Published in *Computer methods and programs in biomedicine*, September 2019, vol. 178, pp. 181-189 which should be cited to refer to this work
DOI:10.1016/j.cmpb.2019.06.016

Classification of diabetes-related retinal diseases using a deep learning approach in optical coherence tomography

Oscar Perdomo[a], Hernán Rios[b], Francisco Rodríguez[b], Sebastián Otálora[c,d], Fabrice Meriaudeau[e], Henning Müller[c,d], Fabio A. González[a,*]

[a] *MindLab Research Group, Universidad Nacional de Colombia, Edificio 453, Laboratorio 207, Bogotá, Colombia*
[b] *Fundación Oftalmológica Nacional, Bogotá, Colombia*
[c] *University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland*
[d] *University of Geneva, Geneva, Switzerland*
[e] *Université de Bourgogne Franche Comté, ImvIA EA7535/IFTIM*

Abstract

Background and objectives: Spectral Domain Optical Coherence Tomography (SD-OCT) is a volumetric imaging technique that allows measuring patterns between layers such as small amounts of fluid. Since 2012, automatic medical image analysis performance has steadily increased through the use of deep learning models that automatically learn relevant features for specific tasks, instead of designing visual features manually. Nevertheless, providing insights and interpretation of the predictions made by the model is still a challenge. This paper describes a deep learning model able to detect medically interpretable information in relevant images from a volume to classify diabetes-related retinal diseases.

Methods: This article presents a new deep learning model, OCT-NET, which is a customized convolutional neural network for processing scans extracted from optical coherence tomography volumes. OCT-NET is applied to the classification of three conditions seen in SD-OCT volumes. Additionally, the proposed model includes a feedback stage that highlights the areas of the scans to support the interpretation of the results. This information is potentially useful for a medical specialist while assessing the prediction produced by the model.

Results: The proposed model was tested on the public SERI-CUHK and A2A SD-OCT data sets containing healthy, diabetic retinopathy, diabetic macular edema and age-related macular degeneration. The experimental evaluation shows that the proposed method outperforms conventional convolutional deep learning models from the state of the art reported on the SERI+CUHK and A2A SD-OCT data sets with a precision of 93% and an area under the ROC curve (AUC) of 0.99 respectively.

Conclusions: The proposed method is able to classify the three studied retinal diseases with high accuracy. One advantage of the method is its ability to produce interpretable clinical information in the form of highlighting the

regions of the image that most contribute to the classifier decision.

Keywords: Optical coherence tomography, deep learning models, interpretability, retinal diseases, medical findings.

1. Introduction

Ophthalmic diseases related to Diabetes Mellitus (DM) are characterized by a vascular permeability of retinal vessels with fluid accumulating in retinal layers [1]. Diabetic Retinopathy (DR) and Diabetic Macular Edema (DME) are two non-exclusive complications that affect the visual field [2]. The diagnosis of these complications is not an easy task, since edema can occur in subjects with and without DM at any stage of DR, with similar symptoms but with different treatment strategies and associated costs [3]. Age-related Macular Degeneration (AMD) is linked to macular changes derived from non-modifiable and modifiable risk factors. The diagnosis is based on typical changes related to aging and visual loss and prognosis is related to the severity of the either geographic atrophy or choroidal neovascular membrane [4].

The Spectral Domain Optical Coherence Tomography (SD-OCT) is a widely accepted noninvasive imaging approach that contains images of the depth of the retina through a set of B-scan (2D images) used to detect abnormalities among the ten retinal layers with an accurate diagnosis of retinal disorders [5]. A typical ophthalmological examination of the retina may include an analysis of eye fundus images and in some cases SD-OCT to locate retinal vascular damage and changes in choroidal thickness [6]. The DR and DME diagnoses are performed by looking for the presence of microaneurysms, intraretinal hemorrhages, exudates and edema [7, 8, 9]. The evaluation of the thickness of the neurosensory retina, retinal pigment epithelium, and choroid are analyzed independently for the AMD diagnosis [10, 11].

Automatic image analysis methods based on machine learning have shown to be a valuable tool to support medical decision making [12, 13]. In particular, deep neural network methods have been explored in several medical domains exhibiting promising results. The results of deep neural networks include: the detection of red lesions in fundus images [14], prediction of breast, lung and stomach cancers using RNA-sequence data [15], early diagnosis of Alzheimer's disease using CT brain images [16] and the recognition of emotions using multimodal physiological signals [17].

Deep learning methods applied to SD-OCT presented outstanding results in automatic segmentation and disease classification tasks. For a segmentation task, the state of the art presents an overall Dice coefficient (mean of all tissues) ranging between 0.90 and 0.95 using known architectures such as VGG [18], U-Net [19, 20, 21, 22] or DenseNet [23, 24]. The classification of SD-OCT volumes has mainly focused on two approaches: (1) the manual or automatic feature extraction combined with ensemble classifiers, and (2) the use of end-to-end deep learning models.

*Corresponding author
e-mail: fagonzalezo@unal.edu.co (Fabio A. González)
URL: https://sites.google.com/a/unal.edu.co/mindlab/ (Fabio A. González)

This paper presents a deep learning-based method with a feedback stage for automatic classification of B-scans inside a volume for three retinal diseases. The method is able to automatically identify visual patterns associated with several pathologies and use them to make accurate predictions. The model has the ability to highlight the patterns in the input image, allowing the expert to better understand the model prediction. The remainder of this article is organized as follows: in Section 2 the main work for retinal disease classification using SD-OCT volumes is summarized. Then, the volume preprocessing and the convolutional neural network-based model architecture are presented in Section 3. The data sets and baseline models used are described in Section 4. The experimental results are reported in Section 5. Finally, Section 6 discusses the outcomes and finishes with conclusions.

2. Related work

The end-to-end OCT-NET model was tested on a data set that contains 32 SD-OCT volumes with healthy and DME patients commonly known as the SERI (Singapore Eye Research Institute) data set as explained in details in subsection 4.1. In this previous work, the OCT-NET model obtained an outstanding performance using a leave-one-patient-out evaluation methodology with an accuracy of 93.75 ± 3.125% and a sensitivity and a specificity of 93.75% [25]. This paper presents an extended version that addresses three main challenges: (1) the qualitative evaluation of B-scans to highlight medical findings using a visualization stage; (2) the quantitative evaluation of SD-OCT volumes with three retinal diseases from two OCT scanners, and (3) the medical feedback of quantitative and qualitative evaluations to validate the usefulness of the methodology.

