DEEP-LEARNING-AIDED DIABETIC RETINOPATHY DIAGNOSIS BASED ON STRUCTURAL AND ANGIOGRAPHIC OPTICAL COHERENCE TOMOGRAPHY

by

Pengxiao Zang

A DISSERTATION

Presented to the

Department of Biomedical Engineering

within the Oregon Health & Science University

School of Medicine

in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

December 2023

Doctor of Philosophy
School of Medicine
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the PhD dissertation of
Pengxiao Zang
has been approved.

Yali Jia, PhD
Advisor

Young Hwan Chang, PhD
Chair

Michelle R. Hribar, PhD
Committee Member

Thomas S. Hwang, MD
Committee Member

David Huang, MD, PhD
Committee Member

TABLE OF CONTENTS

List of Figures

List of Tables

List of Abbreviations

Symbols
2D two-dimensional.
3D three-dimensional.

A
AMD age-related macular degeneration.
AREDS Age-Related Eye Disease Study
AUC area under receiver operating curve.

B
BAM biomarker activation map.
BM Bruch's membrane.

C
CAD computer-assisted diagnostic.
CAM class activation map
CFP color fundus photography.
CNN convolutional neural network.

D
DcardNet densely and continuously connected neural network with adaptive rate dropout.
DCP deep capillary plexus.
DME diabetic macular edema.
dpr dropout rate.
DR diabetic retinopathy.
DSC dice similarity coefficient.
DTD deep Taylor decomposition.

E
ETDRS Early Treatment of Diabetic Retinopathy Study.
EZ ellipsoid zone

F
FA fluorescein angiograph.
FAZ fovea avascular zone.

G
GAN generative adversarial learning.
GCC ganglion cell complex.
GCL ganglion cell layer.

H
HOGs histogram of gradients.

I
ICP intermediate capillary plexus.
ILM inner limiting membrane.
INL inner nuclear layer.
IOU intersection over union.
IPL inner plexiform layer.

L
LRP layer-wise relevance propagation.

N
NFL nerve fiber layer.
NPA non-perfusion area.
NPDR non-proliferative diabetic retinopathy.
nrDR non-referable diabetic retinopathy.
nvtDR non-vision-threatening diabetic retinopathy.
nDME none diabetic macular edema.

O
OCT optical coherence tomography.
OCTA optical coherence tomographic angiography.
ONH optic nerve head.
ONL outer nuclear layer.
OPL outer plexiform layer.

P
PDR proliferative diabetic retinopathy.

R
rDR referable diabetic retinopathy.
ReLU rectified linear unit.
ROC receiver operating characteristic.
RPCP radial peripapillary capillary plexus.
RPE retinal pigment epithelium layer.

S
SSD sum of squared differences.
SSI signal strength index.
SVC superficial vascular complex.
SSIM structural similarity index.

V
vtDR vision threatening diabetic retinopathy.

ABSTRACT

Diabetic retinopathy (DR) is a leading cause of preventable blindness globally. Regular screening for DR is critical for timely management and treatment since fewer than half of DR patients are aware of their condition even when it has advanced to a referable stage. However, the current DR screening process, which is based on manual diagnosis using two different imaging modalities, is expensive and time-consuming for regular DR screening. Automated DR diagnosis is urgently needed. A clinically applicable DR diagnosis should incorporate several design features. Firstly, it should rely on a single imaging modality that can provide sufficient information for DR diagnosis. Secondly, it should be able to classify each input into multiple levels of DR severities. Thirdly, it should possess the capability to distinguish DR from both healthy control cases and other eye diseases. Finally, it should provide clinically meaningful interpretability for each diagnostic result. Therefore, five deep learning systems were developed for automated and interpretable DR diagnosis based on optical coherence tomography (OCT) and its angiography (OCTA). (1) A retinal layer segmentation system which operates volumetrically and uses a customized U-shaped convolutional neural network (CNN) for gross segmentation and a multi-weight graph search algorithm to refine the network output. (2) A two-dimensional multi-level DR classification system that only needs *en face* projections as inputs. A densely and continuously CNN with adaptive rate dropout was developed to classify each input into three clinically relevant severity levels of DR. (3) A three-dimensional (3D) multi-level DR classification system that uses the data volumes as inputs. A customized 3D CNN was developed to detect both referable and vision-threatening DR. (4) A multi-eye-disease detection system that can diagnose DR in a clinical context, i.e., in a dataset that includes healthy eyes and eyes with DR and other diseases. Here, a semi-sequence classifier was used to detect healthy controls, DR, age-related macular degeneration, and glaucoma. (5) An interpretability system that can highlight the biomarkers utilized in the second, third, and fourth modules in a biomarker activation map (BAM). The BAM was generated based on the generative adversarial learning technique. The combination of these systems achieved specialist-level performance in multi-level DR diagnosis and could provide clinically meaningful interpretability for each diagnostic result. In real-world practice, these deep learning systems could reduce vision loss and lower clinical burden by providing time-effective, cost-efficient, and interpretable DR diagnosis.

Acknowledgments

I would like to firstly extend my heartfelt gratitude to my mentor, Dr. Yali Jia, whose unwavering guidance and support have been instrumental in my journey to completing this dissertation. Dr. Jia's diligence and commitment as a mentor have been truly invaluable. Furthermore, Dr. Jia not only nurtured my technical skills but also cared deeply about my career development, generously sharing her experiences and fostering collaborative opportunities that enriched my knowledge and expertise.

I am equally thankful to my esteemed dissertation advisory committee members, Dr. Young Hwan Chang, Dr. Michelle R. Hribar, Dr. Thomas S. Hwang, and Dr. David Huang. Their dedicated involvement and sage advice have profoundly influenced my studies and shaped my path towards a promising career. Their unwavering support and encouragement have been indispensable, and this dissertation would not have been possible without their expert guidance.

I extend my appreciation to Dr. Tristan T. Hormel, whose invaluable guidance on manuscript writing has been instrumental. Dr. Hormel's meticulous feedback and constructive suggestions have consistently improved my writing and presentation skills. I also acknowledge Dr. Thomas S. Hwang and Dr. Steven Bailey for their generous sharing of clinical knowledge, as well as the dedicated research coordinator team, including Denzil Romfh, Humberto Martinez, Chinmay Deshpande, George Pacheco, and Kevin Lathrop, for their commendable efforts in data collection.

A heartfelt thank you goes out to the members of the Center for Ophthalmic Optics & Lasers at Casey Eye Institute. Their camaraderie and assistance have been a constant source of support and friendship throughout my PhD journey.

In closing, I reserve my deepest gratitude for my unwavering pillars of strength and encouragement—my parents and my beloved wife. Their boundless love, unwavering belief in my potential, and unflagging support have been the driving forces behind my academic journey. Throughout the countless late nights, challenging moments, and rigorous studies, their unwavering faith in me has been my constant motivation. My parents' sacrifices and endless encouragement have been instrumental in my pursuit of knowledge, and my wife's unwavering support and understanding have provided me with the peace and balance I needed to navigate the complexities of doctoral research. I am profoundly fortunate to have them by my side, and their love and support have been the true foundations upon which this academic achievement has been built.

1 Introduction

Diabetic retinopathy (DR) is a major microvascular complication of both type 1 and type 2 diabetes, which affects the fine vessels in the eye and causes vision loss [1]. Currently, DR has become a leading cause of preventable blindness globally [1]. About 40% to 45% of diabetic patients are likely to have DR at some point in their life [2]. Treatment of the DR at an early stage can result in the best prognosis, and vision loss can be delayed or deferred [3-5]. However, fewer than half of DR patients are aware of their condition because it may be asymptomatic even in the referable stages [2]. In addition, the prevalence of DR is also growing [6]. Therefore, regular screening for DR is critical for timely management and treatment.

