Taylor & Francis
Taylor & Francis Group

Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/taut20

Annotated retinal optical coherence tomography images (AROI) database for joint retinal layer and fluid segmentation

Martina Melinščak, Marin Radmilović, Zoran Vatavuk & Sven Lončarić

To cite this article: Martina Melinščak, Marin Radmilović, Zoran Vatavuk & Sven Lončarić (2021) Annotated retinal optical coherence tomography images (AROI) database for joint retinal layer and fluid segmentation, Automatika, 62:3-4, 375-385, DOI: 10.1080/00051144.2021.1973298

To link to this article: https://doi.org/10.1080/00051144.2021.1973298

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 31 Aug 2021.

Submit your article to this journal

Article views: 775

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=taut20

AUTOMATIKA
2021, VOL. 62, NOS. 3–4, 375–385
https://doi.org/10.1080/00051144.2021.1973298

OPEN ACCESS Check for updates

Annotated retinal optical coherence tomography images (AROI) database for joint retinal layer and fluid segmentation

Martina Melinščak [a,b], Marin Radmilović[c], Zoran Vatavuk[c] and Sven Lončarić[b]

[a]Department of Mechanical Engineering, Karlovac University of Applied Sciences, Karlovac, Croatia; [b]Department of Electronic Systems and Information Processing, Faculty of Electrical Engineering and Computing, Zagreb, Croatia; [c]Department of Ophthalmology, Sestre milosrdnice University Hospital Center, Zagreb, Croatia

ABSTRACT

Optical coherence tomography (OCT) images of the retina provide a structural representation and give an insight into the pathological changes present in age-related macular degeneration (AMD). Due to the three-dimensionality and complexity of the images, manual analysis of pathological features is difficult, time-consuming, and prone to subjectivity. Computer analysis of 3D OCT images is necessary to enable automated quantitative measuring of the features, objectively and repeatedly. As supervised and semi-supervised learning-based automatic segmentation depends on the training data and quality of annotations, we have created a new database of annotated retinal OCT images – the AROI database. It consists of 1136 images with annotations for pathological changes (fluid accumulation and related findings) and basic structures (layers) in patients with AMD. Inter- and intra-observer errors have been calculated in order to enable the validation of developed algorithms in relation to human variability. Also, we have performed the automatic segmentation with standard U-net architecture and two state-of-the-art architectures for medical image segmentation to set a baseline for further algorithm development and to get insight into challenges for automatic segmentation. To facilitate and encourage further research in the field, we have made the AROI database openly available.

ARTICLE HISTORY
Received 28 February 2021
Accepted 23 August 2021

KEYWORDS
Annotated retinal OCT images; images database; automatic image segmentation; deep learning; age-related macular degeneration

Introduction

Age-related macular degeneration (AMD) is an acquired degeneration of the retina that causes significant central visual impairment in its late-stage and is the leading cause of irreversible blindness in people 50 years of age or older in the developed world [1]. The estimated prevalence of any and late-stage AMD in people aged 45–85 years is 8.69% and 0.37%, respectively [2]. The prevalence of late-stage AMD sharply rises to 7.1% in those 75 years or older [3]. Due to the aging of the population, by 2040 an estimated 288 million people will be affected by AMD [2].

AMD is a progressive disease. The early and intermediate stages are usually asymptomatic, characterized by the accumulation of yellow granular deposits beneath the outermost layer of the retina, the retinal pigment epithelium (RPE) (Figure 1(a)). Advanced or late AMD is defined either by the development of atrophy of the RPE and the overlying photoreceptors or by the development of new blood vessels (neovascular membranes) beneath or above the RPE (Figure 1(b)). These new vessels tend to leak or rupture, with subsequent exudation or haemorrhage accumulating in different retinal layers [4]. These two forms of advanced AMD, the former usually referred to as geographic atrophy (GA), and the latter as exudative, wet, or neovascular AMD (nAMD), can occur alone or together, either simultaneously or sequentially, and both forms can lead to significant visual impairment [5]. There is currently no approved treatment available for GA [6]. Intravitreal anti-vascular endothelial growth factor (anti-VEGF) therapy is the mainstay of nAMD treatment [7].

Optical coherence tomography (OCT) is an imaging technique invaluable for diagnosing AMD and guiding AMD treatment because it provides high-resolution, pseudohistological cross-sectional images of the retina and choroid. In nAMD, OCT is used to detect and visualize specific lesions such as intraretinal fluid (IRF), subretinal fluid (SRF), subretinal hyperreflective material (SRHM), and retinal pigment epithelial detachment (PED). These changes represent exudates (IRF, SRF, some cases of SRHM, some forms of PED), haemorrhage (some cases of SRHM, some forms of PED), or neovascular membranes and fibrosis (some cases of SRHM and some forms of PED) [8]. Previous studies have shown these lesions can serve as OCT biomarkers for the visual function or therapy response [9–14]. In comparison to two-dimensional analyses of central B-scans, volumetric analyses of these pathologic lesions in the entire macular area might more precisely predict these outcomes [13,14].

CONTACT Martina Melinščak martina.melinscak@gmail.com Karlovac University of Applied Sciences, Karlovac 47000, Croatia

