scientific data

OPEN

DATA DESCRIPTOR

OCT5k: A dataset of multi-disease and multi-graded annotations for retinal layers

Mustafa Arikan[1], James Willoughby[1], Sevim Ongun[1], Ferenc Sallo[2], Andrea Montesel[2], Hend Ahmed[3], Ahmed Hagag[1,4], Marius Book[5], Henrik Faatz[6], Maria Vittoria Cicinelli[7,8], Amani A. Fawzi[9], Dominika Podkowinski[10], Marketa Cilkova[4], Diana Morais De Almeida[2], Moussa Zouache[11], Ganesham Ramsamy[12], Watjana Lilaonitkul[13,14,15,16] & Adam M. Dubis[1,11,16] ✉

Publicly available open-access OCT datasets for retinal layer segmentation have been limited in scope, often being small in size, specific to a single disease, or containing only one grading. This dataset improves upon this with multi-grader and multi-disease labels for training machine learning-based algorithms. The proposed dataset covers three subsets of scans (Age-related Macular Degeneration, Diabetic Macular Edema, and healthy) and annotations for two types of tasks (semantic segmentation and object detection). This dataset compiled 5016 pixel-wise manual labels for 1672 OCT scans featuring 5 layer boundaries for three different disease classes to support development of automatic techniques. A subset of data (566 scans across 9 classes of disease biomarkers) was subsequently labeled for disease features for 4698 bounding box annotations. To minimize bias, images were shuffled and distributed among graders. Retinal layers were corrected, and outliers identified using the interquartile range (IQR). This step was iterated three times, improving layer annotations' quality iteratively, ensuring a reliable dataset for automated retinal image analysis.

Background & Summary

Retinal images have become crucial in ophthalmology for detection, tracking progression, and monitoring treatment effectiveness[1–3]. Optical Coherence Tomography (OCT), a non-invasive technique for visualizing tissue in three dimensions, has been extensively used in ophthalmology due to its high resolution and ability to provide detailed images of the retina[4]. Despite significant advancements in OCT technology and analysis techniques, the progress is hampered by the lack of large, high-quality annotated datasets. The need for efficient, accurate, and robust models to analyze large volumes of OCT images is pressing, as manual annotations are not efficiently or economically scalable. Automating this process can improve ophthalmic care by providing fast, consistent, and reliable measurements of retinal layers and disease biomarkers[5–10]. This is particularly important for managing diseases such as Age-related Macular Degeneration[11] (AMD) and Diabetic Macular Edema (DME)[12], where timely and precise monitoring can significantly impact patient outcomes. This highlights the need for automated tools to harness OCT's sensitivity in diagnosing and monitoring DME and AMD.

[1]UCL, Institute of Ophthalmology, London, EC1V 9EL, UK. [2]Jules Gonin Eye Hospital, Department of Ophthalmology, University of Lausanne, Lausanne, Switzerland. [3]University College London Hospitals NHS Foundation Trust, London, UK. [4]Moorfields Eye Hospital NHS Foundation, NIHR Moorfields Biomedical Research Centre, London, EC1V 2PD, UK. [5]Rare Retinal Disease Center, AugenZentrum Siegburg, Siegburg, Germany. [6]Eye Center at St. Franziskus Hospital Münster, Münster, Germany. [7]Department of Ophthalmology, IRCCS San Raffaele Scientific Institute, Milan, Italy. [8]School of Medicine, Vita-Salute San Raffaele University, Milan, Italy. [9]Northwestern University Medical School, Chicago, US. [10]Department of Ophthalmology, Kepler University Clinic, Linz, Austria and Vienna Institute for Research in Ocular Surgery (VIROS), Hanusch Hospital, Vienna, Austria. [11]Department of Ophthalmology & Visual Sciences, University of Utah, Salt Lake City, USA. [12]West Midlands NHS Trust, London, UK. [13]UCL, Global Business School for Health, London, WC1E 6BT, UK. [14]Health Data Research UK (HDR UK), London, NW1 2BE, UK. [15]UCL, Institute of Health Informatics, London, NW1 2DA, UK. [16]These authors contributed equally: Watjana Lilaonitkul, Adam M. Dubis. ✉e-mail: a.dubis@ucl.ac.uk

For quantitative analyses of retinal images to be reliable and facilitate meaningful interpretations[13], accurate and robust segmentation of retinal layers in OCT images is essential, particularly given the variability in layer appearances across different pathologies. For handling large amounts of imaging data and making statistical analysis sufficiently powerful[14,15], fully automated segmentation algorithms[16–20] are preferred. Although images representing retinal layers for a single disease can be segmented automatically with fewer images and annotations for training, more sophisticated algorithms are required to segment retinal layers for a wider set of diseases for better generalization. As a result, publicly available databases and platforms which offer researchers access to image datasets and annotations that can be used to produce sophisticated algorithms for complex automatic ophthalmic image analysis are very desirable[21].

To address the challenges of manual annotation and the lack of high-quality datasets, we introduce the OCT5k dataset, a comprehensive collection of retinal OCT images with multi-disease and multi-graded annotations. This dataset is designed to support the development and validation of automated retinal image analysis algorithms. By providing pixel-wise manual labels for 1672 OCT scans featuring two different diseases (AMD and DME) and healthy subjects, along with 4698 bounding box annotations for a subset of 566 scans across nine classes of disease biomarkers, OCT5k offers a robust foundation for training machine learning models. Our dataset includes annotations from multiple graders to ensure diversity and reduce bias, enhancing the reliability and generalizability of the trained models. Furthermore, the inclusion of both semantic segmentation and object detection tasks in the dataset enables the development of algorithms that can perform comprehensive analysis of retinal images, facilitating diagnosis and monitoring of eye diseases. The OCT5k dataset thus represents a significant step forward in the creation of automated tools for ophthalmic care, promising to improve the accuracy and efficiency of retinal image analysis.

The OCT5k dataset offers a versatile platform for a multitude of research applications. It serves as a benchmark for validating and performing comparative analysis of automated retinal image analysis algorithms, highlighting challenges and potential enhancements. Researchers can identify challenging cases through variability in grader annotations and prioritize these in model training to improve robustness. By training on specific disease categories within the dataset and testing on others, the dataset facilitates studies on model generalization, allowing for the discovery of methods to improve performance across varying pathological conditions. Additionally, comparing a model's agreement with graders to the agreement between different graders can provide insights into the model's reliability. This dataset can be utilized to train models, which can be used to pre-segment and preliminarily annotate new datasets, streamlining the process of annotation correction and validation in studies of new diseases. Additionally, leveraging the multi-grader annotations within the dataset allows for robust analysis techniques that quantify disease severity and progression, by assessing the deviation of diseased samples from healthy controls based on consensus and variability in expert annotation, thereby providing a sophisticated metric for understanding disease impact and trajectory.