In current DR screening, trained specialists need to manually diagnose each patient based on two imaging modalities [3]. The first is fundus photography, which yields a two-dimensional (2D) retina image based on a fundus camera [7-10]. Fundus photography is used for DR staging based on the Early Treatment of Diabetic Retinopathy Study (ETDRS) scale [11, 12]. The second imaging modality is optical coherence tomography (OCT), which yields a three-dimensional (3D) micrometer-scale-resolution image of ocular fundus tissue based on reflectance signals obtained using interferometric analysis of low-coherence light [13]. OCT is used for the detection of diabetic macular edema (DME), which is a vision-threatening manifestation of DR and can be found at any DR stage [14, 15]. However, manual diagnosis by specialists is time-consuming due to a large number of diabetic patients globally [1]. In addition, current screening is also expensive since two diagnostic procedures are needed. Furthermore, regular screening of patients in underdeveloped areas is impossible due to a lack of specialists and infrastructure [16]. Therefore, an automated DR diagnostic system based on only one imaging modality is critical to prevent the blindness of millions of diabetic patients worldwide [1].

According to recent studies, compared to fundus photography, OCT has shown several advantages for DR diagnosis in the absence of allied imaging modalities [17-20]. Firstly, an imaging modality called OCT angiography (OCTA) can be acquired simultaneously with OCT by measuring the decorrelation values to differentiate vasculature from static tissues. OCTA can provide 3D high-resolution images of the microvasculature of the retina [21-23]. By combining OCT and OCTA together, images acquired from just a single procedure can show both static tissue and vascular structure. Recently, numerous investigators explored OCTA in the staging of DR and demonstrated competitive performances when compared to fundus photography [17-20]. Secondly, fundus photography has a low sensitivity (60-73%) and specificity (67-79%) for detecting DME, which accounts for the majority of vision loss in DR [14, 15]. This means that even when an automated DR diagnostic system performs very well against a ground truth generated from fundus photographs, patients with DME may still frequently be misdiagnosed. Therefore, combined OCT and OCTA could instead be used as a single imaging modality for automated DR diagnostic systems since they can provide sufficient information for both DME detection and DR staging.

In recent years, deep learning classifiers have achieved state-of-the-art performances in several disease diagnostic tasks [24-28]. The nonlinearity and complexity based on the hidden layers and activation functions allow deep learning classifiers the ability to fit much more complex models compared to conventional algorithms [24]. Therefore, deep learning classifier should be used as a core technique for automated DR diagnosis based on OCT and OCTA. However, the high performance of the current deep learning classifier often comes at the cost of inscrutable outputs [29-31]. The presence of hidden layers in classifier architectures renders a straightforward account of the classifier's action on inputs inaccessible and makes deep learning classifier outputs difficult to verify. In the absence of heuristic devices, deep learning classifiers cannot be confirmed outside of manual grading, which largely defeats the purpose of automation. The poor interpretability can also obfuscate potential bias that could negatively affect performance in under-represented ethnic groups: a classifier trained only based on one ethnic group may be biased when evaluated on data from others. These issues present a major hurdle for translating deep learning classifiers into the clinic [29-31].

By considering the current circumstances above, we developed 5 deep-learning-aided systems based on OCT and OCTA to achieve automated and clinically interpretable DR diagnosis. In the beginning, the *en face* OCT and OCTA scanned on the macular region was selected as the input for deep-learning-aided DR classifier based on current clinical studies [17-20]. Retinal layer segmentation was then needed for the generation of *en face* OCT and OCTA. However, the first system was mainly focused on the retinal layers segmentation for OCT volume scanned on the optic nerve head (ONH) region [32]. The reason we did not develop this system for the OCT volume scanned on the macula region was that such system has been previously developed by our groups. In addition, the retinal layer segmentation for OCT scanned on ONH region was much more challenging. The successful development of our first system could enable the generation of wide field *en face* OCT and OCTA, which included both macular and ONH regions, to improve the diagnosis performance in the future. The second system was developed as a multi-level DR classifier based on deep learning, which used the *en face* OCT and OCTA as inputs [33]. By combining the first two systems, the automated DR diagnosis has been initially achieved. However, one major concern of the DR diagnosis based on *en face* OCT and OCTA was that the diagnostic performance was influenced by the accuracy of retinal layer segmentation which could be unreliable in severe cases. Therefore, we developed the third system as a deep learning DR classifier based on 3D OCT and OCTA [34]. After we developed automated DR classifiers based on both 2D and 3D inputs, we decided to develop the fourth system as a multi-eye-disease classifier which focused on the identification of DR from healthy eyes and other eye diseases [35]. The successful development of this system enables our DR diagnosis to be more adaptable to real-world clinical settings. To achieve clinically interpretable DR diagnosis, the last system was developed for interpretability, which can highlight the biomarkers utilized in three classification systems on an attention map for each input [36]. In real-world practice, the 5 deep learning systems could reduce vision loss and lower clinical burden by providing time-efficient, cost-efficient, and clinically explainable DR diagnosis. The following sections of this dissertation are organized as follows: the 5 systems are presented in sections 2 to 6, respectively. The discussion and conclusion of these systems are presented in section 7 and 8, respectively.

2 Automated Segmentation of Peripapillary Retinal Boundaries in Optical Coherence Tomography Combining Convolutional Neural Network and Multi-weights Graph Search

2.1 Abstract

Quantitative analysis of the peripapillary retinal layers and capillary plexuses from optical coherence tomography (OCT) and OCT angiography images depend on two segmentation tasks – delineating the boundary of the optic disc and delineating the boundaries between retinal layers. Here, we present a method combining a neural network and graph search to perform these two tasks. A comparison of this novel method's segmentation of the disc boundary showed good agreement with the ground truth, achieving an overall Dice similarity coefficient of 0.91 ± 0.04 in healthy and glaucomatous eyes. The absolute error of retinal layer boundaries segmentation in the same cases was 4.10 ± 1.25 μm.

2.2 Introduction

Optical coherence tomography (OCT) provides noninvasive, structural images of eye fundus tissue based on interferometric analysis of low-coherence light [13]. By considering blood flow induced temporal variation in the signal garnered from OCT, vasculature can be distinguished from static tissue. There are many versions of this technique; collectively they are termed OCT angiography (OCTA) [21-23]. Measurement of retinal layer thickness from structural OCT and analysis of capillary plexuses from OCTA can both help clinical diagnosis and early detection of glaucoma, which is the leading cause of irreversible blindness globally [37-41]. But the clinical utility of such measurements requires accuracy and precision, both of which depend critically on the segmentation of both the optic disc boundary and peripapillary retinal boundaries. Segmentation of these anatomical regions is, then, a critically important task.

Since manual segmentation is time-consuming, several methods to segment the optic disc and peripapillary retinal boundaries have been proposed [42-50]. For peripapillary retinal boundaries segmentation graph search algorithms based on intensity differences between anatomical slabs from structural OCT have been used frequently and show good results. Antony et al. proposed a 3D graph search method for the segmentation of both the optic disc boundary and the peripapillary retinal boundaries [44]. Zang et al. proposed a method which detected the optic disc boundary and segmented peripapillary retinal boundaries separately using a dynamic-programming based graph search algorithm [48]. Gao et al. proposed a method which combined the active appearance model and graph search to segment the peripapillary retinal boundaries [49]. Yu et al. proposed a shared-hole graph search method which first segments the optic disc boundary and then segments the peripapillary retinal boundaries [50]. However, speckle noise and vessel shadows both seriously detrimentally impact segmentation accuracy based just on graph search.

Nowadays, deep learning plays an important role in medical image processing and several learning-based methods exist for segmentation of OCT data [51-56]. Devalla et al. proposed a dilated-residual U-Net to segment optic nerve head tissues such as the lamina cribrosa, choroid, sclera and so on [53]. But the peripapillary retinal boundaries were not segmented in this study. Kugelman proposed a retinal boundary segmentation method for macular OCT based on a combination of recurrent neural networks and graph search [54]. However, the anatomical disruption caused by the optic disc makes peripapillary retinal boundaries segmentation much more challenging than the macular region. Networks trained on macular OCT scans therefore may not generalize well to the peripapillary region.

In this study, we propose an automated segmentation method for optic disc boundary detection and peripapillary retina layer segmentation. We designed two separate neural networks and trained one each to segment the optic disc boundary and peripapillary retinal layers. The final peripapillary retinal boundaries were calculated based on the prediction and gradient maps using a multi-weights graph search algorithm.