The main work reported on the SERI data set is characterized by using deep learning architectures pre-trained on ImageNet[1] combined with ensemble classifiers. First, Awais et al. [26] presented a method that used block-matching and 3D filtering (BM3D) for removing the speckle noise in SD-OCT. The new filtered volumes fed a pre-trained VGG-16 with a k-Nearest Neighbors (kNN) algorithm to classify features from the three dense layers with an accuracy, sensitivity, and specificity of 93%, 87% and 100% respectively. In a similar way, Chan et al. [27] designed a method that applies a BM3D filter and saturation removal. The processed volumes are then used as input of three pre-trained architectures known as Alexnet, VGG and GoogleNet. The last convolutional layers of each model are fused and a feature space reduction is performed using Principal Component Analysis. The volumes are classified using a Support Vector Machine (SVM) with a precision, sensitivity and specificity of 93.75%. Finally, Kamble et al. [28] proposed the fusion of residual connections with an inception architecture termed as inception-ResNet-v2. This model used as an input filtered volumes with a BM3D filter stage and presented a performance of 100% in accuracy, specificity and sensitivity.

The most representative work classifying SD-OCT volumes on the A2A SD-OCT data set is mainly reported in three papers [29, 30, 31]. Sun et al. [29] manually cropped patches based on the annotation of interest points to calculate a Histograms of Oriented Gradient (HOG) and merged them as the training set. Then, Principal Component

[1] http://www.image-net.org/

Analysis (PCA) was performed for reducing the length of the HOG features. Finally, a multiple instance SVM classifier was trained with the PCA-transformed patch representation and tested to classify volumes on the test data set obtaining an accuracy, sensitivity and specificity of 94.4%, 96.8% and 92.1% respectively.

Venhuizen et al. [30] developed an unsupervised clustering stage to extract interest points in 31 B-scans per volume centered at the fovea of 284 SD-OCT volumes as the training data set. The number of 9×9 patches is reduced by the application of a uniform subsampling by a factor of 8. Then, the patches are normalized to zero mean and unit variance before the extraction of 9 principal components through PCA. A bag of words is created using k-means clustering with an experimental value of $k = 100$ on the total set of PCA-transformed patches. Finally, the unsupervised clustering is combined with a supervised training stage that uses a random forest classifier with a number of trees set to 100 trained to differentiate healthy subjects from AMD subjects. The performance on the test data set after classifying 50 AMD and 50 healthy subjects was an AUC of 0.984.

Chakravarty et al. [31] designed a two-stage retinal atlas for macular SD-OCT volumes that comprises a pre-processing and a classification stages. First, the pre-processing step resized the images to a pixel dimensions of 3.6 μm by 8.6 μm. A denoising and intensity standardization is applied to reduce the speckle noise in the volumes. Moreover, a retinal curvature flattening of the SD-OCT volumes is performed, where each B-scan is flattened and aligend across the volume. Finally, the Region of Interest (RoI) by SD-OCT volume is defined to a set of 31 B-scans with the 108 axial scans centered at the macula, where the histogram is calculated and concatenated across all the RoI's. The binary classification stage is done using a linear SVM presenting an accuracy and an AUC of 98% and 0.996 respectively for AMD classification.

The state of the art for classifying real-world scans is mainly focused in three deep learning approaches. De Fauw et al. [32] trained with two private data sets a two-stage deep learning-based pipeline: a deep segmentation network with a three-dimensional U-Net architecture and a deep classification network to predict the diagnosis probability and the referral suggestions using the segmentation. The deep segmentation network was trained using 877 SD-OCT volumes acquired by Moorfields Eye Hospital with a Topcon 3D OCT device, where the three most representative slices were manually segmented in a detailed tissue-segmentation map with 15 classes including anatomy, pathology and image artifacts. The classification network was trained using the 43 most representative slices of $14,884$ SD-OCT volumes obtained from $7,621$ patients referred by the experts as subjects with symptoms suggestive of macular pathology. The model was tested with 997 patients obtaining an area under the ROC curve of 99.21%. Lee et al. [33] used a VGG-16 convolutional neural network applied on a private data set for the classification of normal and AMD. The deep learning model receives as an input a scan with a size of 192×124 and performed 13 convolutional layers with an incremental number of filters and 3 dense layer to classify the two classes. The image database was acquired using a Heidelberg Spectralis OCT device with $80,839$ images in the training set and the test set contains $20,163$ images. The results at image level presented an area under the ROC curve of 92.78% with an accuracy of 87.63%, a sensitivity of 84.63% and a specificity of 91.54%. Additionally, an occlusion test identified the RoIs with the areas contributing most to the deep neural network's probability. Finally, Kermany et al. [34] used a pretrained Inception

V3 from ImageNet to predict four classes: Normal, Choroidal Neovascularization (CNV), DME and Drusen. The method was trained using a public data set acquired by Heidelberg Spectralis OCT device with $108,312$ images for training and tested in $1,000$ with 250 per class. The best results on the test set presented an accuracy of 96.6%, a sensitivity of 97.8%, and a specificity of 97.4%. In addition, a sliding window of 20×20 was systematically moved across 491 images to record the probabilities of the disease.

Although previous work reported very good results in the classification task, the performance of these methods is crucially dependent of the manual extraction of RoIs and in some case limiting the number of scans from the SD-OCT volumes. The proposed model provides the highlighted areas in all scans into volumes with a validation performed by two ophthalmology experts. Distinctively from previous work, our approach automatically classifies AMD but also produces useful medical information at qualitative and quantitative levels inside an SD-OCT volume to support medical decision making in the diagnosis of AMD.

3. Methods

This section presents the details of the SD-OCT classification model based on deep neural networks and more specifically OCT-NET. The method comprises six stages as shown in Figure 1 and it is available in a repository of Github[2]. The first stage (1) receives a raw SD-OCT volume with speckle noise that hinders layers and abnormalities among the layers as an input. Then, the volume preprocessing stage (2) makes the detection of the Internal Limiting Membrane (ILM) and the Retinal Pigment Epithelium (RPE) layers, to resize the volumes in order to crop the relevant raw pixels into the volumes as presented in subsection 3.1. Furthermore, the OCT-NET model (3) performs the feature extraction to classify each B-scan as healthy or non-healthy. Simultaneously, the disease classification stage (4) calculates with a majority rule the prediction for the volume as explained in detail in subsection 3.2. Then, the Class Activation Map (CAM) visualization stage (5) allows to highlight the relevant zones of the scans used by the OCT-NET model to classify a specific retinal disease as reported in subsection 3.3. Finally, the expert feedback stage (6) evaluates the provided information in the disease classification and the CAM visualization stages to qualitatively validate the obtained results.