Publicly available data sets in combination with artificial intelligence methods like deep learning have great potential to transform research on eye diseases and retinal imaging[22]. A major challenge in deep learning is the need for high-quality labels[23]. To train accurate models and compare different algorithms against one another and against expert grading, the quality of the labels is crucial. Therefore, in order to understand the uncertainty[24] properties of data and models, we need to collect annotations for the same images from different graders.

Open-access datasets

Previously, researchers have released datasets however they lack the breadth, depth and robustness of the dataset we described. Li *et al.*[25] introduced a data set for the glaucoma disease for which there was a single annotation for 9 layers and 122 peripapillary OCT scans, however this has been shown to induce non-physiologically plausible changes[26] which may disrupt learning[27]. The data set was augmented with flips to increase its size to 244 images. Hassan *et al.*[28,29] published a dataset with 42 scans and 6 layers annotated. In He *et al.*[30], 1715 annotations, composed of 8 layers each, were collected for 14 healthy and 21 multiple sclerosis OCT volumes. As part of a study published in Anthony *et al.*[31], layer annotations were manually collected from three graders on some 40 mouse OCT slices. The data set also includes automatic segmentation of 6 and 10 layers for 40 dense OCT volumes. In Gholami *et al.*[32] manual annotations for 25 healthy scans are published. Farsiu *et al.*[33] published 269 AMD volumes and 115 control volumes with three layers, which were semiautomatically segmented. In Tian *et al.*[34], the authors have compiled a dataset containing annotations for 100 individual OCT slices taken from healthy individuals. Chiu *et al.*[35] collected two sets of manual gradings for 110 OCT slices and 8 retinal layers. In Chiu *et al.*[36], they collected two sets of manual gradings for 220 OCT slices and three retinal layers. Melinscak *et al.*[37] published a dataset for AMD containing 1136 images annotated with 3 types of fluids and 4 retinal layers. An additional annotation was made for a subset of 75 OCT slices in order to calculate intra-observer and inter-observer error. Morales *et al.*[38] have collected data for 6 retinal layers and 244 slices from rat OCT volume data.

There are two additional datasets, which are large in size but are currently available for the classification task but lack semantic segmentation labels for the retinal layer segmentation task. Kermany *et al.*[39] have around one hundred thousand OCT B-scan images from four categories CNV (choroidal neovascularization), DME (diabetic macular edema), AMD (age-related macular degeneration), and normal scans. Additionally, Kulyabin *et al.*[40] (OCTDL) provide 2064 images covering even more categories (AMD, DME, ERM (epiretinal membrane), Normal, RAO (retinal artery occlusion), RVO (retinal vein occlusion), VID (Vitreo-macular interface disease)).

To address the limitations of existing data sets, we have provided a large set of annotations for 60 OCT volumes - 20 from individuals with AMD, 20 from individuals with diabetic macular edema (DME), and 20 from healthy individuals – taken from a pool of 148 OCT volumes[41]. The volumes were obtained using a Heidelberg Spectralis (Heidelberg Engineering, Germany) device. The axial resolution is 3:5.m and the scan-dimension is

Disease/Group	Number of total scans	Number of volumes x scans per volume
AMD	722	7×19, 4×25, 2×31, 7×61
DME	422	19×19, 1×61
Healthy	530	12×19, 2×25, 3×31, 1×37, 2×61

Table 1. The total number of individual scans that were graded manually by three different annotators is 1672.

8:9.7:4 mm2. The number of scans per volume varies between 19 and 69 scans. The complete list of volumes for each disease/group and the overall number of scans are shown in Table 1.

Additionally, 4698 bounding box annotations for a subset of 566 scans are provided across nine classes. These 9 classes are choroidal folds, retinal fluid, geographic atrophy, hard drusen, hyperfluorescent spots, photoreceptor layer disruption, reticular drusen, soft drusen, and softdrusen PED (Pigment Epithelial Detachment). We have also included automatic retinal layer segmentations for the remaining 88 OCT volumes, giving a total of 2586 annotations.

In conclusion, we provide a set of, multi-disease, multi-grader labeled set of annotations that consists of pixel-wise annotations for retinal layers, as well as bounding box annotations for different classes of biomarkers in OCT images. It can be used to train and evaluate machine learning-based retinal layer segmentation architectures, as well as to segment key retinal biomarkers and serve as a benchmarking dataset for various segmentation and object detection tasks.

Methods

We first describe the annotation process of the retinal layers. In the subsection, titled "Inter-grader Reliability", we present the details of the process preparing multiple gradings for the same scans. In the second subsection, titled "Object Detection Labels", we present the details of the bounding box annotations. In the section titled "Segmentation Model", we describe the baseline method used to segment the retinal layers.

Ground truth annotation. The OCT image dataset used in this study was originally collected by Rasti *et al.*[41] and is publicly available[42]. The dataset consists of OCT scans from normal subjects as well as patients with age-related macular degeneration (AMD) and diabetic macular edema (DME). The goal was to select an equal number of volumes for all three groups (AMD, DME and healthy). The volumes were visually assessed by an expert and categorized into bad, medium and good quality. From the volumes with medium and good quality, 60 volumes were selected randomly.

The ground truth annotation involves human-machine interaction, tooling, finding best practices for annotation and infrastructure and data management. We have built a customized labeling platform based on the Hitachi's open source semantic segmentation editor[43] to allow multiple graders to annotate OCT images. The OCT images are uploaded to the platform and the layers are drawn according a defined layout. The layout defines the number of layers and the colors assigned to the layers. The layout is adjustable, therefore the number of layers can be changed. The number of layers depends on disease and corresponding thickness of certain layers.

The images for annotation were prepared, shuffled and uploaded to the labeling platform. The graders received access to a portion of the images. To speed up annotation we trained a model using a small number of images. We used this model to prepare segmented masks of retinal layers, which were uploaded as well. The annotation masks were replaced by lines and vertices which delineated the layers. The vertices can be dragged and dropped to change the annotation. In addition the vertices can be deleted and new ones can be added. The graders navigated through folders and images and corrected the five retinal layers. To reduce bias, the images from the same volumes were shuffled and distributed among graders. After the correction of the retinal layers the results were compiled. In a second stage outliers were determined and checked and if necessary corrected again. For the outlier detection we used the interquartile range (IQR). We repeated this step three times and and checked the quality of the layer annotations and improved it iteratively (see Fig. 5).

