ± 0.003	0.852 ± 0.007	0.943 ± 0.003	0.942 ± 0.013
SRF (406)	0.783 ± 0.007	0.828 ± 0.040	0.900 ± 0.013	0.915 ± 0.069
ERM (373)	0.705 ± 0.016	0.715 ± 0.086	0.772 ± 0.043	0.729 ± 0.097
HTR (289)	0.491 ± 0.053	0.517 ± 0.041	0.424 ± 0.076	0.443 ± 0.037
Drusen (265)	0.394 ± 0.026	0.491 ± 0.060	0.338 ± 0.036	0.536 ± 0.127
RPE loss (249)	0.471 ± 0.042	0.364 ± 0.096	0.381 ± 0.055	0.318 ± 0.114
SHRM (148)	0.540 ± 0.019	0.410 ± 0.132	0.685 ± 0.052	0.548 ± 0.269

ERM = epiretinal membrane; HTR = hypertransmission; IRF = intraretinal fluid; PED = pigment epithelial detachment; RPE = retinal pigment epithelium; SHRM = subretinal hyperreflective material; SRF = subretinal fluid.

Values represent means ± standard deviations for the 4 rotations of reference standard.

• **STATISTICAL ANALYSIS:** To measure overlap in segmented areas, we used the Dice similarity metric, which is defined as the size of the intersection of 2 areas divided by their average individual size. Therefore, a Dice-score of 1 indicates perfect agreement and a score of 0 indicates disjoint areas. For IRF, SRF, PED, SHRM, and drusen, overlap was calculated on the voxel level. For ellipsoid loss, HTR, RPE-loss, and ERM, only the lateral location (A-scan) of the feature was taken into account. That is, the output of the model was binarized at the optimal threshold, and A-scans with ≥1 voxel above the threshold were regarded as positives. Because not every feature is present in every B-scan in the test set, the Dice score is not always well-defined. Therefore, we calculated, for each feature, a single Dice score for the entire test set, rather than separately per B-scan. The Dice score was not regarded as an appropriate metric for SDD-RPD and HRD, because of their small and focal nature. Therefore, rather than measuring overlap on pixel level, we counted the number of detected/missed features within each B-scan and analyzed this using free-response receiver operating characteristic curves.

The intraclass correlation coefficient (ICC) for absolute agreement was used to measure agreement in size of the regions. The reference area was calculated as the average area of the 3 graders, and the fourth grader was used to estimate human performance. Cases with no segmentation were included as zero area.

Furthermore, by combining the atrophy-related features, it is possible to assess the model's performance in detecting and quantifying complete RPE and outer retinal atrophy (cRORA) as defined by the consensus definition for atrophy associated with AMD on OCT.[13] Following this definition, a B-scan is considered to contain cRORA if it contains 1) a region of HTR ≥250 μm in diameter; 2) a zone of attenuation or disruption of the RPE ≥250 μm in diameter; and 3) an area of ellipsoid zone loss. Following these criteria, we constructed the cRORA feature from the segmentation of RPE-loss, HTR, and ellipsoid-loss, for both the model and the graders. The performance of the model for the detection of cRORA at B-scan level is compared against a reference that requires the consensus of 3 out of 4 graders (≥75% agreement on the presence or absence of cRORA). Furthermore, the model's quantification of the extent of cRORA is compared directly with each grader's assessment.

RESULTS

QUALITATIVE RESULTS OF THE OUTPUT OF THE SEGMENTAtion model are shown in Figures 1 and 2. More results can be found in Supplemental Figures 1–3, and results can be explored interactively online (Supplemental Files Interactive1.html and Interative2.html).