3.1. SD-OCT volume preprocessing

A spectral domain optical coherence tomography is a volumetric array $V(n, a, b)$ that can be defined as a set n of 2D-images called B-scans or cross-sectional scans $I \in \mathbb{R}^{a \times b}$, with a corresponding label $l \subset \{Healthy,\ DME,\ DR - DME\ and\ AMD\}$. The input for the customized OCT-NET was set for scans with size of $224 \times 224 \times 1$ as described in subsection 3.2. Therefore, a set of transformations are needed for automatically extracting a RoI per scan in the SD-OCT volume.

[2]https://github.com/Ojperdomoc/OCT-NET.git

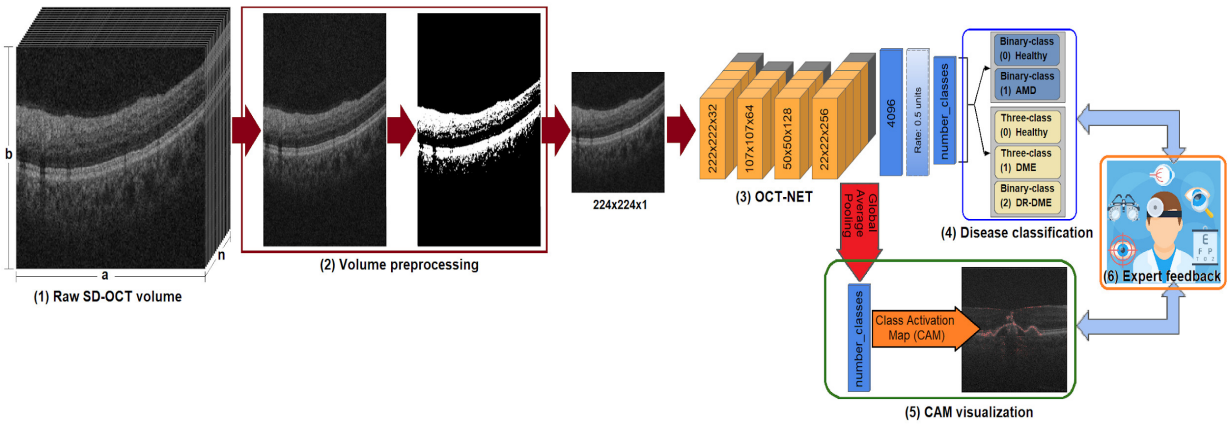

Figure 1: Overview of the six-stage proposed deep learning model for retinal disease classification. (1) The raw SD-OCT volume for a database; (2) The volume preprocessing stage to resize and crop to an input size; (3) the OCT-NET model to extract features from all the scans in a volume; and the interpretability three-stage (4,5,6) to classify and evaluate qualitative and quantitative the scans and the SD-OCT volumes.

A median filter was applied to SD-OCT scans with a threshold to differentiate speckle noise from retinal layers as reported in [25, 35]. First, the RoI was automatically detected using a median filter with a kernel size of 3×3 and a threshold of 0.5 to highlight the top layer (ILM) and bottom layer (RPE) in the volume as shown in stage (2) of Figure 1.

Each image cropped without scaling in such a way that the resulting image fully contains the RoI. This is independently done for each B-scan hence that the process is not affected by a pronounced tilt. Finally, these cropped images are resized keeping aspect ratio to ensure the relevant information in a volume dimension of $V_{input} \in \mathbb{R}^{n \times 224 \times 224}$.

3.2. OCT-NET model

OCT-NET is a customized Convolutional Neural Network (CNN) inspired by the VGG model reported by Simonyan and Zisserman [36]. The proposed model is based on the combination of convolutional and max-pooling layers in four sub-blocks that are responsible for the feature extraction in the CNN and the remaining layers conforming the classification sub-block to classify a scan from an OCT-volume. In summary, the OCT-NET model contains 10 convolutional layers, 4 max-pooling layer, 2 fully connected layers, and 1 dropout layer as shown in detail in Table 1.

The input layer receives an image with a size of $224 \times 224 \times 1$ as reported in subsection 3.1. The number of filters of the convolutional layers in the four sub-blocks is inspired by the VGG model [36], with the difference that OCT-NET has a number of filters f_n defined by an arithmetic series, as described in Eq. (1) as follows:

$$f_n = f_0 + 32 * (n - 1) \tag{1}$$

where the parameter $f_0 = 32$ and n is the number of sub-blocks with $1 \leq n \leq 4$. The cascading of four blocks of convolutional and max-pooling layers provides a translation invariance and a reduction of dimensionality: by applying

Table 1: Structure of OCT-NET with the parameter layer, output shape and trainable parameters of each layer.

N	Layer	Output shape	Number of parameters
0	Input	$224 \times 224 \times 1$	0
1	Conv2D (kernel size = 3×3)	$222 \times 222 \times 32$	320
2	Conv2D (kernel size = 3×3)	$220 \times 220 \times 32$	9248
3	Conv2D (kernel size = 3×3)	$218 \times 218 \times 32$	9248
4	MaxPooling2D (pool size = 2×2)	$109 \times 109 \times 32$	0
5	Conv2D (kernel size = 3×3)	$107 \times 107 \times 64$	18496
6	Conv2D (kernel size = 3×3)	$105 \times 105 \times 64$	36928
7	MaxPooling2D (pool size = 2×2)	$52 \times 52 \times 64$	0
8	Conv2D (kernel size = 3×3)	$50 \times 50 \times 128$	73856
9	Conv2D (kernel size = 3×3)	$48 \times 48 \times 128$	147584
10	MaxPooling2D (pool size = 2×2)	$24 \times 24 \times 128$	0
11	Conv2D (kernel size = 3×3)	$22 \times 22 \times 256$	295168
12	Conv2D (kernel size = 3×3)	$20 \times 20 \times 256$	590080
13	Conv2D (kernel size = 3×3)	$18 \times 18 \times 256$	590080
14	MaxPooling2D (pool size = 2×2)	$9 \times 9 \times 256$	0
15	Dense	4096	84938752
16	Dropout (rate = 0.5)	4096	0
17	Dense	number of classes	8194

a set of f_n learned filters with kernel size of 3×3 and stride of 1×1 and eliminating non-maximal values with pool size of 2×2 and stride of 2×2.

The classification sub-block is composed of three layers: one fully-connected layer with 4096 neurons, one dropout layer with a fraction of deactivation of units during training of 0.5, and a final fully-connected layer with *number of classes* as the number of neurons. The dropout layer allows to learn with different neurons the same information improving the generalization of the model and the number of neurons for the final fully-connected layer is set to 2 or 3 for binary and three-class data sets respectively. Additionally, the disease classification of one SD-OCT volume was determined using a majority rule, as such the volume was affected by the class that was the most preponderant among the B-scans.