2.3 Methods

2.3.1 Patient recruitment and data acquisition

In this study, 46 healthy and 63 participants with glaucoma were recruited and tested at the Casey Eye Institute, Oregon Health & Science University. The diagnoses of all the participants were made by an expert clinical examination. The participants were enrolled after informed consent in accordance with an Institutional Review Board approved protocol. The study was conducted in compliance with the Declaration of Helsinki.

The peripapillary retinal area was scanned using a commercial 70-kHz spectral-domain OCT system (Avanti RTVue-XR, Optovue Inc) with 840-nm central wavelength. The scan regions were 4.5 × 4.5 mm and 1.6 mm in depth (304 × 304 × 640 pixels) centered on the optic disc. Two repeated B-frames were captured at each line-scan location. The blood flow of each line-scan location was detected using the split-spectrum amplitude-decorrelation angiography algorithm based on the speckle variation between two repeated B-frames [23, 57]. The OCT structural images were obtained by averaging two repeated B-frames. For each data set, two volumetric raster scans (one x-fast scan and one y-fast scan) were registered and merged through an orthogonal registration algorithm to reduce motion artifacts [58].

In each OCT data set, the following layers or boundaries are anatomically important: inner limiting membrane (ILM), nerve fiber layer (NFL), ganglion cell layer (GCL), inner plexiform layer (IPL), inner nuclear layer (INL), outer plexiform layer (OPL), outer nuclear layer (ONL), ellipsoid zone (EZ), retinal pigment epithelium (RPE), and Bruch's membrane (BM). In this study, seven boundaries (Vitreous/ILM, NFL/GCL, IPL/INL, INL/OPL, OPL/ONL, ONL/EZ, and RPE/BM) were manually segmented by a human grader.

2.3.2 Neural network designing

The neural network used in this study was designed based on the architecture of the classic U-Net [59, 60] (Fig. 2.1). Three max-pooling and (de)convolution layers were separately used in the down-sampling and up-sampling towers. Because each peripapillary retinal layer can not be identified based just on the upper and lower boundaries, the global position in the whole retina is also an important feature. In order to capture both the relative and absolute location of each peripapillary retinal layer, a 3×3 normal and atrous-convolution layer [61, 62] were cascaded together in each layer of the down-sampling and up-sampling towers. In addition, a global block was also designed to capture the local and global information before the final classification layer. The batch normalization [63] and exponential linear unit (ELU) function [64] were used after each convolution layer (except the output layer) to improve the stability of the final classification.

The Dice similarity coefficient (DSC) for each channel of the output map was used in the loss function:

$$\mathrm{Loss} = 1 - \frac{1}{N_{\mathrm{c}}} \sum_{\mathrm{n}=1}^{N_{\mathrm{c}}} \frac{Out_{\mathrm{n}} \cap Lab_{\mathrm{n}} + eps}{Out_{\mathrm{n}} \cup Lab_{\mathrm{n}} + eps} \tag{1}$$

where N_c is the number of final classes, *eps* is set to 1×10^{-5} to keep the division workable, and Out_{n} and Lab_{n} are the n$^{\mathrm{th}}$ channels of the output map and corresponding label manually segmented by a certified grader. Stochastic gradient descent with Nesterov momentum (momentum = 0.9) was used to optimize the variables in the neural network to find the minimum value of the loss function [65]. The learning rate was halved if the value of the loss function kept increasing over three consecutive training steps (starting from an initial learning rate of 0.1). The same network architecture was used for the segmentation of both the optic disc and peripapillary retinal boundaries.

The designed neural network was trained and tested in Python 3.6, and other image processing was performed in MATLAB 2018b. The workstation used in this study has an Intel (R) Core (TM) i7-8700K CPU @ 3.70GHz, 64.0 GB RAM and NVIDIA RTX 2080 GPU.

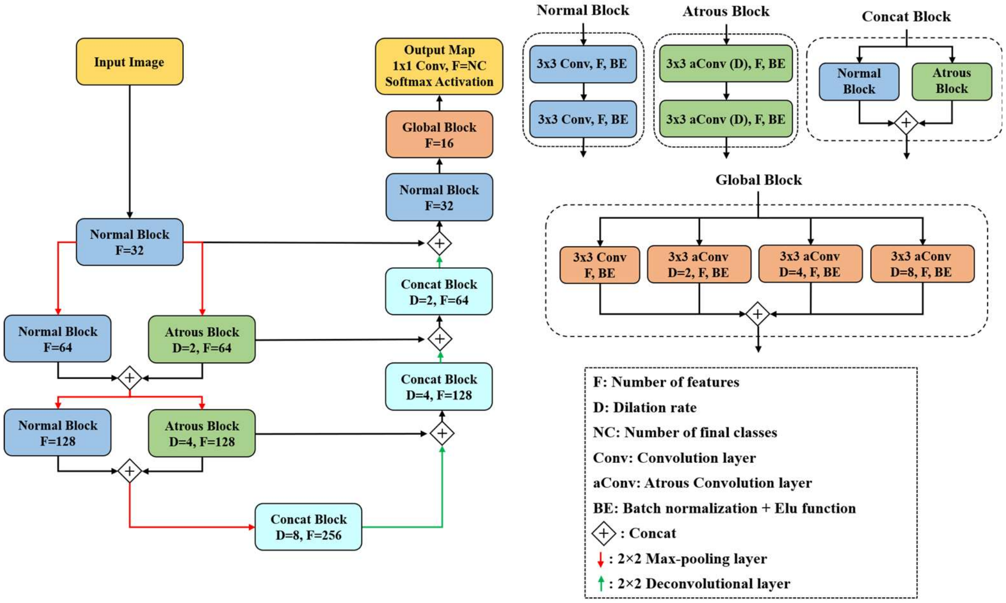


Figure 2.1: The architecture of the designed neural network.

2.3.3 Optic disc boundary segmentation

The major challenge of the peripapillary retinal boundaries segmentation is the special structure of the optic disc, which is totally different from the surrounding retina and varies significantly between eyes. Because the *en face* shape of optic disc is usually approximately circular 180 diametral B-frames were generated based on the detected disc center, thereby ensuring the images used to train the network in optic disc segmentation algorithm were similar.

2.3.3.1 Optic disc center detection

The optic disc center is needed for sampling of the 180 diametral B-frames. However, the optic disc is not always aligned at the exact center of the OCT data volume, and can be far away from the image center (Fig. 2.2(A)). Therefore, we designed an automated localization algorithm for the optic disc that leverages the lack of anatomical layers found in the disk region in order to determine its center. The internal, hierarchical structure of anatomic layers can manifest clearly in OCT images after proper image manipulation. To elucidate these features within our data volumes we designed a convolution kernel k_{hie} to generate a gradient map G_{hie} which demarcates the three strongest retinal layer gradients (Fig. 2.2(C) and 2.2(F)):

$$G_{\text{hie}} = \text{Conv}(B_{\text{normal}}, k_{\text{hie}}) \tag{2}$$

where Conv(•) is the convolution, B_{normal} is each normal B-frame and k_{hie} is a 5×5 kernel with $-\frac{1}{10}$ in the first two rows and $\frac{1}{15}$ in the last three rows. The binary image of each gradient map was then generated by extracting the layers with intensity above an empirically determined threshold (Fig. 2.2(D) and 2.2(G)). Because it lacks internal hierarchical structure, only one layer was detected inside the optic disc (Fig. 2.2(G)). After all the volumetric binary images were generated, we construct an *en face* accumulation image by summing the separate binary images (Fig. 2.2(H)). This leaves the region of the optic disc darker since it retains only one layer after binarization (instead of three), and so obtains lower values in the accumulation image. A binary *en face* image I_b was then generated based on the center region I_c (red box in Fig. 2.2(H)) of the accumulation image to improve the detection stability. This binarization process was defined as:

$$I_b(x,y) = \begin{cases} 1 & 1 - I_c(x,y) > 1.3 \times mean(1 - I_c) \\ 0 & otherwise \end{cases} \tag{3}$$

The optic disc center was then calculated as the geometric center of the binary image just obtained. Though some large vessels might still be visible in the binary image due to the vessel shadows, the calculation of the optic disc center is unaffected due to the approximate rotational symmetry.