Quantitative results for the Dice score and ICC can be found in Table 2, where we summarize, for each feature and for both the model and observer, the mean and standard deviation of the 4 metric scores obtained by rotating the reference. In addition, Bland–Altman analysis for each feature can be found in the Supplemental Material. Fibrosis and OPL descent were excluded from the evaluation because no reliable performance estimate could be made. This was because of the low numbers of annotated occurrences and large grader disagreement (for both features, there were no B-scans in the test set where the graders unanimously agreed on its presence). Averaged over the remaining features, the model obtained a Dice score of 0.63 ± 0.15 (median 0.64) compared with 0.61 ± 0.17 (median 0.60) for the observers. The average ICC

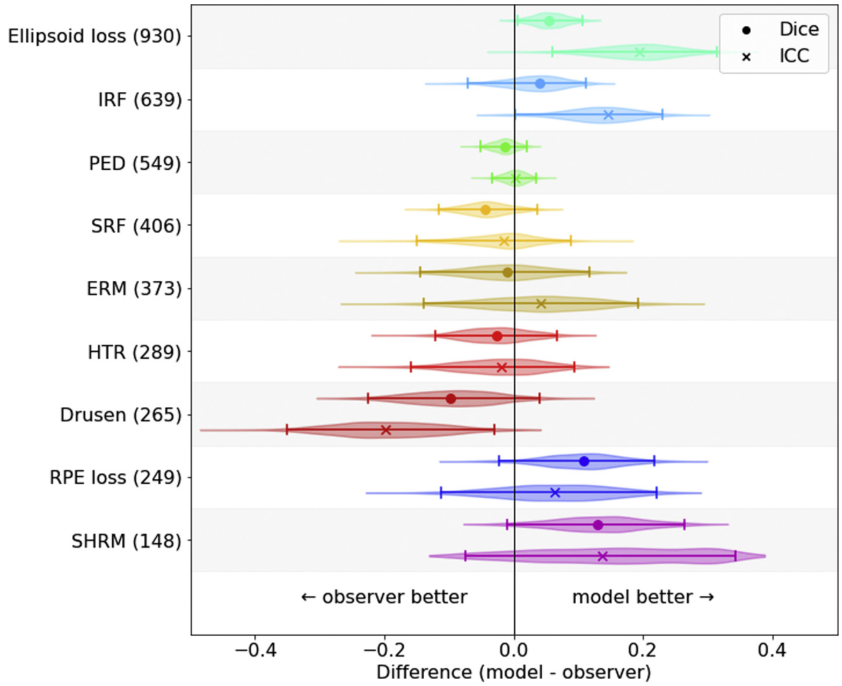

FIGURE 3. Difference in metric score between model and observer for each of the features. Per feature, the distribution of the differences (model minus observer) for Dice (A) and intraclass correlation coefficient (ICC) (B) are displayed. Positive values indicate that the model performs better than the observers. The vertical lines demarcate the 95% confidence intervals of the bootstrapped samples. The dot (Dice) or cross (ICC) represents the difference in actual metric on the full test set as summarized in Table 2. ERM = epiretinal membrane; HTR = hypertransmission; IRF = intraretinal fluid; PED = pigment epithelial detachment; RPE = retinal pigment epithelium; SHRM = subretinal hyperreflective material; SRF = subretinal fluid.

score for the model was 0.66 ± 0.22 (median 0.69) compared with 0.62 ± 0.21 (median 0.55) for the observers.

Differences in metric score between model and observer for each of the features are summarized in Figure 3, for both evaluation metrics. Per feature, the distribution of the differences (model minus observer) was obtained using bootstrapping (1000 bootstrap samples). Differences between model and observer Dice score were within a 95% confidence interval for all features except ellipsoid loss, where model performance was higher ($P = .03$). Regarding ICC, model performance was higher for IRF ($P = .04$) as well as ellipsoid loss ($P = .006$), lower for drusen ($P = .03$), and within the 95% confidence interval of the observer score for other features.

Performance for HRD and SDD-RPD are assessed using free-response receiver operating characteristic curves, which can be found in Figure 4. This figure highlights the sensitivity for both the observers and the model when operating at varying false positive rates, with confidence intervals obtained by bootstrapping (1000 bootstrap samples). For both HRD and SDD-RPD, the 95% confidence intervals of the model overlap with the confidence intervals for each grader. For HRD, the model obtains a sensitivity that is higher than each grader when operating at the same false positive rate.