3.3. Class activation map

The Class Activation Map (CAM) is defined as the sum of the weighted activation maps generated for each image at different spatial locations. The main use of a CAM focuses on the validation of a CNN model that indicates the discriminative image regions for a particular category. Thus, the CAM block adds a Global Average Pooling (GAP) after the last convolutional layer in the CNN model for obtaining an accurate discriminative localization as reported by Bolei et al. [37] and Selvaraju et al. [38]. The CAM highlights the magnitude of the activation at the spatial grid (x, y) to classify an image to class c. The CAM for class c is defined by w_k^c as the weight corresponding to class c for unit k applied to an input image $f_k(x, y)$ described in Equation 2 as follows [37]:

$$M_c(x, y) = \sum_k w_k^c f_k(x, y) \tag{2}$$

4. Experimental evaluation

4.1. SERI+CUHK dataset

The Singapore Eye Research Institute (SERI) database contains 32 SD-OCT volumes with 16 control and 16 DME SD-OCT volumes. Similarly, the Chinese University of Hong Kong (CUHK) database contains 43 SD-OCT volumes with 4 DME and 39 DR-DME SD-OCT volumes. The two data sets were combined into one three-class data set termed in this paper as SERI+CUHK data set. The SERI-CUHK data set was acquired with a CIRRUS SD-OCT device [3] and labeled by certified expert graders as control, DME and DR-DME volumes, according to findings among the retinal layers as shown in Fig 2.

The inclusion criterion was the presence of abnormal retinal thickening, hard exudates, intraretinal cystoid space formation and subretinal fluid among the retinal layers of working-age adult subjects. Finally, each SD-OCT volume contains 128 cross-sectional scans with a resolution of $512 \times 1,024$ pixels. The data set was cropped and resized (keeping the aspect ratio) to a dimension of $128 \times 224 \times 224$ as discussed in subsection 3.1.

[3] https://www.zeiss.com/meditec/int/products/oct-optical-coherence-tomography.html

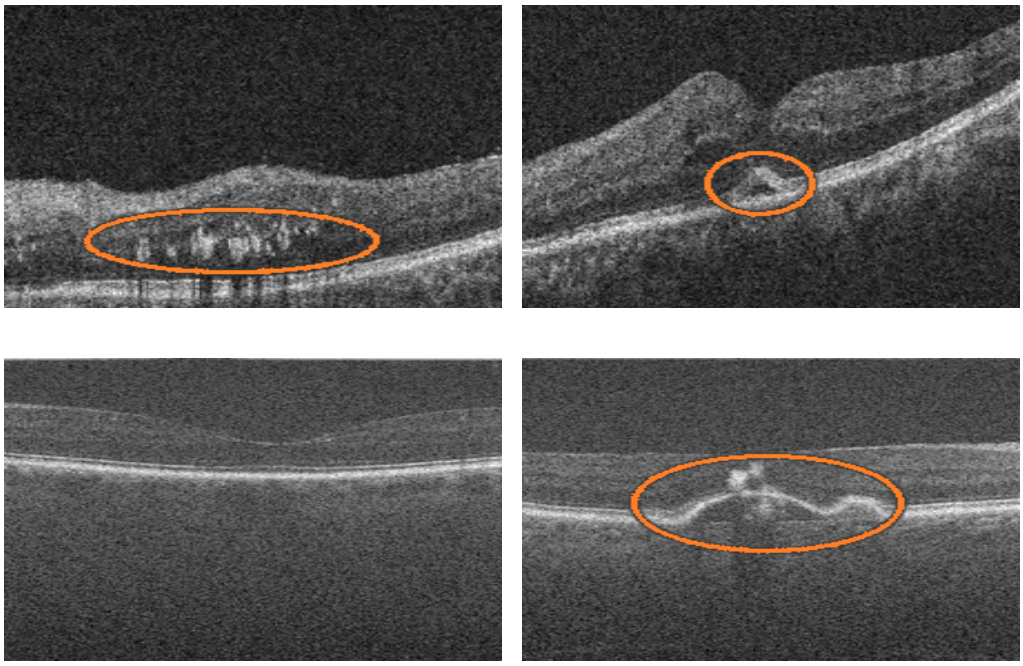

Figure 2: [Top-left] DME SD-OCT scan with hyper-reflective material in the middle layers of the retina, likely by exudates, and [Top-right] DR-DME SD-OCT scan with a retinal pigment epithelium detachment.[Bottom-left] Normal SD-OCT scan and [Bottom-right] AMD SD-OCT scan with a drusenoid detachment with migration of pigment to the inner layers of the retina.

4.2. A2A SD-OCT data set

The A2A SD-OCT is a binary data set from the Age-Related Eye Disease Study 2 (AREDS2) also known as Duke data set [39]. The images from the A2A SD-OCT study obtained the informed consent from all subjects and it was approved by the institutional review boards of the 4 A2A SD-OCT clinics: Devers Eye Institute, Duke Eye Center, Emory Eye Center, and National Eye Institute.

The Duke data set was acquired using imaging systems from Bioptigen, Inc (Research Triangle Park, NC) as shown in Fig. 2. The classification of each volume was done by certified SD-OCT readers. The inclusion criteria were defined as subjects between 50 and 85 years of age, exhibiting intermediate AMD with large drusen ($> 125mm$) in both eyes or large drusen in one eligible eye and advanced AMD in the fellow eye, with no history of vitreoretinal surgery or opthalmic surgery. The Duke data set contains 384 SD-OCT volumes: 269 AMD and 115 control or normal eyes, with 100 B-scans per volume and a resolution of $1,000 \times 512$. The data set was resized to a volume dimension of $100 \times 224 \times 224$ as reported in subsection 3.1.

4.3. Experimental setup

The OCT-NET model was trained with random initialization weights using the Adam optimizer. The batch size, learning rate and number of epochs were experimentally set to 16, $1e-5$ and 10 respectively for all experiments

as reported in a previous article [25]. Moreover, the classification of one SD-OCT volume was determined using a majority rule as explained in subsection 3.2.

The SERI+CUHK data set was randomly split in a stratified way into three independent data sets with 60%, 10% and 30% for training, validation and testing respectively as presented in Table 2. On the other hand, the Duke data set was randomly split into 3 independent data sets with 67% and 10% for training and validation respectively. The remaining 23% corresponds to the test data set, with 50 AMD and 50 Healthy volumes as reported in Table 2.

Table 2: Retinal disease data sets used for training, validation and testing in the experimental evaluation.