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

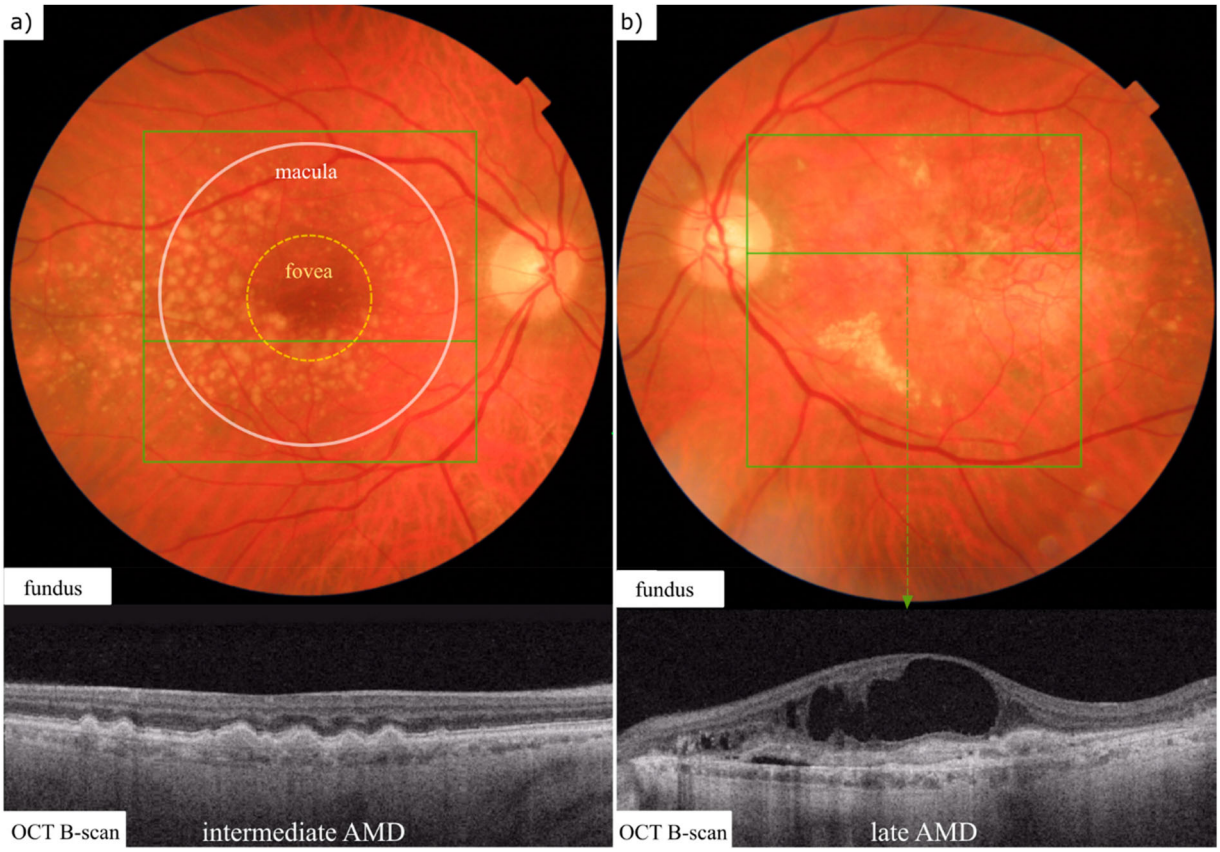

Figure 1. (a) Intermediate AMD. Fundus photography showing numerous yellow granular deposits (drusen) in the macular area. OCT showing the location of drusen beneath the RPE. (b) Late (neovascular) AMD. Fundus photography showing new subretinal vessels and a wide area of exudation. OCT showing extensive structural changes in multiple retinal layers.

Automatic segmentation of retinal layers and of the IRF, SRF, SRHM, and PED (from this point on altogether referred to as fluids) is crucial for detecting and characterization of AMD objectively and in a reproducible way. A reliable automatic OCT system for segmentation is crucial for further development of diagnosing retinal disease. Due to the quantitative analysis of pathological changes, the therapy effectiveness could be predicted [15,16]. The occurrence of intra- and subretinal fluid is an important biomarker that plays a major role in (re-) treatment decisions and is prognostic of visual rehabilitation [17]. Quantitative analysis also allows the prediction of the transition from the middle to the late phase and the prediction of whether in the case of one diseased eye, the disease will affect the other eye [18,19]. Recent studies confirm the correlation between individual fluids and the success of anti-VEGF therapy, making computer segmentation and quantitative analysis even more important [20–22].

Currently, commercially available OCT software algorithms automatically perform retinal layer segmentation to a variable extent and derive basic parameters such as inner or outer retinal thickness, RPE elevation, or central subfield thickness. Although helpful and routinely used to inform clinical decisions, issues of susceptibility to segmentation errors and limited inter-device reproducibility have been raised [23,24]. However, the main issue of these algorithms is the lack of detailed information, as these basic thickness and elevation parameters do not distinguish specific underlying lesions (e.g. retinal fluids in different compartments).

Despite great progress in computer vision and medical image segmentation, a major shortcoming is the lack of publicly available databases of annotated images. In medical image segmentation supervised or semi-supervised methods are still predominant, and their accuracy depends on the quality and scale of annotated data. Most of the published methods are developed on the datasets of images in which no significant pathological changes are present, so it is questionable to what extent the methods are applicable in the case of severe disease. Also, it is difficult to evaluate and compare different methods as there is a lack of publicly available databases of manually labelled images, and results depend significantly on the number of processed images, quality of images (related to the type of OCT device), present disease, etc.

In 2017, MICCAI "Retinal OCT Fluid Challenge (RETOUCH)" was organized [25]. Considering the same dataset and the same metric for evaluation of segmentation and detection, this was a notable improvement in the evaluation of different methods. All applied methods (eight teams were participating) for automatic

segmentation and detection of fluids were based on deep learning methods (segmentation was usually performed with the popular U-net architecture [26]) in combination with other machine learning (ML) methods and image analysis methods. Details can be found on the challenge website [25] and in the accompanying paper [27]. Besides the RETOUCH challenge database, the only publicly available dataset, which some authors use for evaluation of their methods, is the DUKE dataset [28] containing 110 B-scans of 10 DME (diabetic macular oedema) patients acquired with the Spectralis OCT. In a recent review paper, Khan et al. [29] gave an extensive overview of publicly available databases in ophthalmology (up to May 2020) in which authors raise serious concerns about "data poverty" and argue about challenges in data collection.

In this paper, we describe the creation of the openly available **A**nnotated **R**etinal **O**CT **I**mages (AROI) database. We give an overview and analysis of the current state of development of the AROI database, with plans for further improvements and new features. It currently consists of 1136 annotated B-scans (from 24 patients suffering from nAMD) and associated raw high-resolution images. The existing research indicates that the approach in which layers and fluids are jointly segmented aims to take advantage of the interdependence of fluids and layers and thus achieves the best possible segmentation [30–35]. Therefore, we have provided annotations for pathological changes (fluid accumulation and related findings) and basic structures (layers). Also, we have presented results for intra- and inter-observer errors to enable the validation of developed algorithms in relation to human variability. To set a baseline for deep learning (DL) methods we have presented results for automatic segmentation with standard U-net architecture and two state-of-the-art architectures (U-net-like and U-net++) in medical image segmentation. The results indicate that there are major challenges for automatic segmentation in the case of severe pathologies. To the best of our knowledge, there are no such publicly available datasets in terms of scale, and with such exhaustive annotations in patients with severe pathology and still, they are crucial for the introduction of automatic segmentation into clinical practice.

Materials and methods

AROI database

Data collection

For the purpose of this study, we collected the database of manually Annotated Retinal OCT Images (AROI database). In collaboration with Sestre milosrdnice University Hospital Center (Zagreb, Croatia), images were collected and annotated by an ophthalmologist. Selection criteria included patients aged 60 years and older diagnosed with nAMD, irrespective of their anti-vascular endothelial growth factor (anti-VEGF) therapy status, with no significant media opacities precluding adequate retinal imaging, and with no other retinal disorders. The concurrent presence of geographic atrophy (GA) of any extent was not an exclusion criterion, since nAMD and GA occur simultaneously or sequentially in a significant number of patients with advanced AMD. From April 2018 to June 2018, 24 consecutive patients were included in the study. Macular SD-OCT volumes were recorded with the Zeiss Cirrus HD OCT 4000 device. Image quality was checked by an ophthalmologist. Overall signal strength of 6/10 or more and absence of any focal shadow artefacts or out of register artefacts was a prerequisite for further analysis. Each OCT volume consisted of 128 B-scans, spaced 47.24 μm apart, with a resolution of 1024×512 pixels (pixel size 1.96 μm $\times$ 11.74 μm). Retinal fluids and layers were annotated for 1136 B-scans out of a total of 3072 B-scans for 24 patients (37% of B-scans were annotated). Annotations were not provided for each B-scan. The central 10 B-scans around the foveal centre were annotated for each patient, as visual acuity mostly depends on the pathological changes in this area, while more eccentric B-scans were annotated at the ophthalmologist's discretion: in case the adjacent B-scans were deemed similar, the annotations were skipped (performed for every 2nd to 10th scan, depending on the extent and complexity of pathological changes). The average number of annotated scans per patient was 47.3 ± 25.7.

Images within the database are available in PNG format and organized so that each filename is associated with a patient number (1–24) and B-scan ordinal number (0–127) (e.g. patient12_123.png). All raw images, and not just labelled ones, are available to enable 3D automatic segmentation. At this phase, images are not divided into training and test sets.

Data collection adhered to the tenets of the Declaration of Helsinki and the standards of Good Scientific Practice of Sestre milosrdnice University Hospital Center (Zagreb, Croatia). All patients signed informed consent, the images are anonymized and do not contain any additional information about patients. The presented study was approved by the Ethics Committee of the Sestre milosrdnice University Hospital Center (EP-3272/18-11) and the Faculty of Electrical Engineering and Computing (Zagreb, Croatia).