Following the consensus of 3 out of 4 (≥75%) graders, cRORA was present in 7 of 112 B-scans and absent in 96 B-scans (leaving 9 B-scans ambiguous). The model detected cRORA in 13 B-scans, which comprised 6 of the 7 B-scans with cRORA (sensitivity 86%), and 3 of the 96 B-scans without cRORA (specificity 97%). Comparison of the extent (horizontal diameter within a B-scan) of the cRORA for the model and graders is shown in Figure 5, which shows there is considerable variability in assessing diameter of cRORA.

DISCUSSION

WE PRESENT A DEEP LEARNING MODEL FOR SEGMENTATION of 13 features commonly found in neovascular and atrophic AMD that was developed and validated on a large real-world dataset. The model's performance is comparable to, and for some features possibly better than, independent observers, both in terms of overlap and correlation between segmented area and reference area. By comparing against the combined output of multiple observers, we were able to set a reference standard that is more reliable than that of a single observer. Moreover, by rotating the reference standard, the performance of the model was not just

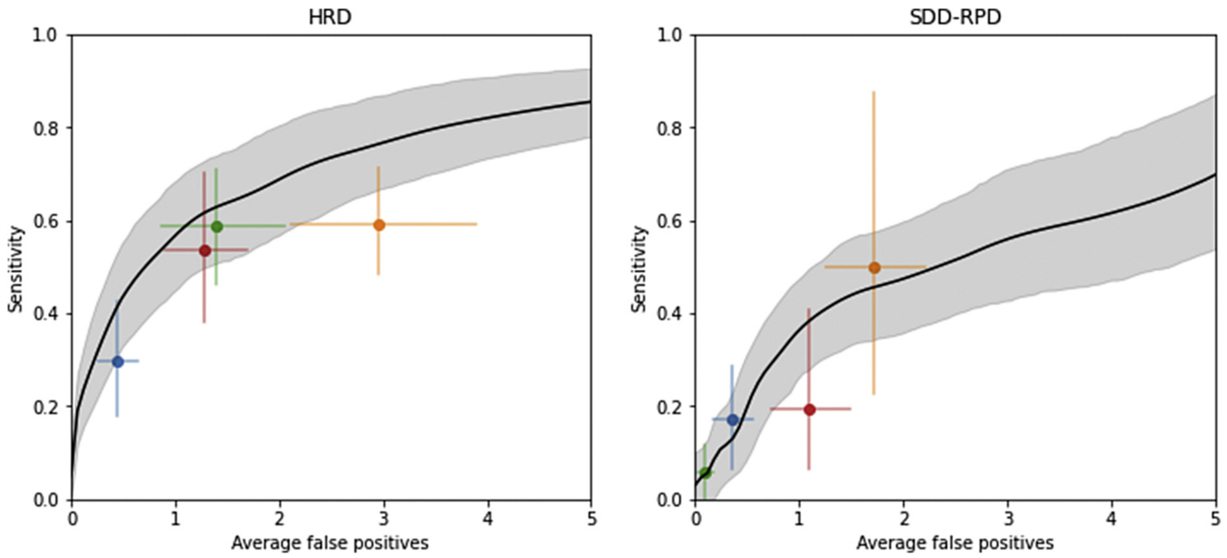

FIGURE 4. Free-response receiver operating characteristic curves for hyperreflective dots (HRDs) and exudates and subretinal drusenoid deposits–reticular pseudodrusen (SDD-RPD). The line represents model sensitivity at different thresholds, with the shaded area representing the 95% confidence interval, obtained by bootstrapping. The dots represent the 4 graders, with error bars representing 95% confidence intervals.