The agreement or disagreement between different expert graders is valuable for model development and for providing feedback to experts. To support this task and further development of the model, we compiled three sets of annotations with the help of different graders. All graders performed the annotations according to the following guidelines. A sample workflow is visualized in 1.

- ILM: This layer is the furthest from the top and should be segmented by following the pattern of the band.
- OPL: OPL is below ILM and is segmented on top of the bright band of the OPL.
- IS-OS: IS-OS layer could be segmented by distinguishing that the IS-OS boundary would be located above the RPE. This layer may also be referred to as EZ (ellipsoid zone).
- OBRPE: Segmenting the RPE (top and bottom): The outermost highly reflective layer is called the Outer boundary of RPE which is segmented along the bottom band of the layer.
- IBRPE: Inner boundary of RPE is segmented along the top band of the Outer Boundary of RPE layer.

The layout of the annotation platform can be easily adjusted and expanded for different retinal layers. This is necessary, since different retinal layers are effected by different eye diseases. Thus, we can use the annotation platform to annotate different OCT scans from different diseases. In Fig. 1 a layout for manual annotation is presented.

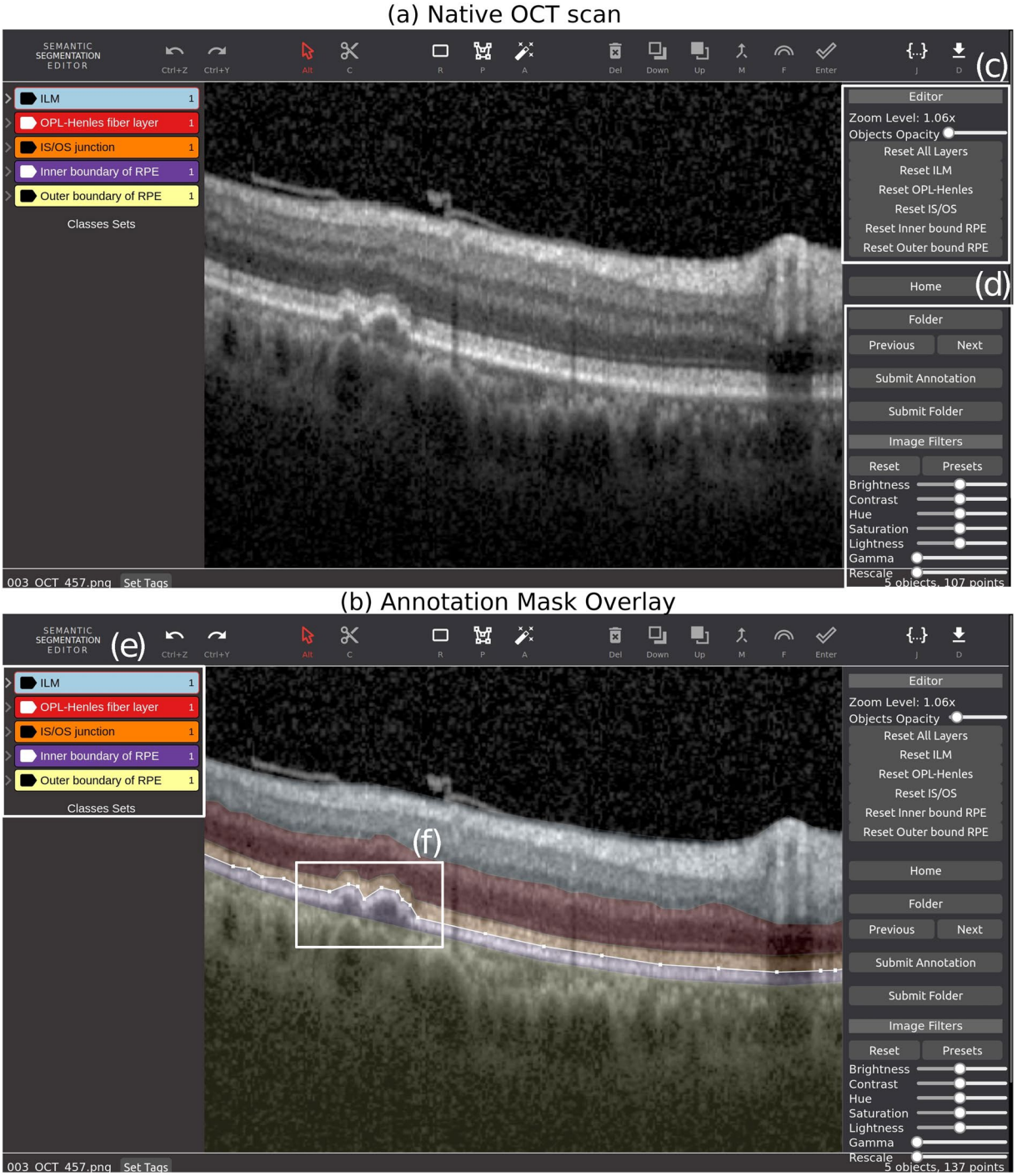

Fig. 1 (**a**) The annotation platform with a layout to annotate five retinal layers, (**b**) and with opacity adjusted, (**c**) functions to change opacity and resetting layers and (**d**) and buttons for navigating, submitting gradings and changing image visualization options such as contrast or hue. (**e**) Custom layout for five retinal layers, (**f**) Area showing editing using vertices.

For the manual annotation and correction of the data set we selected five retinal layers. The graders annotated Internal limiting membrane (ILM), Outer Plexiform Layer (OPL), Photoreceptor Inner Segment and Outer Segment Layers (IS-OS), Inner Boundary of Retinal Pigment Epithelium (IBRPE) and Outer Boundary of Retinal Pigment Epithelium (OBRPE) layers. The retinal layers with a corresponding image and annotation mask is presented in Fig. 2 for AMD and DME.