Fluids and layers annotations

All annotations were done by one expert ophthalmologist. The lesions of interest and annotated in this study were the intraretinal fluid (IRF), subretinal fluid (SRF), subretinal hyperreflective material (SRHM), and retinal pigment epithelial detachment (PED) (Figure 2). Although some of these lesions can reflect processes such as fibrotic scars or fibrovascular membranes, as

they most often reflect exudation or haemorrhage, the term "fluids" was liberally used here to refer to all these lesions. Additionally, as SRF and SRHM share the same location characteristics, and often reflect the same exudative process, the distinction between SRF and SRHM was not performed in this study, and they were annotated jointly as SRF. In the future, development of the enhanced segmentation method is planned in order to discriminate between SRF and SRHM, as well as between different types of SRHM. As the range of SRHM in AMD and other macular disorders includes a number of different lesions, such as neovascular membranes, fibrosis, exudate, haemorrhage, and lipofuscin like material, a new, more extensive dataset supported with multimodal imaging is in preparation.

To detect IRF, SRF, and PED, the knowledge of their location within or outside specific retinal layers can be used to facilitate their detection and differentiation. The retina is histologically divided into 10 layers: (1) internal limiting membrane (ILM), (2) retinal nerve fibre layer (RNFL), (3) ganglion cell layer (GCL), (4) inner plexiform layer (IPL), (5) inner nuclear layer (INL), (6) outer plexiform layer (OPL), (7) outer nuclear layer (ONL), (8) external or outer limiting membrane (ELM or OLM), (9) photoreceptor layer, and (10) retinal pigment epithelium/Bruch's membrane complex (RPE/BM). The layers visible in the OCT scan have been correlated to these histological layers, with the exception of a few additional zones observed in the photoreceptor layer for which the exact histological counterpart is not yet defined [36].

The IRF is a localized extracellular intraretinal fluid accumulation seen as a hyporeflective area located anywhere between ILM and ELM, the SRF is a subretinal fluid accumulation seen as a hyporeflective area between photoreceptor layer and RPE (the SRHM is seen in the same location but as a homogenously or inhomogenously hyperreflective area), and the PED is either a hyporeflective accumulation of fluid or hyperreflective accumulation of other material between RPE and BM (the RPE/BM complex is separated in the case of PED) [8]. However, as pathological processes such as those in AMD can lead to extensive changes in retinal structure, it is often impossible for a reader to determine all the layers in the OCT scan reliably. Therefore, we traced ILM, the inner boundary of RPE, BM, and the boundary between IPL and INL (IPL/INL), as these layers/boundaries could be readily determined in virtually all images (Figure 2). While ILM, the inner boundary of RPE, and BM were chosen as pertinent for IRF, SRF, and PED localization, the IPL/INL boundary was chosen as it could be used to locate the foveal centre in case of an eccentric scan (in patients with poor fixation) or loss of normal foveal depression (in patients with extensive foveal oedema, elevation, traction, or parafoveal atrophy). As the foveal centre consists only of ILM, ONL, photoreceptor layer, and RPE/BM, while other layers taper towards the foveal centre, the centre can be defined as the point of least distance between the ILM and any of these other layers, including the IPL/INL boundary. In our dataset all scans were centred at the fovea and this step was not needed.

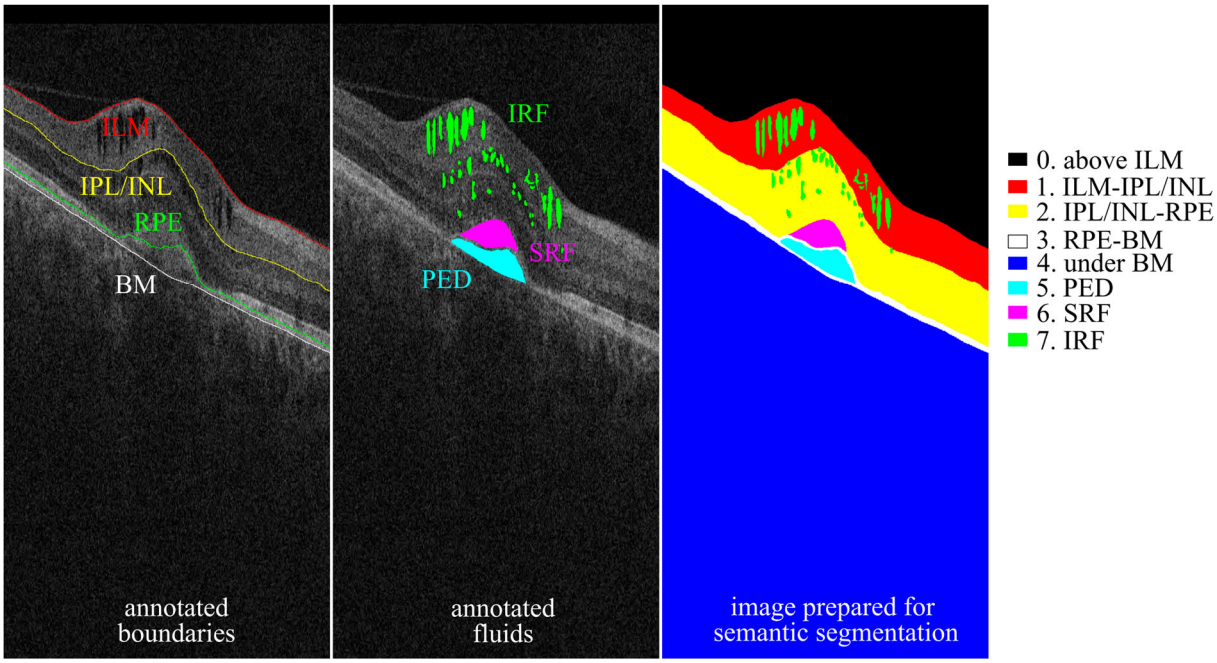

Figure 2. From left to right: example of the image with annotated boundaries, an image with annotated fluids, and an image prepared for semantic segmentation (with eight classes).

Images prepared for semantic segmentation

As simultaneous segmentation of layers and fluids should give better results than separate segmentation of fluids only or layers only, images were prepared as shown in Figure 2 thus reducing the problem of segmentation to semantic segmentation with eight classes: area above ILM (vitreous), area between ILM and INL/IPL, area between IPL/INL and RPE, area below the BM (choroid), and three retinal fluids (PED, SRF, IRF).

Models for automatic segmentation

Architectures

Since its introduction in 2015, the U-net architecture [26] and its various modifications have been the most used architectures for medical image segmentation. The U-net architecture consists of an encoder (contractive path), decoder (expanding path), and skip connections which enables simultaneous capturing of context and localization as it is shown in Figure 3. We will not explain it in detail as it is a well-known architecture for medical image segmentation. We used standard U-net architecture to set a baseline for automatic segmentation.

Although many modifications of U-net architecture were proposed [34,37–39] we chose two state-of-the-art architectures to get further insights into challenges for automatic segmentation. The groundbreaking improvement in computer vision was achieved with ResNet architecture [40] and DenseNet architecture [41], where both reached breakthrough results in classification on the ImageNet dataset [42]. The logical step was to improve the U-net architecture in the ResNet and DenseNet style fashion. We opted for U-net-like architecture [43] and U-net++ [44] as two state-of-the-art architectures. The former combines the good sides of U-net and ResNet architectures and the latter is inspired with DenseNet architecture. A recent paper by Isensee et al. [45], where they proposed a nnU-net (out-of-the-box tool) for biomedical image segmentation that uses U-net-like architecture was another argument for our choice of architecture.