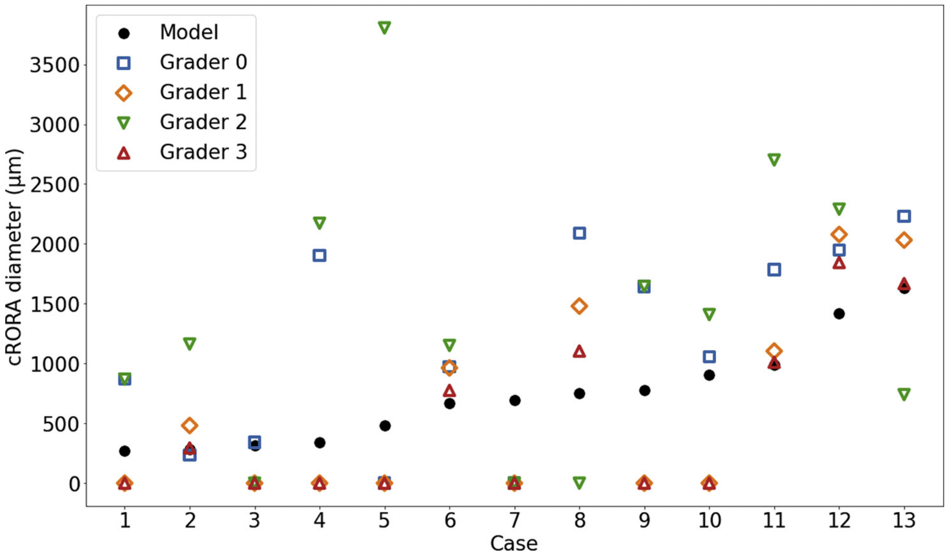

FIGURE 5. Extent of complete retinal pigment epithelium and outer retinal atrophy (cRORA) on the 13 cases with cRORA detected by the model, ordered by their diameter. The diameter of cRORA is obtained from the diameter of colocated retinal pigment epithelium loss and hypertransmission, conditioned on the presence of any ellipsoid loss in the same B-scan. The black dots represent the diameter of the model and the colored shapes represent the diameters obtained by the 4 graders for the same cases.

compared against a single observer who might, for example, grade more conservatively than others. Eight independent graders contributed to the data for model development, which reduced the risk that the subjective opinion of a specific grader transferred to the model. However, because the graders that annotated the test set also contributed to the development data, future external validation of the performance of the presented model may still be required for some applications.

The set of included features in this study is larger than that of most previous work on segmentation in OCT, which predominantly focused on fluid.[14,24,25] The inclusion of separate constituent features of atrophy (ellipsoid loss, RPE-loss, and HTR) is another unique aspect of this