Inter-grader reliability. In the assessment of inter-grader reliability, we compared the results among pairs of annotations, using the metrics Intersection over Union (IoU) and mean distance between layers to visualize the

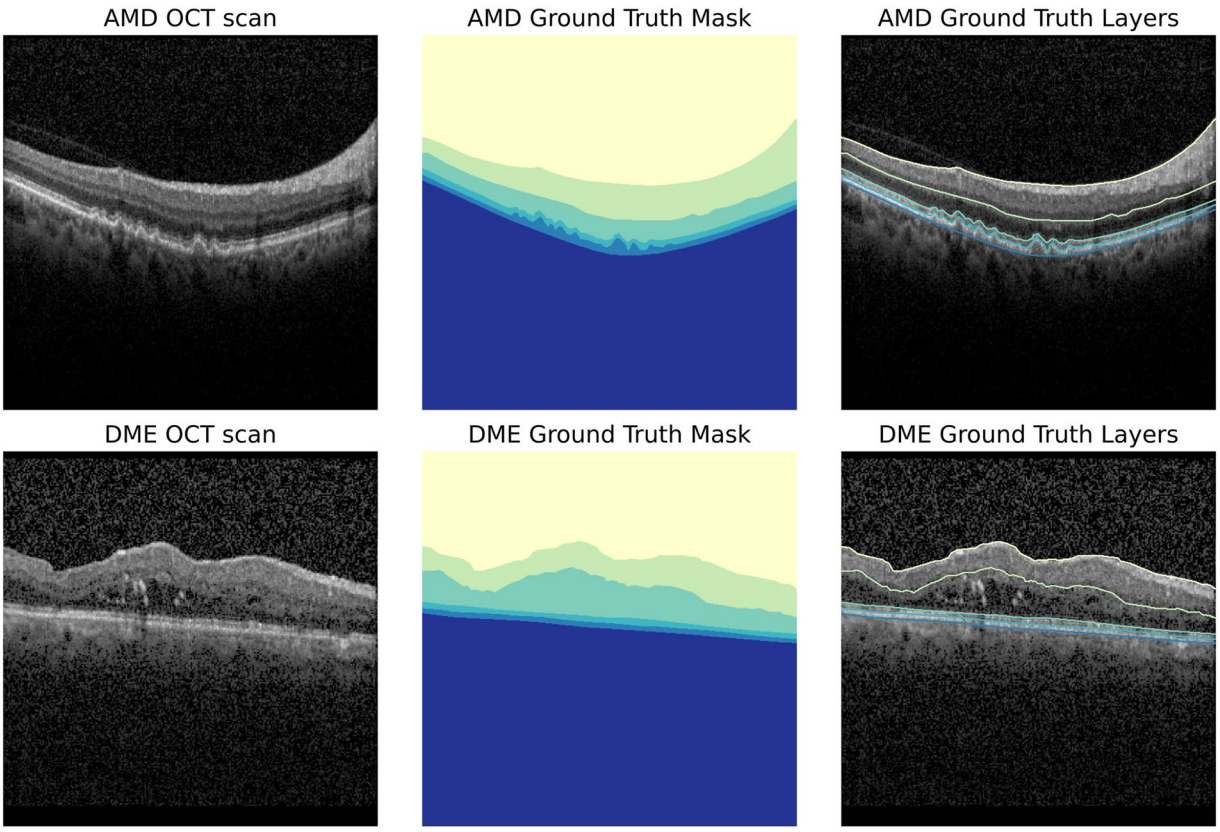

Fig. 2 A single scan from AMD (top) and DME (bottom) (left) and the exported retinal layers from the labeling platform (right). The annotated retinal layers are (from top to bottom): ILM, OPL, IS-OS, IBRPE and OBRPE. And the generated masks for a semantic segmentation task (middle).

degree of agreement between the graders. The analysis, as depicted in Fig. 4, highlighted both agreements and disagreements among the annotations, with notable differences observed particularly in the DME and AMD scans. Specifically, while most upper and lower layer delineations showed relatively low levels of differences, disagreements were more pronounced in the bands IS-OS/IBRPE and IBRPE/OBRPE, especially evident in DME scans. These findings emphasize how complex retinal layer segmentation can be, especially when it comes to identifying subtle details that are crucial for accurate modeling of disease progression.

Object detection labels. Besides retinal layers, we have collected bounding box annotations[44] for 9 different classes and a subset of 566 images. The bounding box annotations were created by a set of experts who were supplied with images from the AMD scans of the Rasti *et al.*[41] and Kermany *et al.*[39] datasets. Each expert was given a unique set of images to annotate with the exception of one set of 100 images which was annotated twice by separate experts. Geographic Atrophy bounding boxes were required to span the entire width of the transillumination defect with the inclusion of up to 5 pixels either side and were required to include the first choroidal vessels on the bottom and minimal pixels at the top. Drusen annotations were required to encompass the peak of the drusen, the Bruch's boundary and both points where the drusen joins the RPE. Each drusen was annotated individually according to these requirements.

The bounding box annotations is composed of 566 Images with a total of 4698 objects distributed across nine classes. The distribution of the objects between the classes is given in the Table 2 below. In Fig. 3 we present two examples of OCT scans containing bounding box annotations.

Segmentation model. We included a baseline code to demonstrate how the data can be utilized for training a segmentation model, opting to showcase it with a popular model and backbone. For the semantic segmentation of retinal layers we use a U-Net with an EfficientNet backbone. EfficientNet[45] proposes scaling up deep learning models in order to improve accuracy and efficiency much more easily. It uses a technique called compound coefficient scaling to modify each dimension uniformly with a fixed set of coefficients. Using this scaling method and AutoML (Automated machine learning), seven models were created, which outperformed most convolutional neural networks and achieved state-of-the-art performance. Thus, we choose EfficientNet[46] to have the ability for training with different image resolutions and easier adoption to different modalities.