The U-net-like architecture is shown in Figure 4. In the down-sampling path, residual blocks contain convolutional filters (3×3) followed by batch normalization (BN) [46] and a ReLU activation function. Down-sampling is accomplished by max-pooling (MP) which reduces image size by half. Skip connection in the residual block is not just an identity connection but it contains a convolutional filter (1×1) with strides equal to 2 and in that way is achieved down-sampling (image size is reduced by half, same as with MP in the main branch of the residual block). At the end of the residual block, the outputs from the main branch (also called a layer) and the skip connection are summarized. Residual blocks in the decoder are implemented in a similar manner: in a layer (the main branch) there are transpose convolutions (3×3), BN, activation function (ReLU), and up-sampling (which double the size of the image). In the skip connection, there is a convolution (1×1) and up-sampling.

The U-net++ architecture (a nested U-net architecture for medical image segmentation) is shown in Figure 5. It is seen how U-net architecture is enhanced with dense blocks and convolution layers between the encoder and decoder. The purpose of modified skip pathways is to reduce the semantic gap between the feature maps of the encoder and decoder before merging. Each circle represents a set of convolution operations. The shaded part shows the original U-net architecture, while the rest (middle part) shows the difference from the original architecture: dense convolution blocks on the skip pathways and deep supervision [44].

Training

Original images size 1024×512 pixels were resized to 512×256 pixels. Categorical cross-entropy loss was

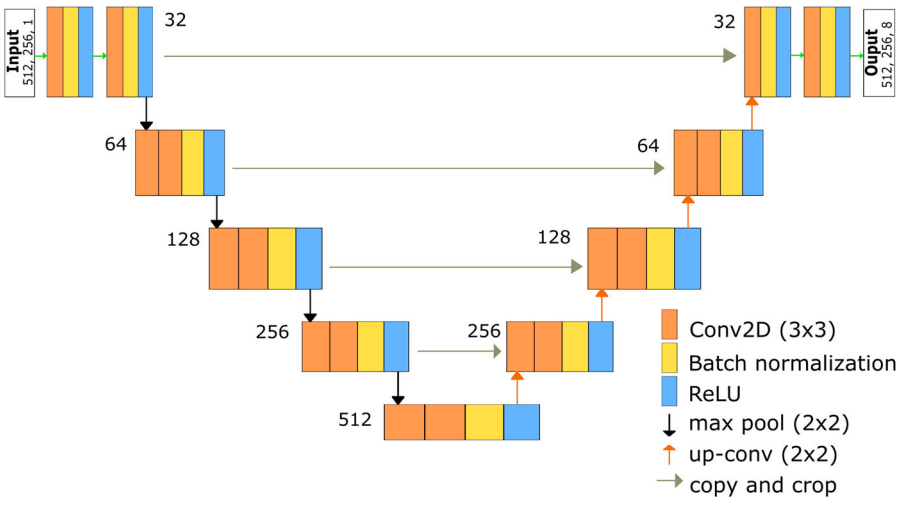

Figure 3. U-net architecture [26].

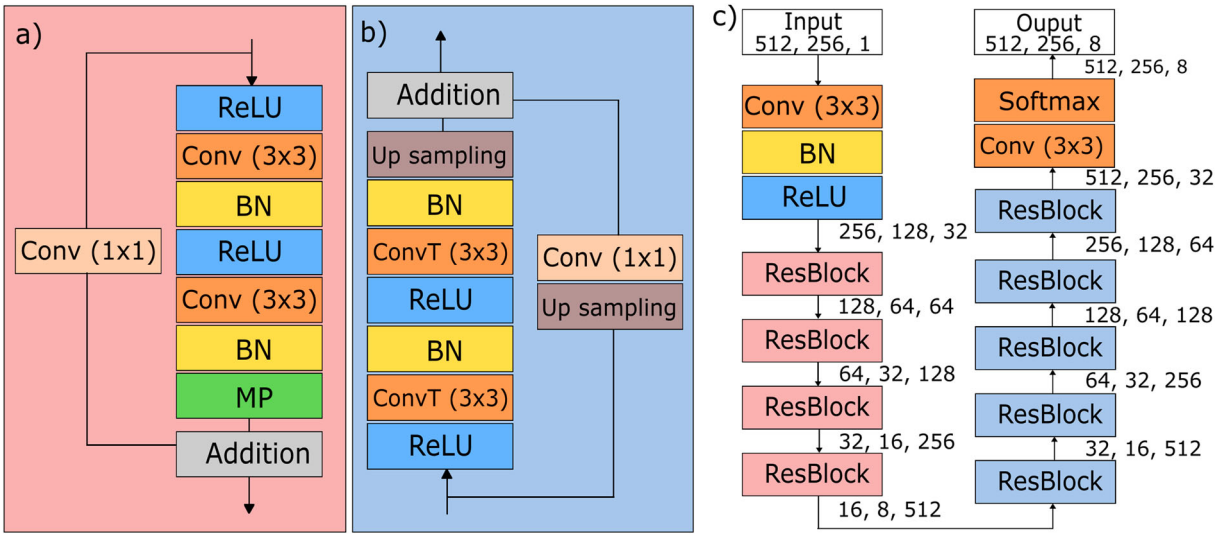

Figure 4. (a) A residual block in the encoder. (b) A residual block in the decoder. (c) The used U-net-like architecture.

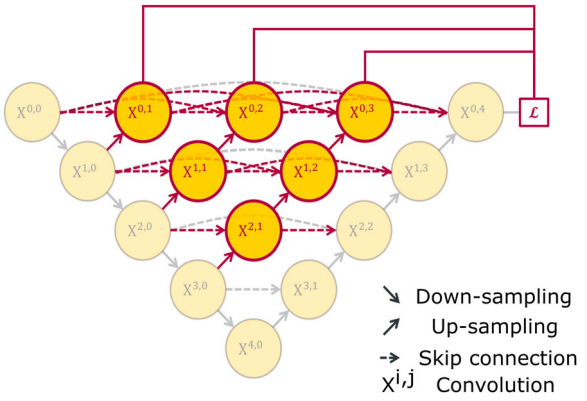

Figure 5. U-net++: a nested U-net architecture [44].

used to train all models. The batch size was set to 4 (we obtained worse results with a larger batch size). The AdaBound optimizer [47] was used (as it combines advantages of Adaptive Moment Estimation (Adam) and Stochastic Gradient Descent (SGD)). Number of trainable parameters for standard U-net, U-net-like, and U-net++ architectures are 7,764,744, 8,230,536, and 9,041,832, respectively. Early stopping was used to prevent overfitting.

K-fold cross-validation was used where each fold contains images from four patients (1st fold contains images from patients 1 to 4, 2nd fold contains images from patients 5 to 8 and so on). K equals 6 in our procedure since, in that way, the test set share is approximately 15% as is a recommendation and common practice in a small data regime. We do not recommend splitting the sets of images from the same patient across training, validation, and test set as adjacent B-scans are similar and that would lead to overestimated validation of the method.

The models were trained on Google Colab [48] with a GPU. The models were implemented in Python, using the Keras library with the TensorFlow backend.

Results

Inter- and intra-observer error

For the purpose of calculating inter- and intra-observer error, annotations were additionally made for 75 B-scans (randomly chosen from the existing dataset). To calculate intra-observer error, the same expert who made all annotations (1st expert) made a re-annotation with no reference to previous annotations, and with a time delay (3 months after finishing the first annotations) for the 75 B-scans. To calculate inter-observer error, annotations were made by another expert (2nd expert) for the same 75 B-scans. Figure 6 shows two examples: an example of good and bad matching. Differences of opinion among experts can be observed. This can partly be explained by the poor quality of images (a large amount of speckle-noise), and by the fact that in the face of extensive disturbance of normal retinal structure, certain OCT findings cannot be reliably discerned or localized without normal anatomic landmarks. Also, with suboptimal image quality, the annotated changes could be confounded with other pathological phenomena (hyperreflective foci, pseudodrusen, outer retinal tubulations, etc.).