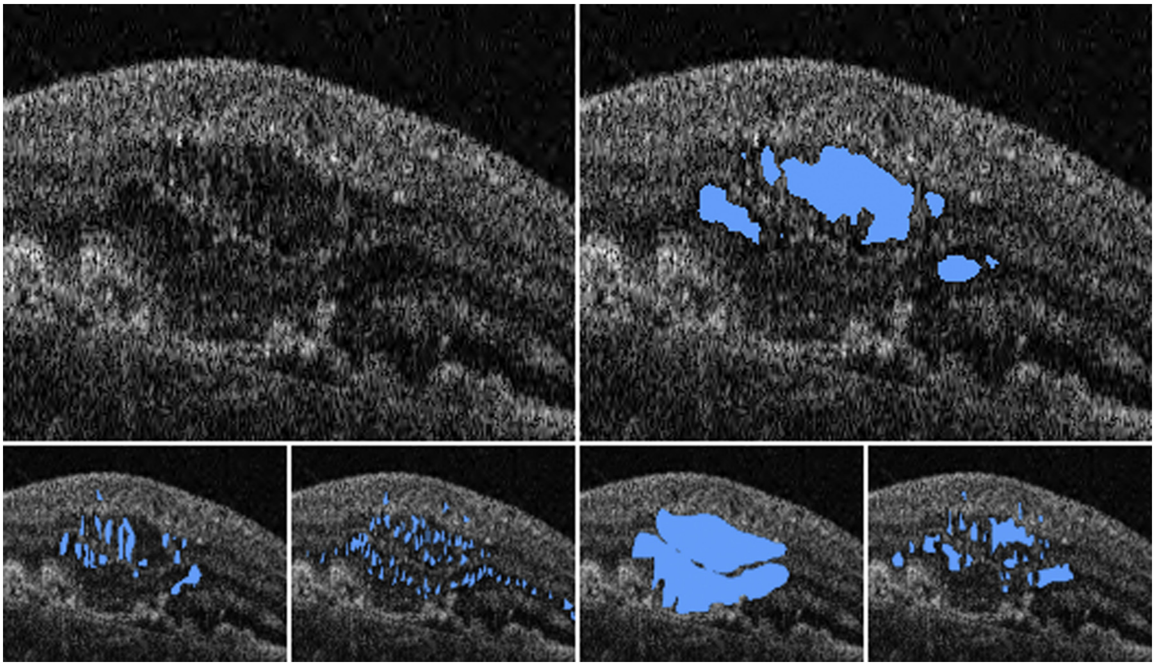

FIGURE 6. Example of segmentation of intraretinal fluid for the model (A), compared to the 4 graders (B). The overlap between the output of the model and each grader is higher on average than the average overlap between each pair of graders.

study, distinguishing it from other models that include a large set of features.[18,19] Direct comparison of performance between models is still challenging because not all models are evaluated using the same metrics, and there is variation in the nature of the data that are used, for example in OCT vendors, included diseases, or annotated features.

In our evaluation, there is a relatively large variability in performance between features, both in terms of Dice score and ICC, with model scores (Dice/ICC) ranging from 0.394/0.338 for drusen to 0.838/0.943 for PED (Table 2). It can be observed from this table that the standard deviation for the model is generally lower than for the observer. This is an artefact of the evaluation scheme where the reference and observers are rotated. The standard deviation for the observer can be interpreted as a measure for the variability in grader agreement (ie, agreement between graders can depend on the subset of graders included in the reference).

Although levels of agreement vary greatly between features, model and observer scores generally follow the same pattern. Therefore, to put model performance in perspective, it is more informative to look at the difference with the observer rather than the absolute metric value. These differences are shown in Figures 3 and 4. The performance of the model appears to exceed that of the graders for some of the features, such as IRF, ellipsoid loss, and HRD. This could be considered remarkable, because the model was trained on data that were generated by the same graders. We identify 3 settings in which the segmentation of the model could outperform that of a human grader. First, the model could have converged to produce results of an average grader, which is closer to either grader separately than graders between themselves (see for example Figure 6). Second, graders might have overlooked small features, such as HRD, leading to lower sensitivity. Third, some of the differences between graders are related to inaccurate delineation of lesion borders, while the output of the model naturally and smoothly follows lesion boundaries.

Performance of the model is slightly below observers' performance for drusen and drusenoid PED. Although drusen are important features related to progression from intermediate to advanced AMD, they are less relevant once the neovascular stage has set in. Therefore, implications of the lower performance for the purpose of this study are limited. A possible explanation of the lower performance is a relative lack of representative training data, as the training data were selected based on eyes receiving treatment with anti-VEGF. The presumption that more training data will improve model performance is further corroborated by the excellent performance of the model for IRF and ellipsoid loss, which were the 2 features with most samples in the training data. Another consideration is that larger variability in agreement between observers could also result in favorable performance for the model. Implementation of artificial intelligence could therefore provide a solution to resolve inherent discrepancies caused by subjective human assessment.

Plausibility of the models' output in ambiguous regions is highlighted in some examples in Figure 7. Although all