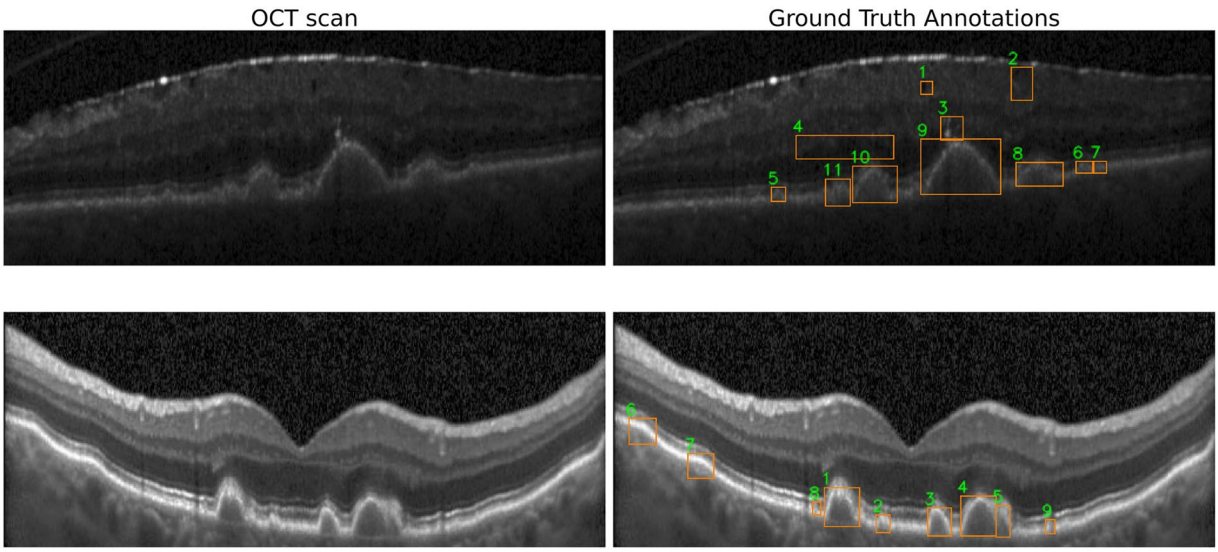

Fig. 3 Examples from two scans with bounding box annotations: Top row has fluid (1,2,4), Hyperfluorescent spots (3), Harddrusen (5,6,7) and Softdrusen(8,9,10,11). Bottom row has Softdrusen PED (1,3,4,5), Softdrusen (2,8,9) and Choroidalfolds (6,7).

Data Records

The dataset is available at the UCL Research Data Repository[47]. The layer annotations are saved in three formats. First, in text files where every line corresponds to a single column of the scan. Second, in grey value image files where every pixel is assigned to a layer. Third, in RGB images, where every pixel is assigned an RGB triplet representing a layer. In addition to annotations we provide scripts to download the original images from Rasti *et al.*[41] and scripts to prepare the images and annotations for model training. The dataset maintains identical folder structure and filenames as the original datasets[41], enabling direct correspondence between our annotations and the source images. We provide an automated script that handles matching of original images with the annotation files. The data[47] set can be downloaded from (https://doi.org/10.5522/04/22128671).

Technical Validation

The annotation strategy for the segmentation data relies on four metrics for annotation quality. Different metrics highlight different aspects of the annotations. They have both advantages and shortcomings. We have used four metrics to analyze the differences between graders for quality assessment and control. Thus, we can apply outlier detection using any of the four metrics for quality assessment and for sending back images for correction.

- Dice similarity coefficient: this is a metric to measure the similarity of two segmentation masks. It can be written as,

$$DSC = \frac{2 \times TP}{(FP + TP) + (TP + FN)}$$

  where TP denotes true positives, FP false positives and FN false negatives.
- Intersection over Union: the IoU metric is related to the DSC and it is calculated by the common pixels between the target and the prediction divided by the total number pixels present in both annotations or masks. The formula is:

$$IoU = \frac{TP}{(TP + FP + FN)}$$

- Mean distance between lines: the mean distance between two lines representing layers is calculated by averaging the sum of absolute differences between two lines. The formula is:

$$MD = \frac{\sum_{n=1}^{N} |Y_1 - Y_2|}{N}$$

where N stands for number of pixels in x-direction and Y_1 and Y_2 stand for position of the layer in y-direction.

Intersection over Union (IoU) is a more robust metric for evaluating the semantic segmentation of retinal layers in OCT images compared to the Dice score. The main reason behind this is that IoU is less sensitive to class imbalance, which is critical for retinal layer segmentation due to the disparity in size between various layers. By calculating the ratio of the area of intersection between the predicted segmentation and the ground truth