Model prediction errors

Some examples of the automatic segmentation results are shown in Figure 7. It is visible that results are good in case there are no significant pathological changes and deformation in retinal structure. In case there are pathological changes, segmentation predictions are deficient in preserving the topology. Also, fluids segmentation should be enhanced. The main cause is the low representation of pixels belonging to these classes in the total number of pixels, especially when it comes to IRF (not present in all patients nor in all B-scans; in addition, it is regularly smaller than SRF and PED).

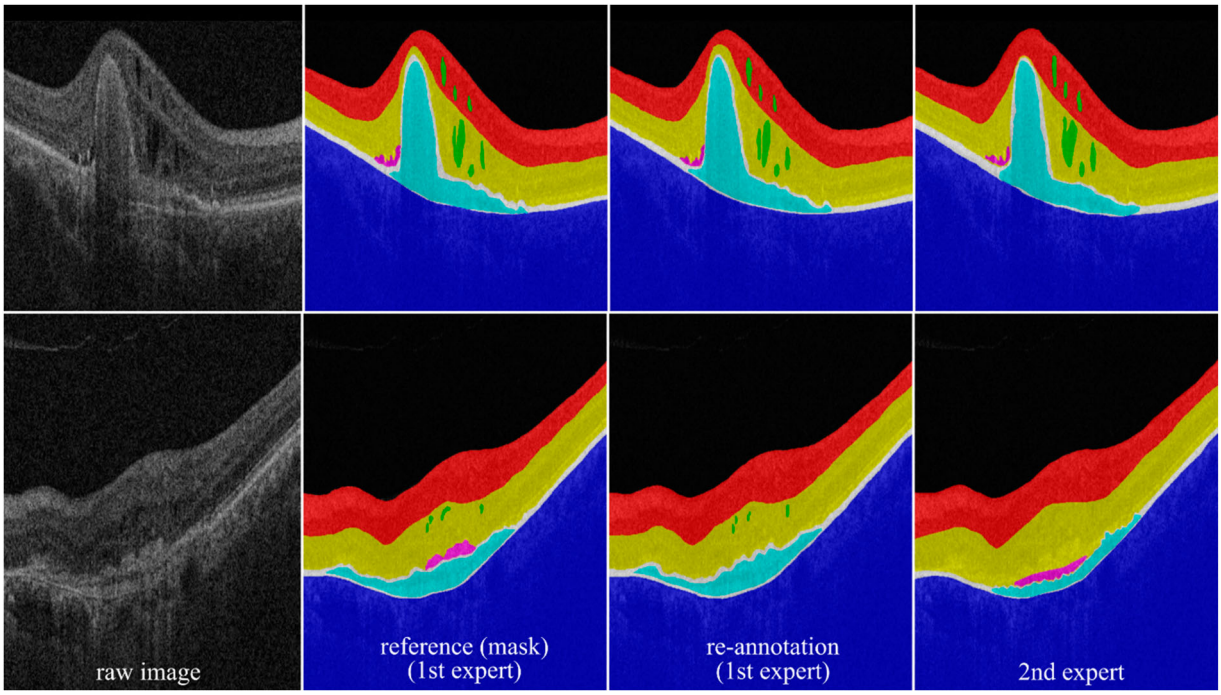

Figure 6. From left to right: raw image, reference (annotations from 1st expert), re-annotations from 1st expert with time delay, annotations from 2nd expert. In the first row, there is an example with good matching. In the second row, there is an example with some differences even between annotations from the same expert (in annotations of PED and SRF) and between two experts (different annotations for fluids). Images are cropped and only the ROI is visible.

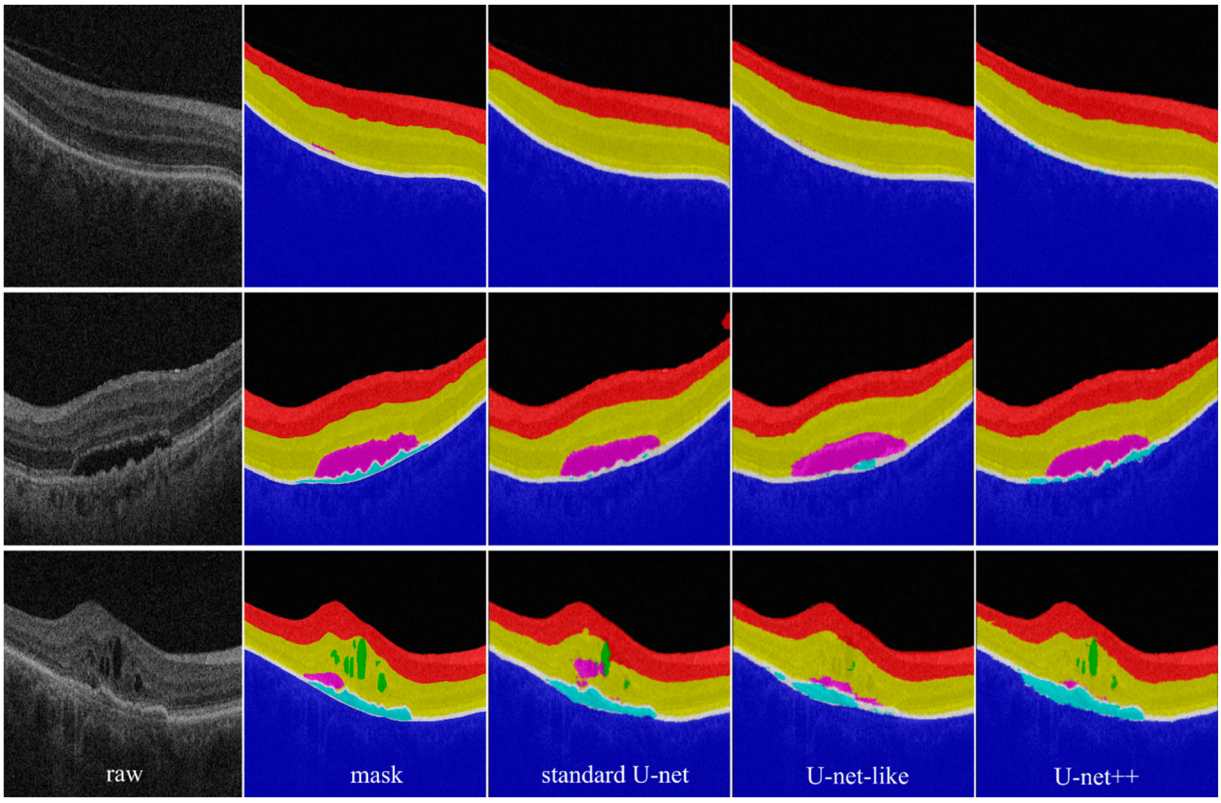

Figure 7. Three examples of the segmentation results. First row: a case with less pronounced pathological changes. Second row: a case with more pathological changes (PED and SRF are present). Third row: a case with extensive pathological changes (PED, SRF, and IRF are present and there is a large distortion of layers). From left to right: raw image, expert annotation (mask), the prediction from the standard U-net architecture, the prediction from the U-net-like architecture, and the prediction from the U-net++ architecture. Images are cropped and only the ROI is visible.

Comparison of inter-observer error and model prediction error

We use the Dice score to evaluate results as it is a similarity measure often used as a metric in the segmentation of medical images. It is calculated according to Equation (1) where TP is true positive, FP false positive, and FN false negative.