Data set	Training	Validation	Test
SERI+CUHK	45 SD-OCT	8 SD-OCT	22 SD-OCT
Duke	246 SD-OCT	38 SD-OCT	100 SD-OCT

4.4. Baseline models and performance metrics

The work proposed by Venhuizen et al. [30], Chakravarty et al. [31] and Kermany et al. [34] were chosen as baseline models applied on the Duke data set as explained in Section 2. In addition, the methods reported by Awais et al. [26] and Kermany et al. [34] were chosen as baseline models for the SERI-CUHK data set. Additionally, we performed a qualitative evaluation for the interpretability stage according to the ability of the proposed model to highlight medical findings in the scans. In this test, 40 SD-OCT volumes from the Duke test data set were randomly split with 20 healthy and 20 AMD samples. Two retina specialists manually labeled each B-scan of this subset without taking into account the given volume label. Finally, the two experts assessed the generated CAM visualization plus the individual prediction from each scan in a volume.

The proposed model was implemented with Keras using a Theano backend on a GeForce GTX TITAN X from NVIDIA. The loss and accuracy metrics were monitored on the training and validation data, and the best performance in the validation set was assessed on the test data set presented in Tables 3 and 4. OCT-NET was evaluated on the test set of the Duke data using accuracy, sensitivity, specificity as performance metrics defined as follows in Equations 3-5. In addition, the AUC was calculated according to the probability that our classifier will rank a randomly selected positive case higher than a randomly chosen negative case.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

OCT-NET was evaluated on the SERI+CUHK test data set using precision, f-score (macro) and Kappa coefficient as multi-class performance metrics (defined in Equations 6-8). Recall was another performance metric evaluated on the SERI+CUHK test data set. It is defined as the true positive rate or sensitivity as explained in Equation 4.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$f - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{7}$$

$$Kappa\ coefficient = \frac{p_o - p_e}{1 - p_e} \tag{8}$$

where,

- TP = True positive (both the ground-truth and predicted are disease class)
- TN = True Negative (both the ground-truth and predicted are healthy class)
- FP = False Positive (predicted as disease class but the ground-truth is healthy class)
- FN = False Negative (predicted as healthy class but the ground-truth is disease class)
- p_o = Probability of correct classification
- p_e = Probability of chance agreement

These performance measures were chosen so that the results can be compared with those reported by the state of the art.

5. Results

5.1. Volume classification performance

The performance classification was reported for the Duke and the SERI+CUHK databases explained in detail in Section 4. Moreover, we tested the performance metrics of the OCT-NET model using the hyper-parameters and monitoring the loss and the accuracy during training as shown in Figure 3. The training was set to 10 epochs as it

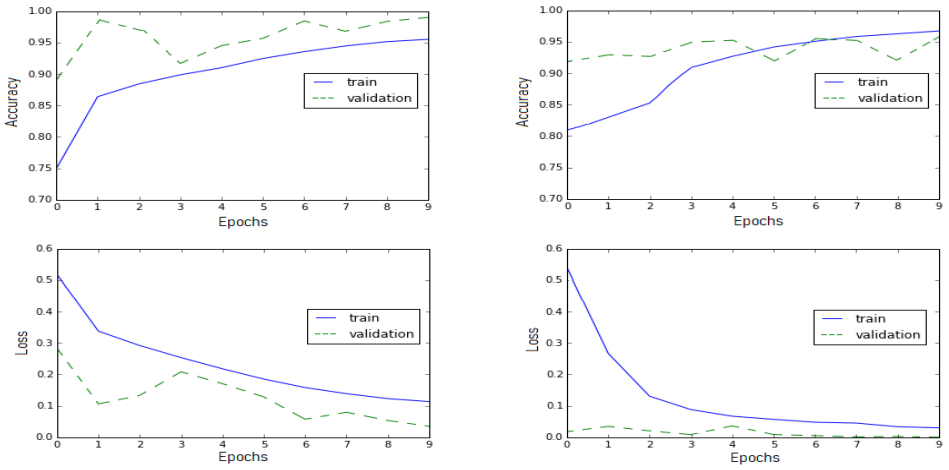

Figure 3: The monitoring of the accuracy and loss during training for the two experiments: [Left] SERI+CUHK, and [Right] Duke data sets. The blue solid and the green dashed lines represent the training and validation sets respectively.

obtained high accuracy while limiting the training time that is on average to 60 minutes per epoch. The computational time to evaluate a single B-scan of the test set from Duke was 0.33 ms, and 0.28 ms in SERI+CUHK.

For the Duke database, the performance of OCT-NET applied on the test data set is reported in Table 3. Additionally, the performance of the proposed method was compared with the main related works using this database as presented in subsection 4.3. The OCT-NET was modified into three architectures with different values in Dropout (DO) layer and the last dense (D2) layer.

The OCT-NET presented a similar AUC metric that method proposed by Chakravarty et al. [31], but outperforms baseline methods in sensitivity, specificity and accuracy as reported in Table 3.

Table 3: Performance measures of the baseline methods and the proposed method on the test data (Duke), bold values show the best score for each architecture.

Model	Sensitivity	Specificity	Accuracy	AUC
Venhuizen et al. [30]	0.96	0.92	0.94	0.984
Chakravarty et al.[31]	0.97	0.98	0.98	**0.99**
Kermany et al.[34]	0.98	0.89	0.94	0.94
OCT-NET	**0.99**	**0.99**	**0.99**	**0.99**
OCT-NET with DO=0.25	0.89	0.89	0.89	0.89
OCT-NET without DO	0.90	0.88	0.88	0.88
OCT-NET with D2=2048	0.95	0.95	0.95	0.95

The precision, recall, f-score Kappa coefficient and AUC were calculated to assess the performance of the pro-

posed model applied to the SERI+CUHK data set, as reported in Table 4. For the SERI+CUHK data set, the best performance of the proposed model on the test data is reported in Table 4. Furthermore, we compared the performance of the baseline model reported by **and** [26] using the three output dense layers (D1, D2 and D3) with three different ensemble classifiers: Decision Trees (DT), and KNNs with $K = 1$ and $K = 3$ applied on the test set.

The OCT-NET architecture presented the best performance and it outperforms baseline methods in precision, recall, f-score and Kappa coefficient as shown in Table 4.

Table 4: Performance metrics of OCT-NET on the test data (SERI+CUHK).