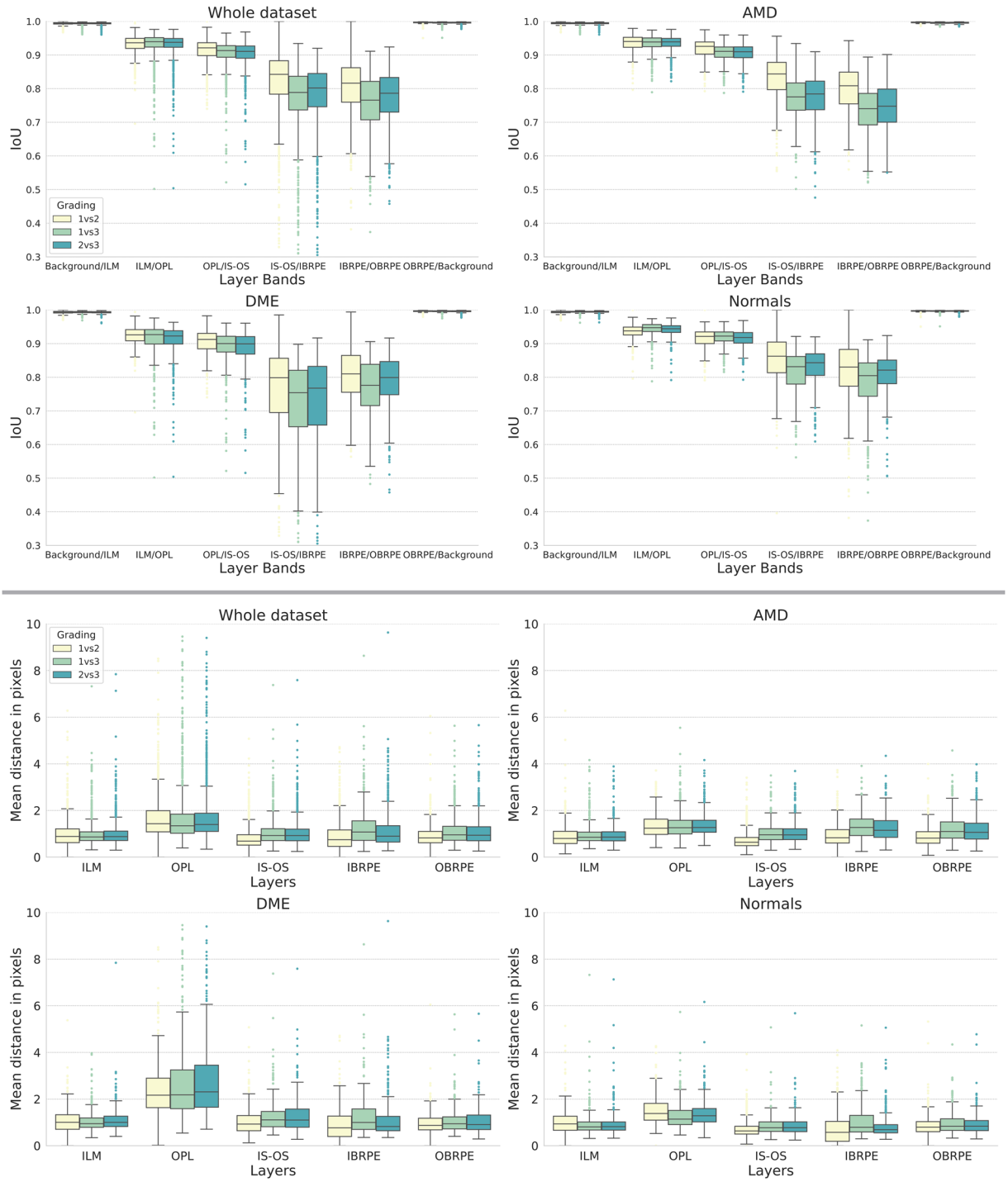

Fig. 4 This figure shows boxplots for two comparison metrics, Intersection over Union (IoU) and Mean Distance, across the four categories of retinal images: AMD, DME, healthy, and the entire dataset. The top row displays IoU results for the whole dataset (top left), AMD (top right), DME (bottom left), and healthy (bottom right). The bottom row displays Mean Distance results for the same categories. Each boxplot compares two gradings for a particular annotation band, providing insight into inter-rater agreement for the dataset.

segmentation to their combined area (union), IoU provides a more accurate representation of the segmentation quality. IoU converges to zero when the ground truth and prediction have no overlap, while the Dice score may still provide non-zero values in such cases, further emphasizing the robustness of IoU for layer segmentation purposes. Therefore, IoU should be selected as the preferred evaluation metric for semantic segmentation of retinal layers in OCT images, since it ensures a more reliable and accurate assessment of the model's performance.

Incorporating mean distance alongside IoU for evaluating semantic segmentation of retinal layers in OCT images addresses IoU's shortcomings when layers disappear or become thin. Mean distance measures average

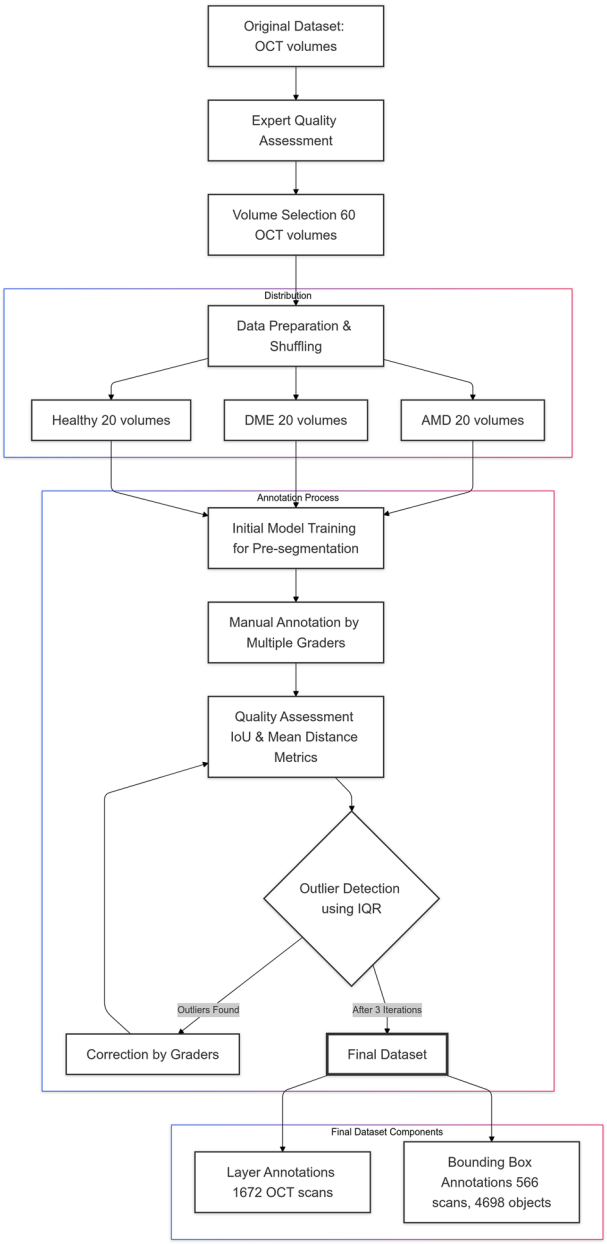

Fig. 5 Flowchart diagram illustrating the dataset creation and quality control process. Starting with the original data containing OCT volumes, expert quality assessment led to the selection of 60 volumes equally distributed across AMD, DME, and healthy categories (20 each). The annotation process began with initial model-assisted pre-segmentation, followed by manual annotation by multiple graders. Quality assessment using IoU and Mean Distance metrics fed into an iterative outlier detection process using IQR, which was repeated three times to ensure annotation quality. The final dataset comprises 1672 OCT scans with layer annotations and 566 scans with 4698 bounding box annotations across nine disease biomarker classes.

boundary distances between predicted and ground truth layers, providing insight into localization accuracy. This additional metric helps detect systematic errors and offers valuable information on boundary consistency in extracted segmentation masks. By combining IoU and mean distance, a more robust and detailed evaluation of segmentation models handling complex tasks is achieved.

Quality assessment. We used quality metrics to calculate the inter-grader agreement of annotations between multiple graders. Outliers in the annotations were detected and corrected in a second stage. We utilized the interquartile range (IQR) for outlier detection and iteratively improved annotation quality. Each grader had access to multiple images, and once they finished annotating, we compared their results using IQR to a second grader. Outliers were identified between two graders' agreement on any of the five layers if they exceeded the IQR criteria. These outliers were then sent back to the graders for correction. This process was repeated three times to ensure high-quality annotations.

Class	Number of Annotations
Soft Drusen	1798
Hard Drusen	1204
Photoreceptor layer disruption	711
Reticular drusen	326
Soft drusen&PED	302
Hyperfluorescent spots	200
Choroidal fold	151
Geographic atrophy	128
Fluid	123

Table 2. The total number of individual bounding box annotations.

We evaluated the inter-grader agreement among annotations by examining both the Intersection over Union (IoU) and the mean distance metric across all six retinal bands. Using multiple metrics provides a more comprehensive quality assessment, as relying solely on the IoU for retinal layer segmentation might be insufficient due to potential misinterpretations that may arise from disappearing or thinning layers. In cases where two layers become connected as a result of disappearing or thinning layers, the IoU metric may yield a low value, indicating low agreement. However, such low agreement can be attributed to the absence of pixels in the band for comparison, rather than genuine disagreement among the graders. Thus, it is essential to employ multiple metrics to efficiently evaluate the quality of the dataset.