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

In Table 1 there are reported Dice scores (mean and standard deviation) for each class and for the inter- and intra-observer error, as well as the prediction errors from the U-net, U-net-like, and U-net++ models. The same results are shown in Figure 8 with bar graphs. We calculated the Dice score for each patient in test fold (that means four patients) but only for those B-scans with a reference segmentation. In case fluids (either SRF or IRF) were not present on a single B-scan of the patient, the Dice score is exempted when calculating the mean and standard deviation of Dice score for that class and that patient. We found it more appropriate than setting it to zero or one as it would lead to overestimating or underestimating the metric value. In the case of patients in whom fluids were present on some of the B-scans but not all, for B-scans where they were not present the value for Dice score was set to zero (the Dice score is not defined in the case of zero division which is a common situation in the absence of the class). It can be observed that for the inter-observer case and in all cases of the automatic segmentation the biggest errors occur in class 3 (surface between RPE and BM) and in classes that represent fluids (PED, SRF, and IRF). One of the factors that contribute to the complexities of automatic segmentation is significant class imbalance: the background (area above ILM and under BM) occupies as much as 83.26% in the total number of pixels, while IRF occupies only 0.12%; surface between RPE and BM occupies 1.07%, SRF 1.05%, PED 1.5%. Further on, out of a total of 1136 scans, PED, SRF, and IRF are present in 1014 (89.26%), 648 (57.04%), and 229 (20.16%) B-scans, respectively.

As Dice score is not an appropriate metric in case of high class imbalance (the Dice score for regions above the ILM and below BM will always be close to one), we also provided results for evaluation in case of converting the layer-segmentation task into a boundary detection problem. In Table 2 there are reported mean square errors (MSE) with belonging standard deviations in the inter-observer case, the intra-observer

Table 1. The Dice score (mean and standard deviation) in the inter-observer case, the intra-observer case, for standard U-net model, U-net-like model, and U-net++ model.

	Above ILM	ILM-IPL/INL	IPL/INL-RPE	RPE-BM	Under BM	PED	SRF	IRF
Inter-observer	0.982 (0.072)	0.950 (0.111)	0.948 (0.112)	0.699 (0.129)	0.989 (0.114)	0.860 (0.301)	0.876 (0.366)	0.735 (0.280)
Intra-observer	0.998 (0.003)	0.973 (0.008)	0.970 (0.117)	0.778 (0.092)	0.998 (0.001)	0.912 (0.242)	0.924 (0.331)	0.844 (0.140)
Standard U-net	0.995 (0.011)	0.950 (0.028)	0.923 (0.083)	0.669 (0.129)	0.988 (0.016)	0.638 (0.173)	0.513 (0.287)	0.480 (0.241)
U-net-like	0.995 (0.004)	0.899 (0.040)	0.890 (0.066)	0.476 (0.132)	0.988 (0.014)	0.533 (0.139)	0.372 (0.293)	0.037 (0.061)
U-net++	0.992 (0.011)	0.944 (0.032)	0.924 (0.064)	0.641 (0.133)	0.986 (0.017)	0.622 (0.159)	0.487 (0.280)	0.419 (0.274)

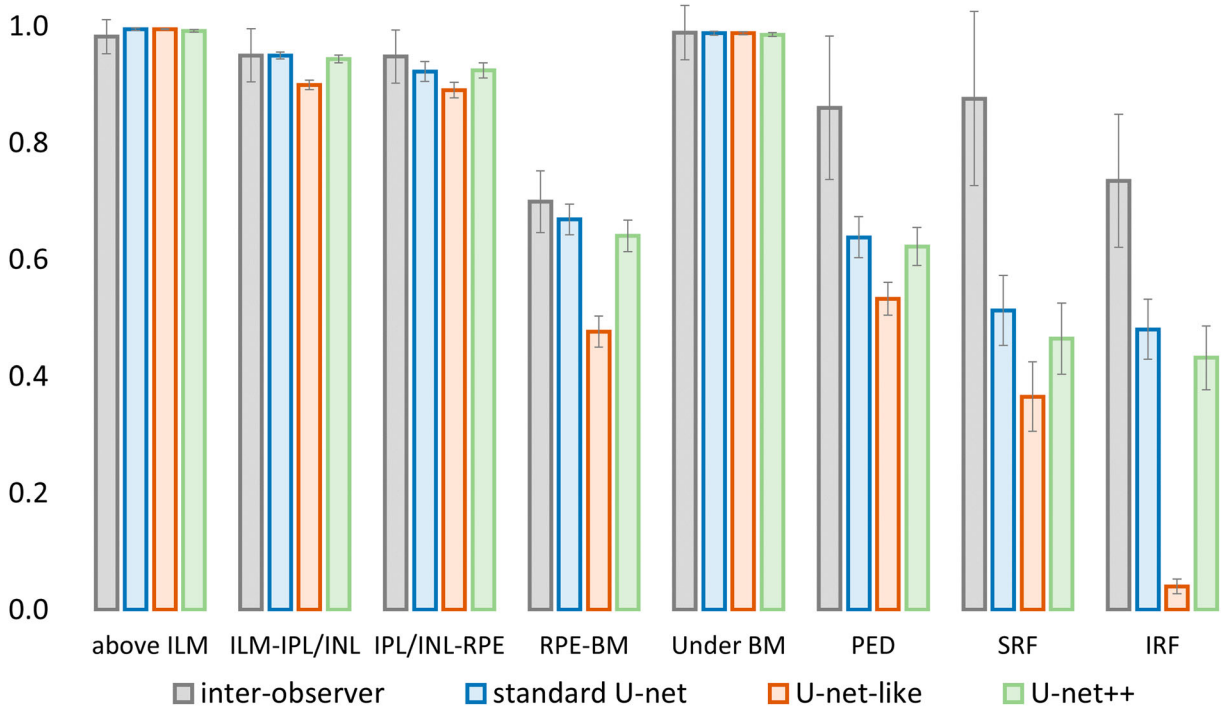

Figure 8. The Dice scores (mean and standard error of the mean) for inter-observer variability, for standard U-net model, U-net-like model, and U-net++ model.

Table 2. The evaluation of the layer-segmentation task as a boundary detection problem: the mean square error (MSE) with belonging standard deviations in the inter-observer case, the intra-observer case, for standard U-net model, U-net-like model, and U-net++ model.

	ILM	IPL/INL	RPE	BM
Inter-observer	5.87 (3.68)	20.46 (47.71)	51.74 (148.15)	12.77 (17.79)
Intra-observer	2.10 (1.43)	5.93 (4.80)	8.28 (9.33)	5.17 (6.68)
Standard U-net	6.23 (3.88)	32.55 (50.22)	60.22 (173.23)	15.88 (19.87)
U-net-like	6.51 (4.22)	40.34 (56.67)	65.47 (177.23)	24.56 (31.22)
U-net++	6.02 (4.01)	37.13 (54.55)	61.55 (165.33)	17.11 (20.12)

case, for standard U-net model, U-net-like model, and U-net++ model. Values are shown in pixels where each pixel corresponds to 1.96 μm along the axial (Z) axis.

Our research suggests that more complex architectures result in only slightly enhanced outcomes, no enhancement at all, or worse outcomes. For classes that represent fluids, the results are noticeably worse, compared to human error. Better results are obtained with U-net++ architecture than with U-net-like architecture, probably because U-net-like architecture lacks skip connections between encoder and decoder (skip connections only exist in each residual block within encoder/decoder) while there are dense blocks and convolutional layers between encoder and decoder in U-net++ architecture. Preliminary, we could conclude that better segmentation accuracy cannot be obtained only with more complex architectures, but rather using some of the efficient techniques in case of distinct class imbalance and the need for preserving the topology. Also, due to the suboptimal quality of images, preprocessing could help in achieving better accuracy.