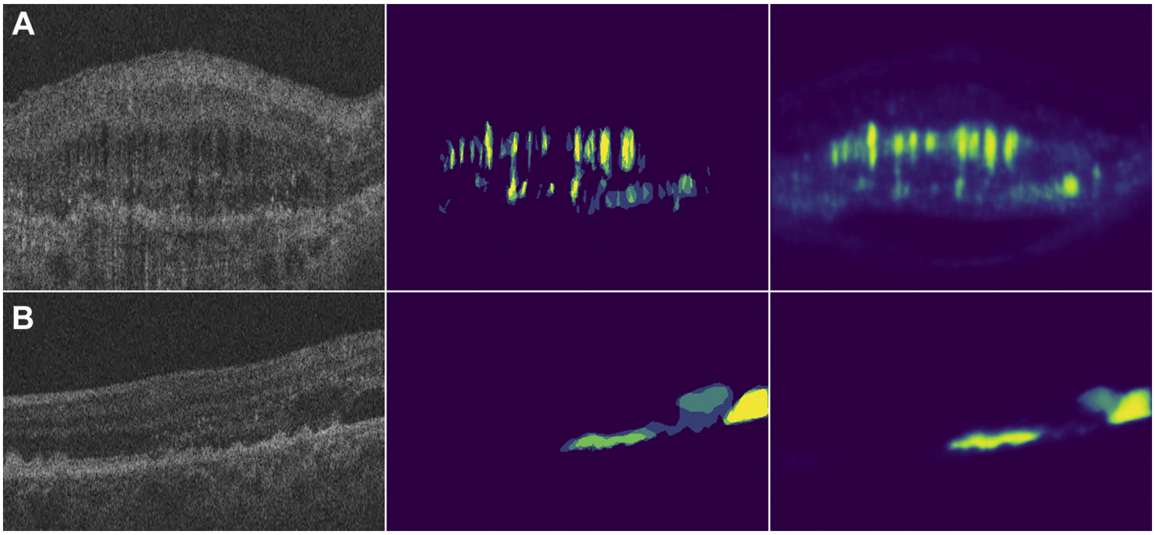

FIGURE 7. Plausibility of model's output based on the grade of agreement between observers, shown for intraretinal fluid (A) and subretinal fluid (B). On the left is the original image; in the center, the heatmap of the graders, with brighter indicating a higher agreement among graders (more graders annotated the voxel); on the right the output of the model, with brighter indicating higher likelihood to belong to the feature as estimated by the model. Assigned likelihood of the model correlates with the agreement of the graders.

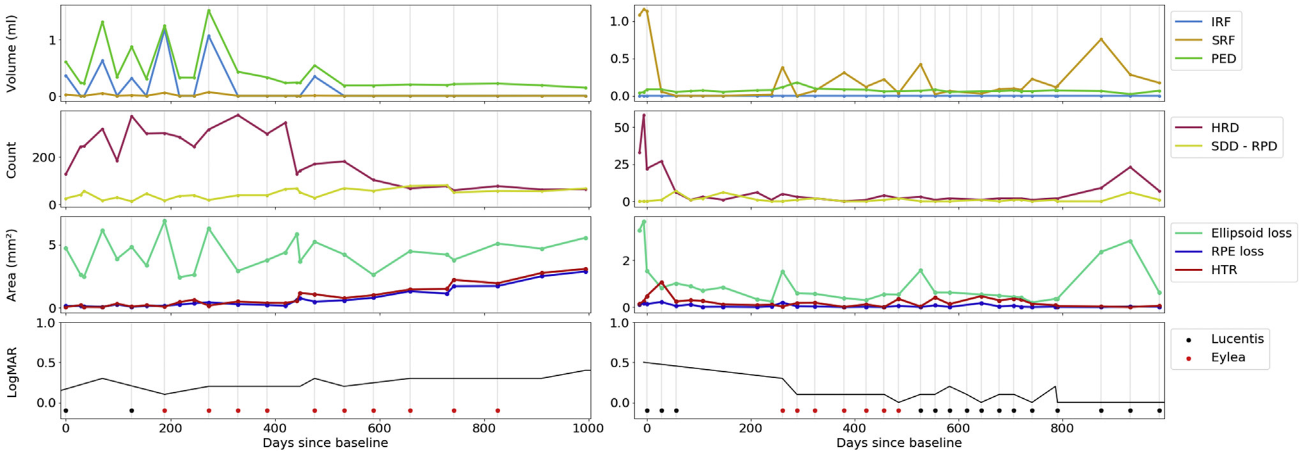

FIGURE 8. Fully automatically generated overview of personal disease history, showing quantified morphologic parameters, visual acuity (black line), and treatment history (vertical lines and black and red dots). (A) A case with initially recurring fluid followed by onset of atrophy. (B) A case with incidental recurrence of subretinal fluid and enduring good visual function. HRD = hyperreflective dot; HTR = hypertransmission; IRF = intraretinal fluid; PED = pigment epithelial detachment; RPE = retinal pigment epithelium; SDD-RPD = subretinal drusenoid deposits–reticular pseudodrusen; SHRM = subretinal hyperreflective material; SRF = subretinal fluid.

quantitative analyses in this study are carried out on binary images, the actual output of the model represents a real-valued probability estimate of presence of each feature for every voxel. This probability estimate correlates with the agreement of the graders: the model assigns higher scores to voxels with univocal grader labels. These scores can be used to select an operating point for the model at various levels of sensitivity, which can be tuned depending on the application.