The quality of annotations between each pair, for the entire dataset as well as for specific groups (AMD, DME, and healthy), is illustrated in Fig. 4. We visualized the IoU scores for inter-grader agreement among gradings 1, 2, and 3 across all six retinal bands. These bands include: Background/ILM, ILM/OPL, OPL/IS-OS, IS-OS/IBRPE, IBRPE/OBRPE, and OBRPE/Background. The agreement for the Background/ILM band and OBRPE/Background approaches 1.0 on average across the entire dataset, indicating high consistency among graders. However, for AMD scans, the agreement for IS-OS/IBRPE and IBRPE/OBRPE decreases to approximately 0.8 and 0.75, respectively, suggesting increased difficulty in annotating these layers. Similarly, for DME scans, the disagreement is more pronounced, with median values close to 0.75 for both IS-OS/IBRPE and IBRPE/OBRPE. These findings indicate that annotating IS-OS/IBRPE is particularly challenging for DME scans, while IBRPE/OBRPE presents difficulty in AMD cases.

In the annotation process, comparing mean distances between lines representing layers offers an additional metric as it accounts for disappearing layers or those connected to either the top or bottom of the neighbouring layer. This requires the use of a post-processing method to maintain the continuity of disappearing layers, ensuring that they appear on the following retinal layer. By combining these metrics and our iterative inter-grader agreement process, we achieved high-quality annotations suitable for training and evaluation of our segmentation method.

Usage Notes

We provide ground truth annotations in three different types. First, as layer annotations with pairs of x and y values representing the layers in pixel coordinates. Second, as annotation masks, where every pixel is assigned a grey value representing a band. The values are between 0 and 5. Third, as annotations masks with RGB values assigned to every pixel. The first type can be used for training sequence based models. The second type can be used for training semantic segmentation models (e.g. U-Net based). The third type is similar to the second type but it is intended for visualization and analysis purposes. The data set consists of annotations only, the original images are downloaded from external data sets. We provide scripts to prepare and combine both annotations and images.

Code availability

Source code for data preparation is available within the data set file[47]. It contains the following scripts:

- Prepare_Images_Download.ipynb: Code for downloading original images and preparing the folder structure.
- Prepare_Data_For_Model_Development.ipynb: Code for generating NumPy arrays containing annotations and images for model training.
- Prepare_Data_For_Detection.ipynb: Code for reading and visualizing bounding box annotations.

Detailed descriptions are in the README.txt within the data set file. A Python environment is needed to be able to run the scripts. We also provide a description for creating a virtual environment. The file "requirements.txt" contains all relevant software libraries and packages with corresponding versions.

Received: 16 August 2024; Accepted: 9 December 2024;
Published online: 14 February 2025

References

1. Puliafito, C. A. *et al.* Imaging of Macular Diseases with Optical Coherence Tomography. *Ophthalmology* **102**, 217–229 (1995).
2. Jaffe, G. & Caprioli, J. Optical coherence tomography to detect and manage retinal disease and glaucoma. *Am. J. Ophthalmol.* **137**, 156–169 (2004).