Model	Precision	Recall	f-score (macro)	Kappa coefficient	AUC
D1 with DT (depth=100) [26]	0.69	0.70	0.69	0.42	0.5
D2 with KNN (K=1)[26]	0.62	0.65	0.63	0.27	0.5
D3 with KNN (K=3)[26]	0.62	0.65	0.63	0.27	0.57
Kermany et al.[34]	0.91	0.78	0.74	0.59	**0.86**
OCT-NET	**0.93**	**0.83**	**0.85**	**0.71**	**0.86**

5.2. Qualitative analysis of CAM

The CAM visualization stage for the proposed model was validated according to the ability of locating medical findings that allow to highlight different retinal disorders as reported in subsection 3.3. The CAM output of the proposed model for the AMD class was highlighted in red with the corresponding medical findings of the ophthalmologist outlined in green as shown in Figure 4. Besides this, the ability of the proposed model to predict the condition of individual scans belonging to an AMD SD-OCT volume, compared with the diagnosis performed by an ophthalmology expert is presented in Figure 5.

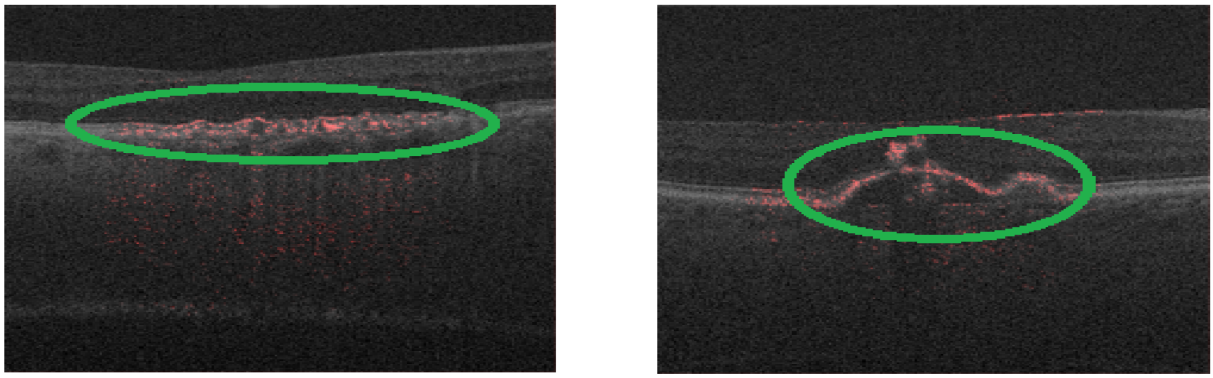

Figure 4: SD-OCT scans for subjects with AMD. [left] three large lesions on the outer layers and, [right] a drusenoid detachment with migration of pigment to the inner layers of the retina.

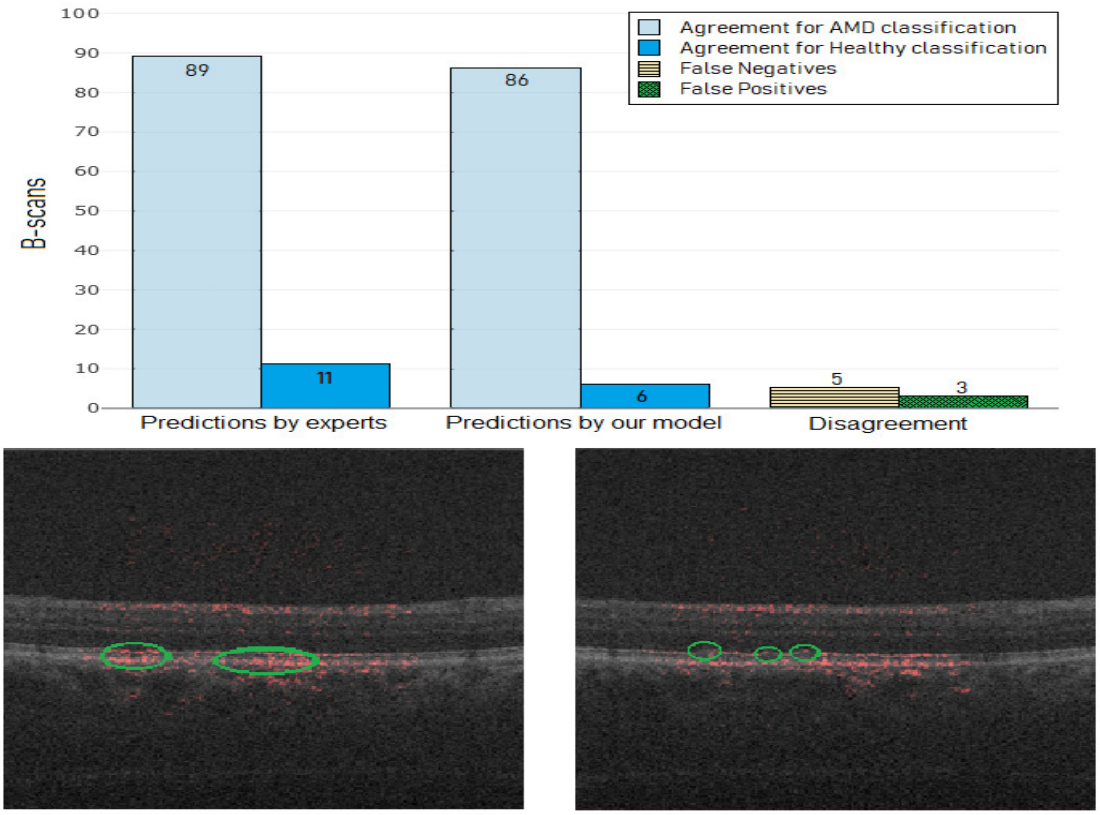

Figure 5: [Top] Classification of B-scans from an SD-OCT volume of an AMD subject by an ophthalmologist and predictions by the proposed model. [Left] False positive or a misclassified B-scan as AMD class due to an RPE layer presenting hyperreflectivity in these areas (green circles). This results in the proposed method incorrectly highlighting the areas like drusen or an RPE elevation. [Right] False negative or a misclassified B-scan as healthy class due to the RPE elevation not having enough hyperreflectivity (green circles); probably the proposed method is not able to detect these tiny drusen.

5.3. Individual B-scan classification

The SD-OCT data sets are commonly labeled to a volume level despite the retinal disease is present in a range of B-scans. This chal elgenmotivated the evaluation of the proposed model to detect healthy and non-healthy scans regardless of the global labels of the volumes. Thus, the proposed method was validated with a subset of 4000 labeled B-scans annotated by the experts as presented in subsection 4.3. Table 5 presents the confusion matrix for the 40 SD-OCT volumes from the Duke test data set as explained in Section 4. The major diagonal is equivalent to the agreements in the classification of the two classes. Otherwise, the subdiagonal represents the erroneous classification of the proposed model. The overall accuracy in the prediction of the two classes was of 89% with a precision of 93% in the detection of AMD scans.

Table 5: Confusion matrix describing the agreement in the predictions for healthy and AMD classes with true negatives (TN) and true positives (TP) respectively. The disagreement between the two classes is measured with false positives (FP) and false negatives (FN).