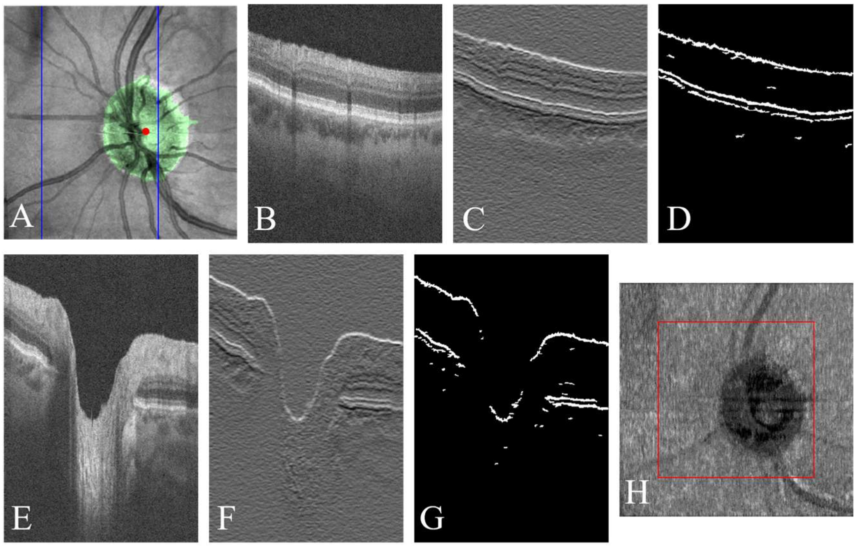


Figure 2.2: Diagram of the optic disc center detection. (A) *En face* average projection of the volumetric OCT. The detected optic disc region is covered by green. The red point is the detected disc center. (B) The normal B-frame corresponding to the position of the left blue line, which is outside of the disc. (C) The gradient map of the B-frame

in (B). (D) The binary image of the layers with highest gradient intensity in (C). (E) The normal B-frame corresponding to the position of the right blue line, which is inside the disc. (F) The gradient map of the B-frame in (E). (G) The binary image of the layers with highest gradient intensity in (F). Note the single band of pixels in the disc region. (H) *En face* accumulation projection based on the volumetric gradient map. The center region with two thirds of the image length is indicated by the red box.

2.3.3.2 Diametral B-frames generation and disc boundary segmentation

The 180 diametral B-frames and corresponding labels were then generated from 1° to 180° based on the detected optic disc center and resized to 416×416 (416 being the pixel length of the image diagonal). After this we cropped the images for network training (Fig. 2.3). Because the optic disc boundary is defined as the Bruch's membrane opening [66], the area of EZ + RPE (cyan region in Fig. 2.3(C)) and the remaining B-frame area constituted the input labels for the disc boundary neural network. The initial *en face* optic disc binary image was then obtained based on the 180 prediction maps from the trained network through a coordinate transformation. The output region so obtained is rough so we performed a multi-angle edge smoothing process on the initial boundary consisting of two steps. First, the bumpy artifacts were removed through a morphological opening process. After that, the convex hull of the disc region was calculated to make sure the final disc region was convex (Fig. 2.4).

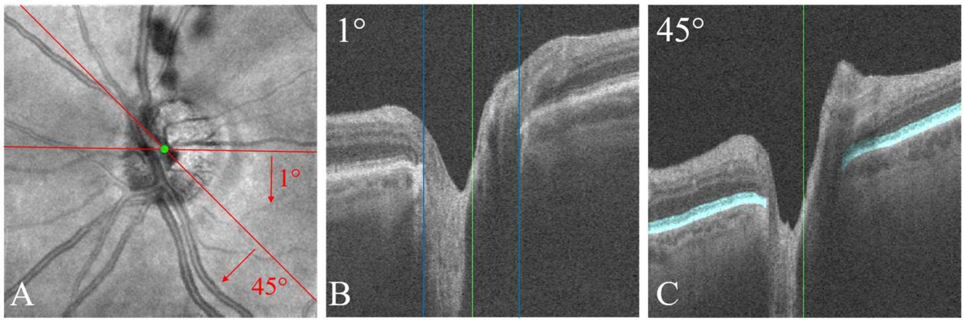


Figure 2.3: Generation of diametral B-frames. (A) *En face* average projection of a volumetric OCT scan from a glaucoma patient. The green point is the automatically detected optic disc center. The two red lines with angle and arrows indicate planes along which the diametral B-frames are generated. (B) The diametral B-frame corresponding to the red line at 1°. The green line corresponds to the optic disc center (green point) in (A). The region between two blue lines is the optic disc. The peripapillary retina is to the left and right of the blue lines. (C) The generated diametral B-frame corresponding to the red line at 45° in (A). The manually segmented EZ + RPE are colored in cyan.

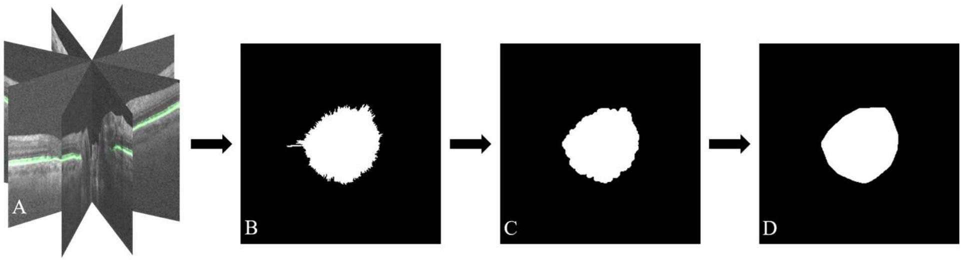


Figure 2.4: Smoothing process of the initial optic disc boundary. (A) Volumetric prediction maps of EZ + RPE. (B) Initial optic disc region based on the *en face* projection of (A). (C) The bump artifacts were removed using morphological opening. (D) The final optic disc region after the convex hull calculation.

2.3.4 Peripapillary retinal layer segmentation

The training data set for the peripapillary retinal boundaries segmentation network was obtained based on the manually delineated internal boundaries between retinal layers. In order to provide extra features for learning and help to mitigate errors due to layer distortion and vessel shadows near the disc we organized the input data as the combination of several adjacent B-frames. Therefore, each input image in the training data set contained channels with size 416×304×5, from a combination of five adjacent B-frames (Fig. 2.5(A)). Each input label in the training data set was calculated based on the manually segmented retinal layers of the middle (i.e., third, marked by red arrow in Fig. 2.5(A)) B-frame of the corresponding image. The size of each input label was 416×304×7, with the first channel corresponding to the area outside the retina. The other channels are the regions of the six main retinal layers (Fig. 2.5(C)).

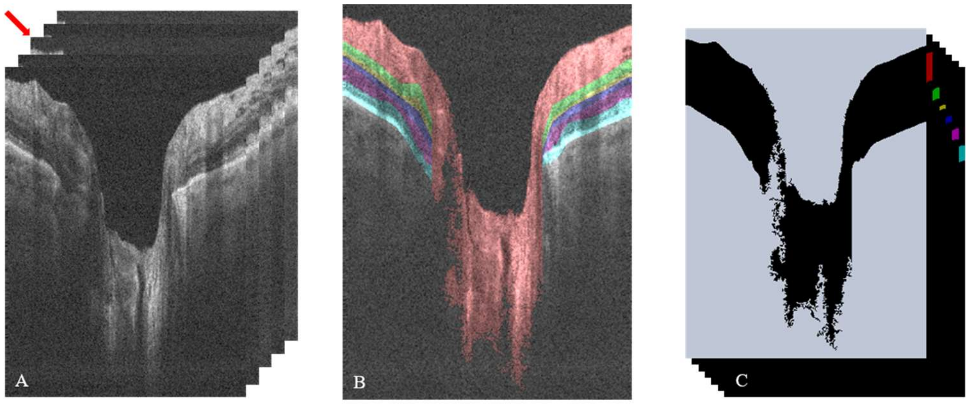


Figure 2.5: The image and corresponding label in the training data set for the designed neural network for peripapillary retinal boundaries segmentation. (A) Image constructed from five adjacent B-frames. (B) Colormap of the peripapillary retinal layers based on the manually delineated boundaries of the B-frame marked by red arrow in (A). Six major layers are shown: NFL (red), IPL (green), INL (yellow), OPL (blue), ONL (purple), and EZ + RPE (cyan). (C) The seven channel labels based on the manual delineation of the third channel of (A).