Although there are numerous potential applications, this study does not provide a direct validation against a clinically relevant outcome. As an example of a practical application, the model could generate reproducible results for improved reporting and decision making when integrated in the current clinical care. Examples of fully automatically generated personal disease history, obtained by applying the model to all available OCT volumes for patients receiving anti-VEGF treatment, can be found in

Figure 8. This visualization of changes in retinal morphology could provide the treating clinician with valuable information in guiding treatment decisions, for example in determining treatment intervals or when considering switching to different treatments or guiding entry into clinical trials. Besides the added value of automatically estimated parameters in clinical routine, automatic segmentation of retinal morphology could constitute a pivotal role in the study of underlying disease mechanisms and the identification of targets for new therapeutic strategies.[9,21]

A limitation of this study is that the model has been validated only on neovascular AMD. However, training data included other diseases in which anti-VEGF therapy is used, such as diabetic macular edema or retinal vein occlusion. Validation on these diseases is left for future work. Moreover, the model was developed and validated specifically on Topcon OCT volumes. Application of the model to other OCT vendors has not been explored but would likely require more data and minor adaptations in methodology.[18]

Another possible limitation, when applying the presented model on large-scale real-world data, without any human supervision, is that the model is not able to detect possible scanning artefacts or poor image quality. These factors could have a negative impact on the reliability of the produced quantified parameters. Furthermore, the model does not produce quantified results of the location of features with respect to the foveal center, which would allow for more specific summarizing of volumetric measures and improved diagnostics. We are currently investigating how the model can be augmented to also provide an estimate for scan quality and the location of the fovea within the scan.

In conclusion, we present a fully automatic segmentation model for 13 features related to neovascular AMD that performs at the level of experienced graders. The application of this model will open numerous new opportunities for study of morphologic retinal changes and treatment efficacy in real-world settings. Furthermore, it can facilitate structured reporting in the clinic, which will reduce subjectivity in clinicians' assessments and enable implementation of refined treatment guidelines. This could ultimately lead to increased speed of interpretation, a reduction of cost, and improved personalized care.

ALL AUTHORS HAVE COMPLETED AND SUBMITTED THE ICMJE FORM FOR DISCLOSURE OF POTENTIAL CONFLICTS OF INTEREST. Funding/Support: Supported by a grant from Novartis Pharmaceuticals. Funding was obtained from the Automation in Medical Imaging project, a collaborative project of the Fraunhofer Gesellschaft and the Radboud University and University Medical Center. This research was funded by German Research Foundation grant no. MU4279/2-1 (to P.L.M.). The sponsor or funding organization had no role in the design or conduct of this research. Financial Disclosures: A.N.W. is supported by the UCL Wellcome Trust PhD Programme for Clinicians. R.S. received travel expenses from Allergan and is a board member of Eye-Wise Diagnostics. K.B. received grants from Roche and Bayer and speaker fees from Novartis, Bayer, Topcon, and Heidelberg Engineering. All authors attest that they meet the current ICMJE criteria for authorship.