		Prediction by model	
		Healthy	**AMD**
GroundTruth	**Healthy**	TN=2001	FP=332
	AMD	FN=125	TP=1542

6. Discussion and conclusion

OCT-NET outperforms the state of the art methods for AMD diagnosis reported in [30, 31] in sensitivity, specificity, accuracy but it presents a similar AUC compared to the model proposed by Chakravarty et al. [31] as shown in Table 3. The main two advantages of the proposed model compared to the two-stage method [30] and the retinal atlas [31] are the automatic classification of raw scans without manual annotation of interest points or regions and the generation of qualitative and quantitative information to support medical decision making in a diagnosis of AMD as presented in Figure 4.

The experimental results of OCT-NET on the SERI+CUHK data set overcome the performance of the approach presented by Kermany et al. [34] in precision, recall, f-score and Kappa coefficient as reported in Table 4. The proposed method shows an outstanding performance compared to the Inception-v3 pretrained with weights from ImageNet [34] without requiring a large database for training or selecting a limited numbers of scans with the condition by a patient. Finally, the Kappa coefficient or inter-rater agreement presented a substantial level of agreement of 0.71 between the model and the expert for the classification of healthy, DME and DR-DME SD-OCT volumes as reported in Section 5.

The global label of an SD-OCT volume is used without questioning local labels for each scan or the specific range of scans that contain the retinal disorder. We evaluated the prediction of B-scans belonging to an SD-OCT volume inspired in the manual classification performed by an ophthalmology expert as presented in subsection 5.3. The experimental results shown an agreement of 92.5% for AMD and 85.8% for healthy scans of a total of 4000 scans compared to the manually labeled scans, as reported in Table 5. In addition, the range of scans with the retinal disorder in the volume, and the highlighted areas by CAM stage present a strong agreement with the delineations of the ophthalmologists as shown in subsection 5.2. This suggests that the information provided by the model can potenti llbaye useful to deal with the lack of interpretability in deep learning models applied to medical images.

Despite the very good results for the SD-OCT volume AMD classification as reported in subsection 5.1, the evaluation of the qualitative analysis of CAM and the individual B-scan classification provided useful feedback about the medical findings in scans classified as false positives and false negatives as presented in subsection 5.2 and 5.3 respectively. The false positives were misclassified mainly in non-centered scans or poor resolution among the scans,

which means that the layers are not defined in some scans inside the volume as shown in the scan [B] of Figure 5. On the other hand, false negatives could be due to the presence of subtle findings in some images. We hypothesized that the tiny drusen may be misleading the proposed method to classify these images as healthy as presented in scan [C] of Figure 5.

The speckle noise in images from medical devices was different among the SD-OCT volumes. The OCT-NET model was trained with random weights presenting a better model generalization in classification tasks without being affected by the speckle noise. Finally, we want to emphasize that our approach was assessed with SD-OCT volumes acquired from different devices, with a different populations and several retinal diseases. However, these datasets are relatively small and our study lacks an evaluation over larger datasets, this will be part of our future work.

7. Conflict of interest statement

The authors declare no conflicts of interest related to this research work.

8. Acknowledgment

Oscar Perdomo thanks COLCIENCIAS for funding this research with a doctoral grant. We appreciate the efforts devoted by: Prof. Sina Farsiu and Prof. Cynthia A. Toth from Duke University; Carol Cheung and Tien Y Wong from the Chinese University of Hong Kong (CUHK) and Singapore Eye Research Institute (SERI) to collect SD-OCT volumes. This work was partially supported by Nvidia.

References

[1] Undurti N Das. Diabetic macular edema, retinopathy and age-related macular degeneration as inflammatory conditions. *Archives of medical science: AMS*, 12(5):1142, 2016.

[2] Mads Fonager Nørgaard and Jakob Grauslund. Automated screening for diabetic retinopathy–a systematic review. *Ophthalmic research*, 2018.

[3] Ryan Lee, Tien Y Wong, and Charumathi Sabanayagam. Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *Eye and vision*, 2(1):17, 2015.

[4] Yali Jia, Steven T Bailey, David J Wilson, Ou Tan, Michael L Klein, Christina J Flaxel, Benjamin Potsaid, Jonathan J Liu, Chen D Lu, Martin F Kraus, et al. Quantitative optical coherence tomography angiography of choroidal neovascularization in age-related macular degeneration. *Ophthalmology*, 121(7):1435–1444, 2014.

[5] Ryan L Shelton, Woonggyu Jung, Samir I Sayegh, Daniel T McCormick, Jeehyun Kim, and Stephen A Boppart. Optical coherence tomography for advanced screening in the primary care office. *Journal of biophotonics*, 7(7):525–533, 2014.

[6] Zhuo Zhang, Ruchir Srivastava, Huiying Liu, Xiangyu Chen, Lixin Duan, Damon Wing Kee Wong, Chee Keong Kwoh, Tien Yin Wong, and Jiang Liu. A survey on computer aided diagnosis for ocular diseases. *BMC medical informatics and decision making*, 14(1):80, 2014.

[7] Martin M Nentwich and Michael W Ulbig. Diabetic retinopathy-ocular complications of diabetes mellitus. *World journal of diabetes*, 6(3): 489, 2015.

[8] Desire Sidibe, Shrinivasan Sankar, Guillaume Lemaitre, Mojdeh Rastgoo, Joan Massich, Carol Y Cheung, Gavin SW Tan, Dan Milea, Ecosse Lamoureux, Tien Y Wong, et al. An anomaly detection approach for the identification of dme patients using spectral domain optical coherence tomography images. *Computer methods and programs in biomedicine*, 139:109–117, 2017.

[9] Gabriela Samagaio, AiAda Estévez, Joaquim de Moura, Jorge Novo, Maria Isabel Fernandez, and Marcos Ortega. Automatic macular edema identification and characterization using oct images. *Computer Methods and Programs in Biomedicine*, 2018.

[10] E Talisa, Marco A Bonini Filho, Adam T Chin, Mehreen Adhi, Daniela Ferrara, Caroline R Baumal, Andre J Witkin, Elias Reichel, Jay S Duker, and Nadia K Waheed. Spectral-domain optical coherence tomography angiography of choroidal neovascularization. *Ophthalmology*, 122(6):1228–1238, 2015.

[11] Adeel M Syed, Taimur Hassan, M Usman Akram, Samra Naz, and Shehzad Khalid. Automated diagnosis of macular edema and central serous retinopathy through robust reconstruction of 3d retinal surfaces. *Computer methods and programs in biomedicine*, 137:1–10, 2016.