After the trained neural network obtained initial boundaries B_{initial} based on the prediction maps of each B-frames in the volumetric OCT, the final eight boundaries were obtained by refining the initial boundaries using a multi-weights graph search (Fig. 2.6). (The EZ/RPE boundary was not segmented by the neural network, but we added it at this step in order to obtain a complete segmentation.) To improve the accuracy and stability of this graph search, weights were calculated not just based on the search direction but also based on the vertical distance to initial boundaries. The multi-weights graph search was defined as

$$P(x,z)=\arg\min(P(x-1,z+d(i))+G(x,z)\times(w(i)+\left|z+d(i)-B_{\text{initial}}(x-1)\right|\times 0.1)$$
$$i=[1,\ 2,\ ...,\ n]\quad d=[-3,\ -2,\ -1,\ 0,\ 1,\ 2,\ 3]$$
$$w=[1.4,\ 1.2,\ 1.0,\ 1.0,\ 1.0,\ 1.2,\ 1.4] \tag{4}$$

where $P(x,z)$ is the cost of the shortest path from the first column to the coordinate (x,z) in x^{th} column, $G(x,z)$ is the pixel value in the corresponding gradient map (examples in Fig. 2.2(C) and 2.2(F)), $z + d(i)$ is the row of one of the n neighboring pixels in $(x\text{-}1)^{\text{th}}$ column, and $w(i)$ is the empirically determined weight assigned to each search direction.

Near the optic disc, there is large variation in the vitreous/ILM boundary location. Furthermore, in this region we require the boundaries converge to the Bruch's membrane opening. To achieve these goals, we modified the search weights in this region (between the orange and blue lines in Fig. 2.6(B)) according to Eq. 5 and 6:

For Vitreous/ILM:

$$n=21\quad d=[-10,\ -9,\ ...,\ 0,\ ...,\ 9,\ 10]$$
$$w=[1.8,\ 1.8,\ ...,\ 1.8,\ 1.4,\ 1.2,\ 1.0,\ 1.0,\ 1.0,\ 1.2,\ 1.4,\ 1.8,\ ...,\ 1.8,\ 1.8] \tag{5}$$

For the NFL/GCL, IPL/INL, INL/OPL, OPL/ONL, and ONL/EZ:

$$n=17\quad d=[-8,\ -7,\ ...,\ 0,\ ...,\ 7,\ 8]$$
$$w=[1.4,\ 1.4,\ 1.2,\ 1.2,\ 1.1,\ 1.1,\ 1.0,\ 1.0,\ 1.2,\ 1.4,\ 1.6,\ 1.8,\ 2.0,\ 2.2,\ 2.4,\ 2.6,\ 2.8] \tag{6}$$

The searching order of the eight boundaries was RPE/BM → Vitreous/ILM → NFL/GCL → ONL/EZ → INL/IPL → OPL/ONL → IPL/INL → EZ/RPE, and the search region included the initial estimate plus the six pixels above and below. For a boundary without an initial value, the search area was changed to [B_{pre} - 6, B_{pre} + 6], in which B_{pre} was the just segmented boundary of the last B-frame. In addition, the area of each boundary did not exceed the region based on the associated slab's upper and lower limit. For the region inside the optic disc, just the top and bottom boundaries were segmented based on the binary image of the whole retina. These weights and parameters were empirically chosen just based on the training dataset and will be used in segmentation of future data. After the boundary segmentation, each boundary was smoothed by a mean filter with size 5×5.

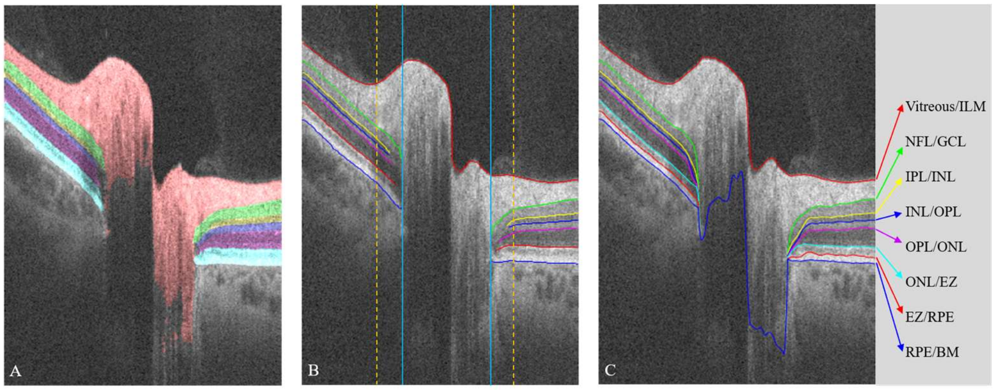


Figure 2.6: The initial boundaries were refined by a multi-weights graph search. (A) The prediction map generated from the trained neural network. (B) The initial boundaries based on the prediction map in (A). The optic disc region, as automatically determined by the algorithm, is indicated by the solid light blue vertical lines. The region between these lines and the orange dotted lines is where refined weights in the graph search are used to ensure convergence to the Bruch's membrane opening. This region covers one quarter of the distance between the edge of the image and the optic disc. (C) The final boundaries after the multi-weights graph search and smoothing.

2.4 Results

In this study, 78 eyes from 46 healthy individuals and 104 eyes from 63 glaucoma patients were scanned. Among the data set, 30 scan volumes each from different healthy participants and glaucoma patients were chosen for the training data set (10800 inputs for optic disc boundary segmentation and 18000 inputs for peripapillary retinal boundaries segmentation). The training batch size was set to 4. Among the 4 inputs, two of them were randomly chosen from the glaucoma training data and another two inputs were randomly chosen from the normal training data. The trained model was obtained after 18000 training steps. The rest of the data set was used to test the performance of this segmentation method. In addition, there was no overlap between the cases used in the training and testing dataset.

2.4.1 Qualitative analysis

In Fig. 2.7, the segmented optic disc is shown in green. The region corresponds to the area expected from visual inspection.

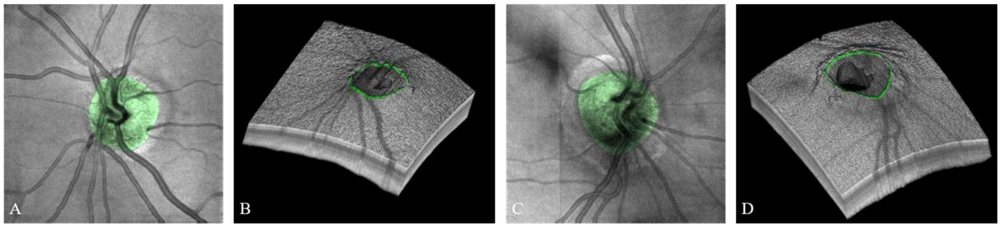


Figure 2.7: The segmentation results of the optic disc boundary. In each part, the optic disc or its boundary is shown in green. (A) The *en face* average projection of the volumetric OCT scanned from a healthy participant. (B) The

bottom-to-top 3D view of the volumetric OCT of (A). (C) The *en face* average projection of the volumetric OCT scanned from a glaucoma patient. (D) The bottom-to-top 3D view of the volumetric OCT of (C).

The segmented boundaries of peripapillary retinal from a healthy participant is shown in Fig. 2.8. In addition, the anatomical structures outside and inside the optic disc are clearly shown in Fig. 2.8(B) and 2.8(C). The superficial vascular complex (SVC), defined as the inner 80% of ganglion cell complex (GCC), includes all structures between the ILM and IPL/INL border [67]. An *en face* SVC angiogram was generated by projecting the maximum decorrelation within the same slab [68-70]. In addition, the segmentation results based on the OCT data scanned from a glaucoma patient are shown in Fig. 2.9. The angiogram of the NFL slab, which is critically important to the detection and diagnosis of glaucoma, was defined as the radial peripapillary capillary plexus (RPCP). Notably, the glauomatous wedge shaped defect can be visualized on both RPCP angiogram (Fig. 2.9(B)) and NFL thickness map (Fig. 2.9(C)) [37-41]. The superotemporal area with capillary loss could also clearly be seen in the RPCP (marked by a green line in Fig. 2.9(B)).