3. Adhi, M. & Duker, J. Optical coherence tomography – Current and future applications. *Curr. Opin. Ophthalmol.* **24**, 213–221 (2013).
4. Drexler, W. & Fujimoto, J. G. State-of-the-art retinal optical coherence tomography. *Prog. Retin. Eye Res.* **27**, 45–88 (2008).
5. Schmidt-Erfurth, U. *et al*. Guidelines for the management of neovascular age-related macular degeneration by the European Society of Retina Specialists (EURETINA). *Br. J. Ophthalmol.* **98**, 1144–1167 (2014).
6. Schmidt-Erfurth, U. *et al*. Guidelines for the management of diabetic macular edema by the European Society of Retina Specialists (EURETINA). *Bulg. Rev. Ophthalmol.* **66**, 3 (2022).
7. Hogarty, D. T., Mackey, D. A. & Hewitt, A. W. Current state and future prospects of artificial intelligence in ophthalmology: a review. *Clin. Exp. Ophthalmol.* **47**, 128–139 (2019).
8. Kal, M., Płatkowska-Adamska, B., Zarębska-Michaluk, D. & Rzymski, P. Reduced vessel density and enlarged foveal avascular zone in the macula as a result of systemic hypoxia caused by SARS-CoV-2. *infection. J. Pers. Med.* **13**, 73 (2023).
9. Pascual-Prieto, J. *et al*. Utility of optical coherence tomography angiography in detecting vascular retinal damage caused by arterial hypertension. *Eur. J. Ophthalmol.* **30**, 579–585 (2019).
10. Lee, J. Y. *et al*. Optical coherence tomography angiography as a potential screening tool for cerebral small vessel diseases. *Alzheimers Res. Ther.* **12**, 73 (2020).
11. Keane, P. A. *et al*. Evaluation of age-related macular degeneration with optical coherence tomography. *Surv. Ophthalmol.* **57**, 389–414 (2012).
12. Browning, D. J., McOwen, M. D., Bowen, R. M. Jr. & Tisha, L. O. Comparison of the clinical diagnosis of diabetic macular edema with diagnosis by optical coherence tomography. *Ophthalmology* **111**, 712–715 (2004).
13. Podkowinski, D. *et al*. Neuroretinal atrophy following resolution of macular oedema in retinal vein occlusion. *Br. J. Ophthalmol.* **103**, 36–42 (2019).
14. De Fauw, J. *et al*. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
15. Kermany, D. S. *et al*. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131 (2018).
16. Gu, Z. *et al*. CE-Net: Context encoder network for 2D medical image segmentation. *IEEE Trans. Med. Imaging* **38**, 2281–2292 (2019).
17. Fang, L. *et al*. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search. *Biomed. Opt. Express* **8**, 2732 (2017).
18. Roy, A. G. *et al*. ReLayNet: Retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional network. *Biomed. Opt. Express* **8**, 3627 (2017).
19. Chiu, S. *et al*. Automatic segmentation of seven retinal layers in SDOCT images congruent with expert manual segmentation. *Opt. Express* **18**, 19413–19428 (2010).
20. Pekala, M. *et al*. Deep learning based retinal OCT segmentation. *Comput. Biol. Med.* **114**, 103445 (2019).
21. Tong, Y., Lu, W., Yu, Y. H. & Shen, Y. Application of machine learning in ophthalmic imaging modalities. *Eye Vis.* **7**, 22 (2020).
22. Schmidt-Erfurth, U. *et al*. Artificial intelligence in retina. *Prog. Retin. Eye Res.* **67**, 1–29 (2018).
23. Dubis, A. M. *et al*. Democratizing deep learning research through large publicly available datasets and tools. *Invest. Ophthalmol. Vis. Sci.* **62**, 1809–1809 (2021).
24. Arikan, M. *et al*. Uncertainty-based deep active learning for retinal layer segmentation. *Invest. Ophthalmol. Vis. Sci.* **62**, 2554–2554 (2021).
25. Li, J. *et al*. Multi-scale GCN-assisted two-stage network for joint segmentation of retinal layers and discs in peripapillary OCT images. *Biomed. Opt. Express* **12**, 2204–2220 (2021).
26. Lujan, B. *et al*. Directional optical coherence tomography provides accurate outer nuclear layer and Henle fiber layer measurements. *Retina* **35**, 1511–1520 (2015).
27. Gong, C., Wang, D., Li, M., Chandra, V. & Liu, Q. KeepAugment: A simple information-preserving data augmentation approach. in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 1055–1064 (IEEE, 2021).
28. Hassan, T., Akram, M., Masood, M. & Yasin, U. Deep structure tensor graph search framework for automated extraction and characterization of retinal layers and fluid pathology in retinal SD-OCT scans. *Comput. Biol. Med.* **105**, 112–124 (2018).
29. Hassan, T., Akram, M. U., Werghi, N. & Nazir, M. N. RAG-FW: A hybrid convolutional framework for the automated extraction of retinal lesions and lesion-influenced grading of human retinal pathology. *IEEE J. Biomed. Health Inform.* **25**, 108–120 (2021).
30. He, Y. *et al*. Retinal layer parcellation of optical coherence tomography images: Data resource for multiple sclerosis and healthy controls. *Data Brief* **22**, 601–604 (2019).
31. Antony, B. J. *et al*. Automated segmentation of mouse OCT volumes (ASiMOV): Validation & clinical study of a light damage model. *PLOS ONE* **12**, e0181059 (2017).
32. Gholami, P., Parthasarathy, M. K., Roy, P. & Lakshminarayanan, V. OCTID citation. Scholars Portal Dataverse V1 (2018).
33. Farsiu, S. *et al*. Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. *Ophthalmology* **121**, 162–172 (2014).
34. Tian, J. *et al*. Real-time automatic segmentation of optical coherence tomography volume data of the macular region. *PLOS ONE* **10**, e0133908 (2015).
35. Chiu, S. J. *et al*. Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema. *Biomed. Opt. Express* **6**, 1172–1194 (2015).
36. Chiu, S. J. *et al*. Validated automatic segmentation of AMD pathology including drusen and geographic atrophy in SD-OCT images. *Invest. Ophthalmol. Vis. Sci.* **53**, 53–61 (2012).
37. Melinščak, M., Radmilovic, M., Vatavuk, Z. & Loncaric, S. Annotated retinal optical coherence tomography images (AROI) database for joint retinal layer and fluid segmentation. *Automatika* **62**, 375–385 (2021).
38. Morales, S. *et al*. Retinal layer segmentation in rodent OCT images: Local intensity profiles & fully convolutional neural networks. *Comput. Methods Programs Biomed.* **198**, 105788 (2021).
39. Kermany, D. S., Zhang, K. & Goldbaum, M. H. Large dataset of labeled optical coherence tomography (OCT) and chest X-ray images. Mendeley Data v2 (2018).
40. Kulyabin, M. *et al*. OCTDL: Optical coherence tomography dataset for image-based deep learning methods. IEEE Dataport (2023).
41. Rasti, R., Rabbani, H., Mehridehnavi, A. & Hajizadeh, F. Macular OCT classification using a multi-scale convolutional neural network ensemble. *IEEE Trans. Med. Imaging* **37**, 1024–1034 (2018).
42. Rasti, R., Rabbani, H., Mehridehnavi, A. & Hajizadeh, F. Dataset for OCT classification (50 normal, 48 AMD, 50 DME). Available at: https://hrabbani.site123.me/available-datasets/dataset-for-oct-classification-50-normal-48-amd-50-dme (2017).
43. Hitachi-Automotive-And-Industry-Lab. Semantic segmentation editor. GitHub https://github.com/Hitachi-Automotive-And-Industry-Lab/semantic-segmentation-editor (2022).
44. Willoughby, J. *et al*. Object detection on medical images with the aid of contrastive gated attention. *Invest. Ophthalmol. Vis. Sci.* **63**, 2998–F0268 (2022).
45. Tan, M. & Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. in *Proceedings of the 36th International Conference on Machine Learning* 6105–6114 (PMLR, 2019).
46. Yakubovskiy, P. Segmentation models. *GitHub* https://github.com/qubvel/segmentation_models (2019).
47. Arikan, M. OCT5k: A dataset of multi-disease and multi-graded annotations for retinal layers. UCL Research Data Repository v4. https://doi.org/10.5522/04/22128671.v4 (2023).

Acknowledgements
The authors acknowledge the use of the Joint Academic Data science Endeavour (JADE) Tier 2 computing facility funded by the Engineering and Physical Sciences Research Council (EPSRC), UK. This research was supported by the National Institute for Health and Care Research (NIHR) Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust, UCL Institute of Ophthalmology and UCL Institute of Healthcare Engineering. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. This work was supported by National Institutes of Health Care Grant (EY014800), and an Unrestricted Grant from Research to Prevent Blindness, New York, NY, to the Department of Ophthalmology & Visual Sciences, University of Utah.

Author contributions
M.A. built the annotation platform, compiled the data set, developed the baseline segmentation method and wrote the manuscript with support from A.D. and W.L. J.W. prepared the object detection labels and reviewed the manuscript. W.L. conceived the annotation procedure to reduce bias, helped developing the baseline method and reviewed the manuscript. A.D. planned the overall annotation, monitored the annotation process, selected OCT volumes and reviewed the manuscript. S.O., F.S., A.M., H.A., A.H., M.B., H.F., M.C., A.F., D.P., M.C., D.A., M.Z. and G.R. annotated scans, gave feedback for improving the platform and the manuscript.

Competing interests
The authors declare the following competing interests: A.M.D. has a commercial relationship with DeepEye Medical GmbH. Additionally, A.M.D., W.L., J.W. and M.A. are inventors on patents assigned to UCL Business Ltd., including WO-2023047118-A1, WO-2022208060-A3, and WO-2023047117-A1.

Additional information
Correspondence and requests for materials should be addressed to A.M.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.