REFERENCES

1. Coleman HR, Chan C, Ferris FL III, et al. Age-related macular degeneration. *Lancet* 2008;372:1835–1845.
2. Rosenfeld PJ, Brown DM, Heier JS, et al. Ranibizumab for neovascular age-related macular degeneration. *N Engl J Med* 2006;355:1419–1431.
3. Colijn JM, Buitendijk GH, Prokofyeva E, et al. Prevalence of age-related macular degeneration in Europe: the past and the future. *Ophthalmology* 2017;124:1753–1763.
4. Wong WL, Su X, Li X, et al. Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *Lancet Glob Health* 2014;2:e106–e116.
5. Martin DF, Maguire MG, Fine SL, et al. Ranibizumab and bevacizumab for treatment of neovascular age-related macular degeneration: two-year results. *Ophthalmology* 2012;119:1388–1398.
6. Holz FG, Amoaku W, Donate J, et al. Safety and efficacy of a flexible dosing regimen of ranibizumab in neovascular age-related macular degeneration: the sustain study. *Ophthalmology* 2011;118:663–671.
7. Mitchell P, Korobelnik J, Lanzetta P, et al. Ranibizumab (lucentis) in neovascular age-related macular degeneration: evidence from clinical trials. *Br J Ophthalmol* 2010;94:2–13.
8. Keane PA, Patel PJ, Liakopoulos S, et al. Evaluation of age-related macular degeneration with optical coherence tomography. *Surv Ophthalmol* 2012;57:389–414.
9. Schmidt-Erfurth U, Waldstein SM. A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration. *Prog Retin Eye Res* 2016;50:1–24.
10. Schmidt-Erfurth U, Klimscha S, Waldstein S, et al. A view of the current and future role of optical coherence tomography in the management of age-related macular degeneration. *Eye* 2017;31:26–44.
11. Waldstein SM, Philip A, Leitner R, et al. Correlation of 3-dimensionally quantified intraretinal and subretinal fluid with visual acuity in neovascular age-related macular degeneration. *JAMA Ophthalmol* 2016;134:182–190.
12. Arnold JJ, Markey CM, Kurstjens NP, et al. The role of subretinal fluid in determining treatment outcomes in patients with neovascular age-related macular degeneration - a phase IV randomised clinical trial with ranibizumab: the FLUID study. *BMC Ophthalmol* 2016;16:31.
13. Sadda SR, Guymer R, Holz FG, et al. Consensus definition for atrophy associated with age-related macular degeneration on OCT: classification of atrophy report 3. *Ophthalmology* 2018;125:537–548.
14. Bogunović H, Venhuizen F, Klimscha S, et al. RETOUCH: the retinal OCT fluid detection and segmentation

benchmark and challenge. *IEEE Trans Med Imaging* 2019;38:1858–1874.

15. Toth CA, Decroos FC, Ying G, et al. Identification of fluid on optical coherence tomography by treating ophthalmologists versus a reading center in the comparison of age-related macular degeneration treatments trials (CATT). *Retina* 2015;35:1303.
16. 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* 2016;316:2402–2410.
17. Burlina PM, Joshi N, Pekala M, et al. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol* 2017;135:1170–1176.
18. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018;24:1342–1350.
19. Yim J, Chopra R, Spitz T, et al. Predicting conversion to wet age-related macular degeneration using deep learning. *Nat Med* 2020;26:892–899.
20. Schmidt-Erfurth U, Bogunovic H, Sadeghipour A, et al. Machine learning to analyze the prognostic value of current imaging biomarkers in neovascular age-related macular degeneration. *Ophthalmol Ret* 2018;2:24–30.
21. Schmidt-Erfurth U, Vogl W, Merrill Jampol L, et al. Application of automated quantification of fluid volumes to anti–VEGF therapy of neovascular age-related macular degeneration. *Ophthalmology* 2020;127:1211–1219.
22. Chen X, Niemeijer M, Zhang L, et al. Three-dimensional segmentation of fluid-associated abnormalities in retinal OCT: probability constrained graph-search-graph-cut. *IEEE Trans Med Imaging* 2012;31:1521–1531.
23. Roy AG, Conjeti S, Karri SPK, et al. ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. *Biomed Opt Express* 2017;8:3627–3642.
24. Lee H, Kang KE, Chung H, et al. Automated segmentation of lesions including subretinal hyperreflective material in neovascular age-related macular degeneration. *Am J Ophthalmol* 2018;191:64–75.
25. Venhuizen FG, van Ginneken B, Liefers B, et al. Deep learning approach for the detection and quantification of intraretinal cystoid fluid in multivendor optical coherence tomography. *Biomed Opt Express* 2018;9:1545–1569.
26. Amoaku W, Chakravarthy U, Gale R, et al. Defining response to anti-VEGF therapies in neovascular AMD. *Eye* 2015;29:721–731.