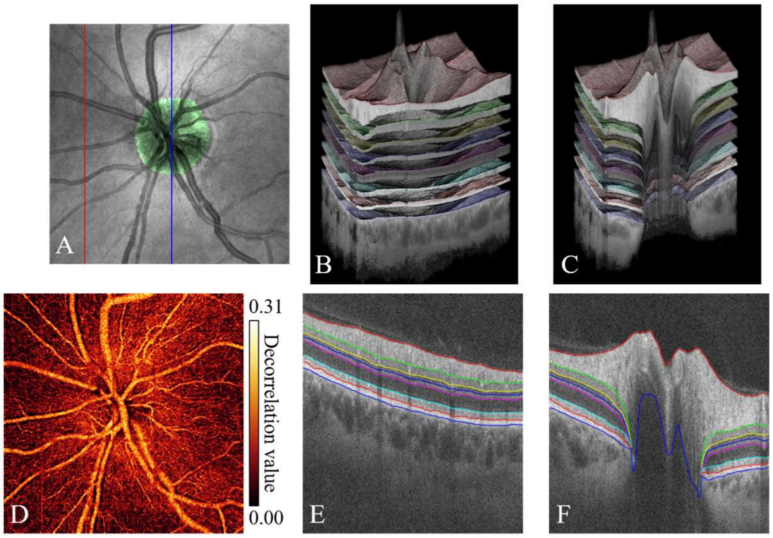


Figure 2.8: Segmentation results of the left eye of a healthy participant. (A) The *en face* average projection, with the segmented optic disc region overlaid in green. (B) The 3D anatomical map of the entire volumetric OCT based on the segmented peripapillary retinal layers. (C) Cutaway from (B) at the blue line location in (A), clearly showing the anatomic structure inside the disc. (D) *En face* SVC angiogram based on the segmented boundaries. (E) B-frame corresponding to the red line in (A) with segmented peripapillary retinal boundaries. (F) Corresponding image for the blue line in (A). The slab boundaries are, from top to bottom, the Vitreous/ILM (red), NFL/GCL (green), IPL/INL (yellow), INL/OPL (blue), OPL/ONL (magenta), ONL/EZ (cyan), EZ/RPE (red) and RPE/BM (blue).

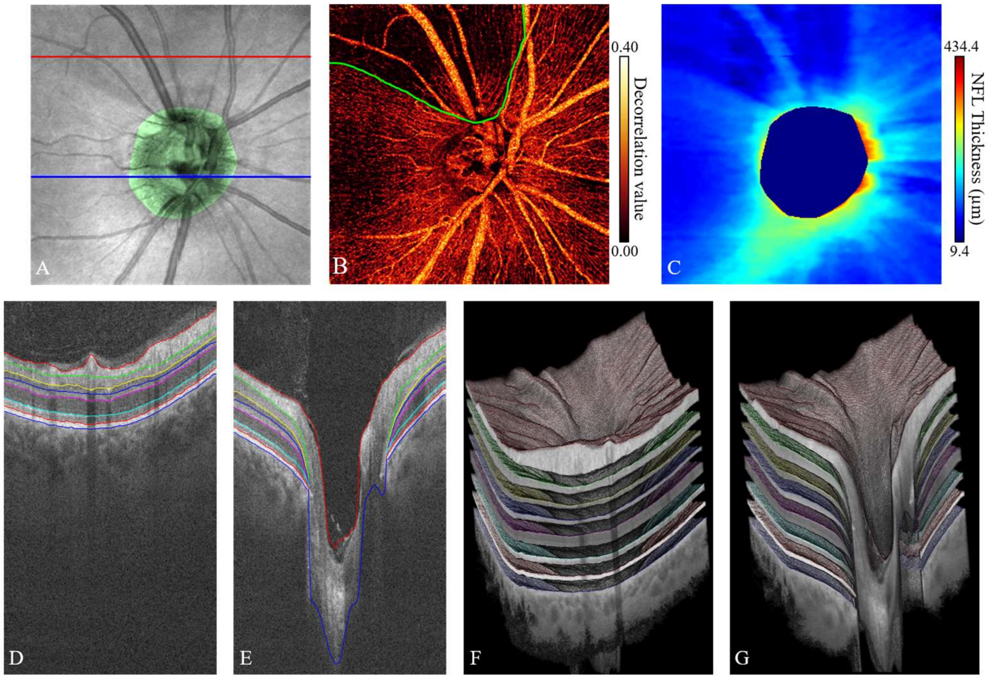


Figure 2.9: Segmentation results for the right eye of a glaucoma patient. (A) *En face* average projection image, with the segmented optic disc region overlaid in green. (B) *En face* RPCP angiogram based on the segmented boundaries. Capillary loss in the superotemporal area is marked with a green line. (C) NFL thickness map based on the segmented peripapillary retinal boundaries. (D) B-frame corresponding to the red line in (A) with segmented peripapillary retinal boundaries. (E) Corresponding image for the blue line in (A). (F) The 3D anatomical map of whole volumetric OCT based on the segmented peripapillary retinal layers. (G) Cutaway from (F) at the blue line location in (A), clearly showing anatomic structure inside the optic disc. The slab boundaries are, from top to bottom, the Vitreous/ILM (red), NFL/GCL (green), IPL/INL (yellow), INL/OPL (blue), OPL/ONL (magenta), ONL/EZ (cyan), EZ/RPE (red) and RPE/BM (blue).

2.4.2 Quantitative analysis

We tested 21960 diametral B-frames generated from 122 volumetric OCT scans to assess the performance of the neural network used in the optic disc boundary detection. The mean ± standard deviation of the testing loss (Eq. 1) between the predication maps and ground truth labels was 0.033 ± 0.028. We also calculated the Dice similarity coefficient (DSC) between the predicted final disc boundaries and corresponding manual delineations. The DSC was 0.92 ± 0.03 in normal and 0.91 ± 0.05 in glaucomatous eyes.

For the performance of peripapillary retinal boundaries segmentation, we calculated the absolute errors (μm, based on 3.125 μm/pixel) of the peripapillary retinal boundaries between our method and manual delineation (Table 2.1). The overall absolute errors were similar for both healthy and glaucomatous eyes. Because the NFL thickness is a critical feature for the detection and diagnosis of glaucoma, the NFL thickness based on our method was calculated and compared with the gold standard based on the manual delineation. The mean ± standard deviation value of the

NFL thickness differences (manual minus automated) was 2.14 ± 1.45 μm in glaucomatous and 1.67 ± 1.83 μm in normal eyes.

Table 2.1: Segmentation accuracy of our method

Boundaries	Healthy (Mean ± Std; μm)	Glaucoma (Mean ± Std; μm)
Vitreous/ILM	2.83 ± 1.12	3.25 ± 0.92
NFL/GCL	6.42 ± 0.36	6.64 ± 0.27
IPL/INL	5.59 ± 0.34	5.70 ± 0.23
INL/OPL	4.93 ± 0.66	4.97 ± 0.47
OPL/ONL	4.59 ± 0.90	5.37 ± 0.32
ONL/EZ	3.90 ± 0.83	4.22 ± 0.67
EZ/RPE	3.31 ± 0.77	3.74 ± 0.82
RPE/BM	3.52 ± 1.00	3.46 ± 0.94
Overall	4.09 ± 1.34	4.11 ± 1.16

As another test of performance for the algorithm presented here, we also compared our results to those obtained with our previous method, which was based exclusively on the graph search algorithm [48]. The comparisons of the segmentation accuracy of peripapillary retinal boundaries is shown in Table 2.2.

Table 2.2: Comparison of the peripapillary retinal boundaries segmentation

	Healthy		Glaucoma	
	NFL/GCL	All layers	NFL/GCL	All layers
Only graph search	9.34 ± 1.35 μm	4.78 ± 3.51 μm	14.26 ± 3.73 μm	11.45 ± 7.84 μm
With neural network	6.42 ± 0.36 μm	4.09 ± 1.34 μm	6.64 ± 0.27 μm	4.11 ± 1.16 μm
P-Value	0.006	0.09	0.004	0.002

Through Table 2.2, it is clear that the segmentation accuracy and stability were both improved after combining the neural network with the classic graph search.