Discussion

The AROI database contains a large sample of 1136 B-scans with exhaustive annotations (both fluids and retinal layers were annotated by an expert ophthalmologist). As images are collected from patients suffering from nAMD where pathologic biomarkers and large distortion of the retinal structure are present, automatic segmentation of such images presents a significant challenge. Figure 9 shows the confusion matrix for inter-observer error, and for model (U-net, U-net-like, and U-net++) prediction error. In the case of automatic segmentation, it is observed that IRF is often misclassified as the surface between IPL/INL and RPE (class 2) as it is in that area. In case the IRF is smaller or does not differ significantly in intensity and texture from the surface between IPL/INL and RPE, the model does not recognize it as a separate class. Also, it is observed that in a similar way SRF is often misclassified as PED or surface between IPL/INL and RPE. PED is mostly misclassified as an area under BM, and with further inception of individual predictions, it is apparent that it happens when Bruch's membrane is not clearly visible (due to geographic atrophy or some other pathological changes). These results of automatic segmentation for three models (standard U-net architecture and two state-of-the-art architectures) can serve as a baseline for further development of deep learning models.

Also, intra- and inter-observer errors were calculated to enable the validation of algorithms for automatic segmentation. However, it is still not clear what level of segmentation accuracy we need in clinical practice as manual segmentation is rarely performed in clinical practice. There is probably no universal rule, and the required accuracy of segmentation will depend on the purpose – whether it is a diagnosis or prediction of the outcome of anti-VEGF therapy or prediction of another eye disease.

We have limited ourselves to collecting images from only one type of OCT device and from patients suffering from one type of disease. However, developing algorithms on as large and as diverse databases as possible would provide for more robust algorithms that could be implemented in commercial OCT device software. We hope that open access will become a common practice for the majority of research groups in the future and that online image collections and repositories will contribute to building a single database covering various diseases, various types of devices, and various retinal structures annotated from different experts.

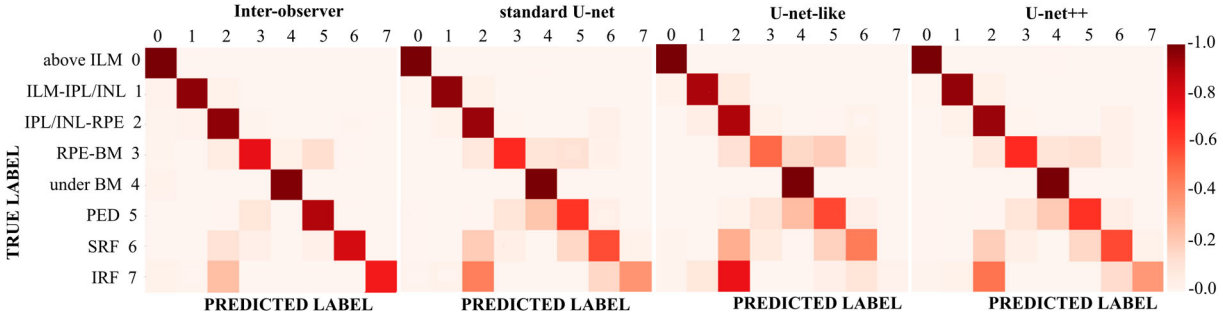

Figure 9. The confusion matrix for inter-observer error, for U-net model, U-net-like model, and U-net++ model prediction error.

As a lack of publicly available databases is one of the major obstacles to introducing AI and deep learning to ophthalmology, we consider the development of the AROI database as a step forward to introducing automatic segmentation in clinical practice and thus enabling quantitative analysis and more successful diagnosis and therapy.

Acknowledgements

Special thanks to Aida Kasumović, MD (Department of Ophthalmology, Sestre milosrdnice University Hospital Center, Zagreb, Croatia) who did image labelling for the purpose of calculating inter-observer error. Also, authors would like to thank Filip Melinščak for valuable discussions and comments on the text.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data availability statement

Code is available on GitHub: https://github.com/mmelinscak/OCT-images-segmentation. Dataset is available at: https://ipg.fer.hr/ipg/resources/oct_image_database [49] and https://doi.org/10.17605/OSF.IO/5WYR3 [50].

ORCID

Martina Melinščak http://orcid.org/0000-0001-5128-3213

References

[1] Jager RD, Mieler WF, Miller JW. Age-related macular degeneration. N Engl J Med. 2008;358(24):2606–2617.

[2] 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 Global Health. 2014 Feb;2(2):e106–e116. DOI:10.1016/S2214-109X(13)70145-1

[3] Klein R, Klein BEK, Linton KLP. Prevalence of age-related maculopathy. Ophthalmology. 1992 Jun;99(6):933–943. DOI:10.1016/S0161-6420(92)31871-8.

[4] Coleman HR, Chan C-C, Ferris FL, et al. Age-related macular degeneration. Lancet. 2008 Nov;372(9652):1835–1845. DOI:10.1016/S0140-6736(08)61759-6

[5] Kaszubski P, Ben Ami T, Saade C, et al. Geographic atrophy and choroidal neovascularization in the same eye: a review. Ophthalmic Res. 2016;55(4):185–193. DOI:10.1159/000443209

[6] Patel HR, Eichenbaum D. Geographic atrophy: clinical impact and emerging treatments. Ophthalmic Surg Lasers Imaging Retina. 2015 Jan;46(1):8–13. DOI:10.3928/23258160-20150101-01

[7] Brown D, Heier JS, Boyer DS, et al. Current best clinical practices—management of neovascular AMD. J Vitreo-Retin Dis. 2017 Sep;1(5):294–297. DOI:10.1177/2474126417725946

[8] Spaide RF, Jaffe GJ, Sarraf D, et al. Consensus nomenclature for reporting neovascular age-related macular degeneration data. Ophthalmology. 2020 May;127(5):616–636. DOI:10.1016/j.ophtha.2019.11.004

[9] Schmidt-Erfurth U, Waldstein SM. A paradigm shift in imaging biomarkers in neovascular age-related macular degeneration. Prog Retin Eye Res. 2016 Jan;50:1–24. DOI:10.1016/j.preteyeres.2015.07.007

[10] Ritter M, Simader C, Bolz M, et al. Intraretinal cysts are the most relevant prognostic biomarker in neovascular age-related macular degeneration independent of the therapeutic strategy. Br J Ophthalmol. 2014 Dec;98(12):1629–1635. DOI:10.1136/bjophthalmol-2014-305186

[11] Ristau T, Keane PA, Walsh AC, et al. Relationship between visual acuity and spectral domain optical coherence tomography retinal parameters in neovascular age-related macular degeneration. Ophthalmologica. 2013 Oct;231(1):37–44. DOI:10.1159/000354551

[12] Willoughby AS, Ying G-S, Toth CA, et al. Subretinal hyperreflective material in the comparison of age-related macular degeneration treatments trials. Ophthalmology. 2015 Sep;122(9):1846–1853.e5. DOI:10.1016/j.ophtha.2015.05.042

[13] Waldstein SM, Philip A-M, 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 Feb;134(2):182. DOI:10.1001/jamaophthalmol.2015.4948

[14] Lee H, Jo A, Kim HC. Three-dimensional analysis of morphologic changes and visual outcomes in neovascular age-related macular degeneration. Invest Ophthalmol Vis Sci. 2017 Feb;58(2):1337. DOI:10.1167/iovs.16-20637