[12] Oliver Faust, Yuki Hagiwara, Tan Jen Hong, Oh Shu Lih, and U Rajendra Acharya. Deep learning for healthcare applications based on physiological signals: a review. *Computer methods and programs in biomedicine*, 2018.

[13] Eli Gibson, Wenqi Li, Carole Sudre, Lucas Fidon, Dzhoshkun I Shakir, Guotai Wang, Zach Eaton-Rosen, Robert Gray, Tom Doel, Yipeng Hu, et al. Niftynet: a deep-learning platform for medical imaging. *Computer methods and programs in biomedicine*, 158:113–122, 2018.

[14] Jose Ignacio Orlando, Elena Prokofyeva, Mariana del Fresno, and Matthew B Blaschko. An ensemble deep learning based approach for red lesion detection in fundus images. *Computer methods and programs in biomedicine*, 153:115–127, 2018.

[15] Yawen Xiao, Jun Wu, Zongli Lin, and Xiaodong Zhao. A deep learning-based multi-model ensemble method for cancer prediction. *Computer methods and programs in biomedicine*, 153:1–9, 2018.

[16] Xiaohong W Gao, Rui Hui, and Zengmin Tian. Classification of ct brain images based on deep learning networks. *Computer methods and programs in biomedicine*, 138:49–56, 2017.

[17] Zhong Yin, Mengyuan Zhao, Yongxiong Wang, Jingdong Yang, and Jianhua Zhang. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Computer methods and programs in biomedicine*, 140:93–110, 2017.

[18] Cecilia S Lee, Doug M Baughman, and Aaron Y Lee. Deep learning is effective for classifying normal versus age-related macular degeneration oct images. *Ophthalmology Retina*, 1(4):322–327, 2017.

[19] Hyungwoo Lee, Kyung Eun Kang, Hyewon Chung, and Hyung Chan Kim. Automated segmentation of lesions including subretinal hyperreflective material in neovascular age-related macular degeneration. *American journal of ophthalmology*, 191:64–75, 2018.

[20] Sripad Krishna Devalla, Prajwal K Renukanand, Bharathwaj K Sreedhar, Giridhar Subramanian, Liang Zhang, Shamira Perera, Jean-Martial Mari, Khai Sing Chin, Tin A Tun, Nicholas G Strouthidis, et al. Drunet: a dilated-residual u-net deep learning network to segment optic nerve head tissues in optical coherence tomography images. *Biomedical optics express*, 9(7):3244–3265, 2018.

[21] Freerk G Venhuizen, Bram van Ginneken, Bart Liefers, Mark JJP van Grinsven, Sascha Fauser, Carel Hoyng, Thomas Theelen, and Clara I Sánchez. Robust total retina thickness segmentation in optical coherence tomography images using convolutional neural networks. *Biomedical optics express*, 8(7):3292–3316, 2017.

[22] Avi Ben-Cohen, Dean Mark, Ilya Kovler, Dinah Zur, Adiel Barak, Matias Iglicki, and Ron Soferman. Retinal layers segmentation using fully convolutional network in oct images. *RSIP Vision*, 2017.

[23] Mike Pekala, Neil Joshi, David E Freund, Neil M Bressler, Delia Cabrera DeBuc, and Philippe M Burlina. Deep learning based retinal oct segmentation. *arXiv preprint arXiv:1801.09749*, 2018.

[24] Yufan He, Aaron Carass, Bruno M Jedynak, Sharon D Solomon, Shiv Saidha, Peter A Calabresi, and Jerry L Prince. Topology guaranteed segmentation of the human retina from oct using convolutional neural networks. *arXiv preprint arXiv:1803.05120*, 2018.

[25] Oscar Perdomo, Sebastian Otálora, Fabio A González, Fabrice Meriaudeau, and Henning Müller. Oct-net: A convolutional network for automatic classification of normal and diabetic macular edema using sd-oct volumes. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, pages 1423–1426. IEEE, 2018.

[26] Muhammad Awais, Henning Müller, Tong B Tang, and Fabrice Meriaudeau. Classification of sd-oct images using a deep learning approach. In *Signal and Image Processing Applications (ICSIPA), 2017 IEEE International Conference on*, pages 489–492. IEEE, 2017.

[27] Genevieve CY Chan, Ravi Kamble, Henning Müller, Syed AA Shah, TB Tang, and Fabrice Mériaudeau. Fusing results of several deep learning architectures for automatic classification of normal and diabetic macular edema in optical coherence tomography. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 670–673. IEEE, 2018.

[28] Ravi M Kamble, Genevieve CY Chan, Oscar Perdomo, Fabio A González, Manesh Kokare, Henning Müller, and Fabrice Mériaudeau. Automated diabetic macular edema (dme) analysis using fine tuning with inception-resnet-v2 on oct images. In *Conference proceedings:... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, volume 2018, pages 2715–2718. IEEE, 2018.

[29] Weiwei Sun, Xiaoming Liu, and Zhou Yang. Automated detection of age-related macular degeneration in oct images using multiple instance learning. In *Ninth International Conference on Digital Image Processing (ICDIP 2017)*, volume 10420, page 104203V. International Society for Optics and Photonics, 2017.

[30] Freerk G Venhuizen, Bram van Ginneken, Bart Bloemen, Mark JJP van Grinsven, Rick Philipsen, Carel Hoyng, Thomas Theelen, and Clara I Sánchez. Automated age-related macular degeneration classification in oct using unsupervised feature learning. In *Medical Imaging 2015: Computer-Aided Diagnosis*, volume 9414, page 94141I. International Society for Optics and Photonics, 2015.

[31] Arunava Chakravarty, Divya Jyothi Gaddipati, and Jayanthi Sivaswamy. Construction of a retinal atlas for macular oct volumes. In *International Conference Image Analysis and Recognition*, pages 650–658. Springer, 2018.

[32] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342, 2018.

[33] Cecilia S Lee, Doug M Baughman, and Aaron Y Lee. Deep learning is effective for classifying normal versus age-related macular degeneration oct images. *Ophthalmology Retina*, 1(4):322–327, 2017.

[34] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5): 1122–1131, 2018.

[35] R Natarajan et al. Comparative analysis of optical coherence tomography retinal images using multidimensional and cluster methods. *Biomedical Research*, 26(2), 2015.

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[37] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[38] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[39] Sina Farsiu, Stephanie J Chiu, Rachelle V O'Connell, Francisco A Folgar, Eric Yuan, Joseph A Izatt, Cynthia A Toth, Age-Related Eye Disease Study 2 Ancillary Spectral Domain Optical Coherence Tomography Study Group, et al. Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. *Ophthalmology*, 121(1):162–172, 2014.