2.4.3 Neural network analysis

Inside the neural network, the addition of the atrous-convolution layer in each atrous-block and the global block greatly improved the performance of the neural networks. In order to further analyze the neural network design, we compared the validation accuracy (based on DSC) of the peripapillary retinal layers segmentation between the four architectures below: original U-Net, U-Net + global block, U-Net + cascaded atrous-block, and proposed architecture (Table 2.3). Clearly, adding the cascaded atrous-convolution layers in the down and up sampling towers and global block at the end of the network critically improved the convergence of the neural network. In addition, the validation accuracies of the healthy and glaucoma data based on the inputs using only one channel (the middle one) instead of the 5 used in our algorithm were 84.11% and 83.53% respectively. These accuracies were about 2% lower than the accuracies shown in the last column of Table 2.3 which proved the five channels input design was effective.

Table 2.3: Comparison of the validation accuracy between different architectures

	Original U-Net	U-Net + global block	U-Net + cascaded atrous-block	Proposed Architecture
Healthy	23.14%	62.52%	81.79%	86.47%
Glaucoma	21.87%	61.26%	79.92%	85.31%

Fig. 2.10 shows example feature maps learned by the network in the normal convolution layers of the global block. It is clear that in each map the network is learning different retinal layers, as each map highlights specific layers or combinations thereof. The result of each map then yields a complete segmentation.

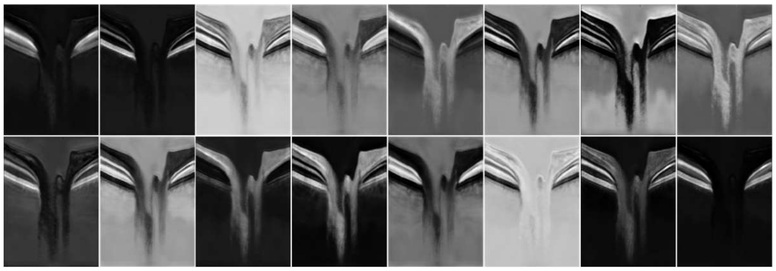


Figure 2.10: The sixteen feature maps of normal layers in the Global block.

2.5 Discussion and conclusion

The structure inside the optic disc, layer distortion near the optic disc, and vessel shadows constitute three major difficulties for peripapillary retinal boundaries segmentation. First, the optic disc needs to be segmented before the peripapillary retinal boundaries segmentation due to its unique anatomical structure. We solved this challenge by utilizing a geometric reorientation (diametral B-frames) and training a neural network on this more amenable geometry. The generated diametral B-frames have a high degree of structural consistency, which greatly increased the segmentation accuracy and stability of the optic disc boundary. In addition, the smoothing method that conformed to the anatomical features of optic disc also guaranteed the fidelity of the boundary. In the peripapillary retinal boundaries segmentation stage, the reason of not using diametral B-frames was that the diametral B-frames have the same directions with large vessels. The large vessel shadows could hardly influence the segmentation of single EZ + RPE layer but will influence the segmentation accuracy of six adjacent layers.

For the network architecture, the atrous-convolution layers and global block in the neural network could capture both local and global information at each pixel. The combination of the input data and neural network used in the design guaranteed that the peripapillary retinal boundaries segmentation would not be influenced by either disc distortion or vessel shadows.

Though the segmentation accuracy was greatly improved by using the neural network, limitations were also obvious. The performance of this method was limited by the depth and breadth of the training data set. In order to use this method on other OCT devices with different scan patterns or data from patients with different eye diseases, the training data set would need to be expanded. However, the complexity of the network architecture should be sufficient to learn either new pathologies or instruments, since even in these situations the OCT scans have nearly the same overall structure. In a future study, this method will be used on an expanded training data set to broaden its capabilities.

2.6 Conclusion

We combined a neural network with the traditional graph search method to segment both the optic disc and paripalliary retina boundaries in an optic disc centered volumetric OCT scan. The addition of the neural network greatly improved both segmentation accuracy and stability. The quantified tissue information, especially the NFL thickness and analysis of capillary plexuses, have the potential to pose a significant improvement in the diagnosis and early detection of glaucoma.

3 Diabetic Retinopathy Classification at Multiple Levels Based on Structural and Angiographic Optical Coherence Tomography

3.1 Abstract

Optical coherence tomography (OCT) and its angiography (OCTA) have several advantages for the early detection and diagnosis of diabetic retinopathy (DR). However, automated, complete DR classification frameworks based on both OCT and OCTA data have not been proposed. In this study, a convolutional neural network (CNN) based method is proposed to fulfill a DR classification framework using *en face* OCT and OCTA. A densely and continuously connected neural network with adaptive rate dropout (*DcardNet*) is designed for the DR classification. In addition, adaptive label smoothing was proposed and used to suppress overfitting. Three separate classification levels are generated for each case based on the International Clinical Diabetic Retinopathy scale. At the highest level the network classifies scans as referable or non-referable for DR. The second level classifies the eye as non-DR, non-proliferative DR (NPDR), or proliferative DR (PDR). The last level classifies the case as no DR, mild and moderate NPDR, severe NPDR, and PDR. We used 10-fold cross-validation with 10% of the data to assess the network's performance. The overall classification accuracies of the three levels were 95.7%, 85.0%, and 71.0% respectively. A reliable, sensitive and specific automated classification framework for referral to an ophthalmologist can be a key technology for reducing vision loss related to DR.

3.2 Introduction

Optical coherence tomography (OCT) can generate depth-resolved, micrometer-scale-resolution images of ocular fundus tissue based on reflectance signals obtained using interferometric analysis of low coherence light [13]. By scanning multiple B-frames at the same position, change in the OCT reflectance properties can be measured as, e.g., decorrelation values to differentiate vasculature from static tissues. This technique is called OCT angiography (OCTA), and it can provide high-resolution images of the microvasculature of retina [21-23]. Numerous investigators explored OCTA in the detection and diagnosis of various ocular diseases, and demonstrated many advantages when compared to traditional imaging modalities such as fundus photography or fluorescein angiography [17-20]. Among these is diabetic retinopathy (DR), which affects the retinal capillaries and is a leading cause of preventable blindness globally [1]. OCT-based biomarkers such as central macular thickness and OCTA-based biomarkers such as avascular areas have demonstrated superior potential for diagnosing and classifying DR compared to traditional imaging modalities

[17-20]. However, recently emerged automated deep-learning classification methods were largely based on color fundus photography (CFP) [7-10]. Therefore, taking advantages of both powerful deep learning tools and innovative structural and angiographic information, we developed an automated framework that can perform a full DR classification (across datasets including all DR grades) based on *en face* OCT and OCTA projected from the same volumetric scans.

In order to improve classification accuracy and reliability, a new convolutional neural network architecture was designed based on dense and continuous connection with adaptive rate dropout (*DcardNet*). The system produces three classification levels to fulfill requests in clinical diagnosis. Non-referable and referable DR (nrDR and rDR) are classified in the first level. No DR, non-proliferative DR (NPDR), and proliferative DR (PDR) are in the second classification level. No DR, mild and moderate NPDR, severe NPDR, and PDR are in the third level. While training *DcardNet*, adaptive label smoothing was used to reduce overfitting. To improve interpretability and help understand which regions contribute to the diagnosis, class activation maps (CAM) were also generated for each DR class [71].

3.3 Related works

Several methods for the automated classification of DR severity have been proposed since the convolutional neural network (CNN) became the most widely used solution for image classification problems [7-10, 72-76]. Most of these methods are based on CFP, which is a traditional and commonly used technique capable of DR diagnosis. R. Gargeya *et al*. proposed a machine learning based method to classify CFP images as healthy (no retinopathy) or having DR [7]. They used a customized *ResNet* architecture [77] to extract features from the input CFPs. The final classification was performed on a decision tree classification model by using the combination of extracted features and three metadata variables. They achieved a 0.97 area under receiver operating curve (AUC) after 5-fold stratified cross-validation. In addition, a visualization heatmap was generated for each input CFP based on visualization layer in the end of their network [71]. V. Gulshan *et al*. used Inception-v3-based transfer learning to classify the CFP mainly as rDR and nrDR [9]. In the validation tests on two publicly available datasets (eyePACS-1 and Messidor-2), they achieved an AUC of 0.991 and 0.990, respectively. M. D. Abramoff *et al*. also proposed a CNN-based method to classify CFP images as rDR and nrDR and achieved an AUC of 0.980 during validation [8]. For more detailed DR classification, R. Ghosh *et al.* proposed a CNN-based method to classify the CFP images into both two-class (no DR vs DR) and five severities:

no DR, mild NPDR, moderate NPDR, severe-NPDR, and PDR [10]. They achieved an overall accuracy of 85% for the classification into five severities.