[15] Bogunović H, Waldstein SM, Schlegl T, et al. Prediction of anti-VEGF treatment requirements in neovascular AMD using a machine learning approach. Invest Ophthalmol Vis Sci. 2017 Jun;58(7):3240–3248. DOI:10.1167/iovs.16-21053

[16] Adamis AP, Brittain CJ, Dandekar A, et al. Building on the success of anti-vascular endothelial growth factor therapy: a vision for the next decade. Eye. 2020 Jun. DOI:10.1038/s41433-020-0895-z

[17] Sheyman A, Fawzi AA, editors. Retinal vascular disease. Singapore: Springer Singapore; 2020. DOI:10.1007/978-981-15-4075-2

[18] Peng Y, Keenan TD, Chen Q, et al. Predicting risk of late age-related macular degeneration using deep learning. arXiv:2007.09550 [cs, eess]; 2020 Jul. [cited 2020 Jul 25]. Available from: http://arxiv.org/abs/2007.09550

[19] Yim J, Chopra R, Spitz T, et al. Predicting conversion to wet age-related macular degeneration using deep learning. Nat Med. 2020 May. DOI:10.1038/s41591-020-0867-7

[20] Siedlecki J, Fischer C, Schworm B, et al. Impact of subretinal fluid on the long-term incidence of macular atrophy in neovascular age-related macular degeneration under treat & extend anti-vascular endothelial growth factor inhibitors. Sci Rep. 2020 Dec;10(1):8036. DOI:10.1038/s41598-020-64901-9

[21] Kim KT, Lee H, Kim JY, et al. Long-term visual/anatomic outcome in patients with fovea-involving fibrovascular pigment epithelium detachment presenting choroidal neovascularization on optical coherence tomography angiography. JCM. 2020 Jun;9(6):1863. DOI:10.3390/jcm9061863

[22] Schmidt-Erfurth U, Vogl W-D, Jampol LM, et al. Application of automated quantification of fluid volumes to anti-VEGF therapy of neovascular age-related macular

degeneration. Ophthalmology. 2020 Mar:S0161642020302682. DOI:10.1016/j.ophtha.2020.03.010
[23] Sadda S, Wu Z, Walsh AC, et al. Errors in retinal thickness measurements obtained by optical coherence tomography. Ophthalmology. 2006 Feb;113(2):285–293. DOI:10.1016/j.ophtha.2005.10.005
[24] Waldstein SM, Gerendas BS, Montuoro A, et al. Quantitative comparison of macular segmentation performance using identical retinal regions across multiple spectral-domain optical coherence tomography instruments. Br J Ophthalmol. 2015 Jun;99(6):794–800. DOI:10.1136/bjophthalmol-2014-305573
[25] RETOUCH – grand challenge. grand-challenge.org. [cited 2020 Jun 30]. Available from: https://retouch.grand-challenge.org/
[26] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. arXiv:1505.04597 [cs]; 2015 May. Available from: http://arxiv.org/abs/1505.04597
[27] Bogunovic H, Venhuizen F, Klimscha S, et al. RETOUCH – the retinal OCT fluid detection and segmentation benchmark and challenge. IEEE Trans Med Imaging. 2019:1. DOI:10.1109/TMI.2019.2901398
[28] Chiu SJ, Allingham MJ, Mettu PS, et al. Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema. Biomed Opt Express. 2015 Apr;6(4):1172. DOI:10.1364/BOE.6.001172
[29] Khan SM, Liu X, Nath S, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. Lancet Digital Health. 2020 Oct:S2589750020302405. DOI:10.1016/S2589-7500(20)30240-5
[30] Montuoro A, Waldstein SM, Gerendas BS, et al. Joint retinal layer and fluid segmentation in OCT scans of eyes with severe macular edema using unsupervised representation and auto-context. Biomed Opt Express. 2017 Mar;8(3):1874. DOI:10.1364/BOE.8.001874
[31] 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 Aug;8(8):3627. DOI:10.1364/BOE.8.003627
[32] Wei H, Peng P. The segmentation of retinal layer and fluid in SD-OCT images using mutex dice loss based fully convolutional networks. IEEE Access. 2020;8:60929–60939. DOI:10.1109/ACCESS.2020.2983818
[33] De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med. 2018 Sep;24(9):1342–1350. DOI:10.1038/s41591-018-0107-6
[34] Wang J, Chen C, Li F, et al. S-D Net: joint segmentation and diagnosis revealing the diagnostic significance of using entire RNFL thickness in glaucoma; p. 10.
[35] Mishra Z, Ganegoda A, Selicha J, et al. Automated retinal layer segmentation using graph-based algorithm incorporating deep-learning-derived information. Sci Rep. 2020 Dec;10(1):9541. DOI:10.1038/s41598-020-66355-5
[36] Staurenghi G, Sadda S, Chakravarthy U, et al. Proposed lexicon for anatomic landmarks in normal posterior segment spectral-domain optical coherence tomography. Ophthalmology. 2014 Aug;121(8):1572–1578. DOI:10.1016/j.ophtha.2014.02.023
[37] Gu Z, Cheng J, Fu H, et al. CE-Net: context encoder network for 2D medical image segmentation. IEEE Trans Med Imaging. 2019 Oct;38(10):2281–2292. DOI:10.1109/TMI.2019.2903562
[38] Orlando JI, Seeböck P, Bogunović H, et al. U2-Net: a Bayesian U-Net model with epistemic uncertainty feedback for photoreceptor layer segmentation in pathological OCT scans. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019); 2019 Apr; p. 1441–1445. DOI:10.1109/ISBI.2019.8759581
[39] Liu W, Sun Y, Ji Q. MDAN-UNet: multi-scale and dual attention enhanced nested U-Net architecture for segmentation of optical coherence tomography images. Algorithms. 2020 Mar;13(3):60. DOI:10.3390/a13030060
[40] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. 2016 Jun; p. 770–778. DOI:10.1109/CVPR.2016.90
[41] Huang G, Liu Z, van der Maaten L, et al. Densely connected convolutional networks. arXiv:1608.06993 [cs]; 2018 Jan. [cited 2020 Jul 19]. Available from: http://arxiv.org/abs/1608.06993
[42] ImageNet. [cited 2020 Sep 14]. Available from: http://www.image-net.org/
[43] K. Team. Keras documentation: image segmentation with a U-Net-like architecture. [cited 2020 Sep 14]. Available from: https://keras.io/examples/vision/oxford_pets_image_segmentation/
[44] Zhou Z, Siddiquee MMR, Tajbakhsh N, et al. UNet++: a nested U-Net architecture for medical image segmentation. arXiv:1807.10165 [cs, eess, stat]; 2018 Jul. [cited 2020 Mar 30]. Available from: http://arxiv.org/abs/1807.10165
[45] Isensee F, Jaeger PF, Kohl SAA, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18(2):203–211. DOI:10.1038/s41592-020-01008-z.
[46] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift; p. 11.
[47] Luo L, Xiong Y, Liu Y, et al. Adaptive gradient methods with dynamic bound of learning rate. arXiv:1902.09843 [cs, stat]; 2019 Feb. [cited 2021 Feb 4]. Available from: http://arxiv.org/abs/1902.09843
[48] Google colaboratory. [cited 2020 May 17]. Available from: https://colab.research.google.com/notebooks/intro.ipynb
[49] OCT image database – image processing group. [cited 2021 Jan 28]. Available from: https://ipg.fer.hr/ipg/resources/oct_image_database#
[50] Melinscak M. Retinal OCT image segmentation. 2021 Feb. DOI:10.17605/OSF.IO/5WYR3