However, all of the above methods were based on the CFP. Compared to CFP, OCT and OCTA can provide more detailed information (i.e. 3D, high-resolution, vascular and structural imaging). An automated DR classification framework based on OCT/OCTA could reduce the number of procedures that must be performed in the clinic if OCT/OCTA can deliver the same diagnostic value as other modalities, which will ultimately reduce clinical burden and healthcare costs. Therefore, an automated framework for DR classification based on OCT and OCTA data is desirable.

H. S. Sandhu *et al.* proposed a computer-assisted diagnostic (CAD) system based on quantifying three OCT features: retinal reflectivity, curvature, and thickness [72]. A deep neural network was used to classify each case as no DR or NPDR based on those three retinal features and achieved an overall accuracy of 93.8%. The same group also proposed a CAD system for DR classification based on quantified features from OCTA [73]: blood vessel density, foveal avascular zone (FAZ) area, and blood vessel caliber and trained a support vector machine with a radial basis function kernel. They achieved an overall accuracy of 94.3%. However, these systems examined and classified only no DR and NPDR cases. M. Alam *et al.* proposed a support vector machine-based DR classification CAD system using six quantitative features generated from OCTA: blood vessel tortuosity, blood vascular caliber, vessel perimeter index, blood vessel density, foveal avascular zone area, and foveal avascular zone contour irregularity [74]. They achieved 94.41% and 92.96% accuracies for control versus disease (NPDR) and control versus mild NPDR. In addition, they achieved 83.94% accuracy for multiclass classification (control, mild NPDR, moderate NPDR, and severe NPDR). However, as only pre-determined features were incorporated into this model, it could not learn from the much richer feature space latent in the entire OCTA data. In addition, CAD systems based on only empirically selected biomarkers have limited potential for further improvements even as the number of available datasets grows. M. Heisler *et al.* proposed a DR classification method based on *en face* OCT and OCTA images using ensemble networks [75]. Each case was classified as nrDR or rDR and they achieved an overall accuracy of 92.0%. In addition, the CAM of each *en face* image was generated. However, only 2-class classification was performed in this study. Therefore, an OCT and OCTA based DR classification framework capable of fulfilling different clinical requests and generating CAMs is needed.

There are two major challenges for OCT and OCTA-based DR classification. First, OCTA generates a much greater detailed image of the vasculature than traditional CFP. Extracting classification related features from such detailed information is much more challenging compared with the CFP-based classification. The second challenge is the relatively small size of the available OCT and OCTA dataset, compared to the very large CFP dataset used in the previous CFP-based networks. This challenge can lead to a severe overfitting problem during the training of the network. Addressing these challenges requires a network architecture with not only efficient convergence but also low overfitting. We designed a densely and continuously connected neural network with adaptive rate dropout and used it to perform a DR classification in three levels. We also produced corresponding CAMs in this study. In addition, adaptive label smoothing was proposed to further reduce overfitting. The main contributions of the present work are as follows:

- We present an automated framework for the DR classification and CAM generation based on both OCT and OCTA data. In this framework, three DR classification levels are performed for the first time.
- We propose a new network architecture based on dense and continuous connections with adaptive rate dropout.
- We propose an adaptive label smoothing to suppress overfitting and improve the performance generalization of the trained network.

3.4 Materials

In this study, 303 eyes from 250 participants, including healthy volunteers and patients with diabetes (with or without DR) were recruited and examined at the Casey Eye Institute, Oregon Health & Science University. Masked trained retina specialists graded the disease severity based on Early Treatment of Diabetic Retinopathy Study (ETDRS) scale [11] using corresponding 7-field fundus photography. Based on the recent studies on referable retinopathy level shown in the International Clinical Diabetic Retinopathy scale [12], we defined referable retinopathy as the equivalent ETDRS grade, which is grade 35 or worse. The participants were enrolled after informed consent in accordance with an Institutional Review Board (IRB # 16932) approved protocol. The study was conducted in compliance with the Declaration of Helsinki and Health Insurance Portability and Accountability Act.

The macular region of each eye was scanned once or twice (after a one-year gap) using a commercial 70-kHz spectral-domain OCT system (Avanti RTVue-XR, Optovue Inc) with 840-nm central wavelength. The scan regions were 3.0 × 3.0 mm and 1.6 mm in depth (304 × 304 × 640 pixels) centered on the fovea. Two repeated B-frames were captured at each line-scan location to calculate the OCTA decorrelation values. The blood flow of each line-scan location was detected using the split-spectrum amplitude-decorrelation angiography algorithm based on the speckle variation between two repeated B-frames [23, 57]. The OCT structural images were obtained by averaging two repeated B-frames. For each data set, two volumetric raster scans (one x-fast scan and one y-fast scan) were registered and merged through an orthogonal registration algorithm to reduce motion artifacts [58].

For each pair of OCT and OCTA data, the following retinal layers were automatically segmented (Fig. 3.1) based on the commercial software in the spectral-domain OCT system (Avanti RTVue-XR, Optovue Inc): inner limiting membrane (ILM), nerve fiber layer (NFL), ganglion cell layer (GCL), inner plexiform layer (IPL), inner nuclear layer (INL), outer plexiform layer (OPL), outer nuclear layer (ONL), ellipsoid zone (EZ), retinal pigment epithelium (RPE), and Bruch's membrane (BM). In addition, for the cases with severe pathologies, the automated layer segmentation was manually corrected by graders using the customized COOL-ART software [78].

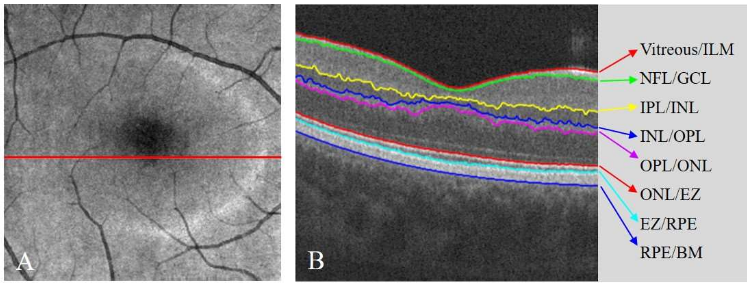


Figure 3.1: The automated retinal layer segmentation from OCT structural image scanned from a healthy participant. (A) The *en face* average projection of the whole OCT structure. (B) The B-frame corresponding to the position of red line in (A). The eight boundaries of the seven main retinal layers were segmented.

Based on the segmented boundaries, six *en face* projections from OCT reflectance signals and OCTA decorrelation values were obtained and used to build a six-channel input data (Fig. 3.2). The first three channels were the inner retinal thickness map (z-axis distance between the Vitreous/ILM and OPL/ONL), inner retinal *en face* average projection (Vitreous/ILM to OPL/ONL) and EZ *en face* average projection (ONL/EZ to EZ/RPE) based on the volumetric OCT (Fig. 3.2(A)-(C)). The last three channels were the *en face* maximum projections of the superficial vascular complex (SVC), intermediate capillary plexus (ICP), and deep capillary plexus (DCP) based on the volumetric OCTA. (Fig. 3.2(D)-(F)) [70]. The SVC was defined as the inner 80% of the ganglion cell complex (GCC), which included all structures between the ILM and IPL/INL border. The ICP was defined as the outer 20% of the GCC and the inner 50% of the INL. The DCP was defined as the remaining slab internal to the outer boundary of the

OPL [18, 41]. In addition, the projection-resolved OCTA algorithm was applied to all OCTA scans to remove flow projection artifacts in the deeper plexuses [68, 69].

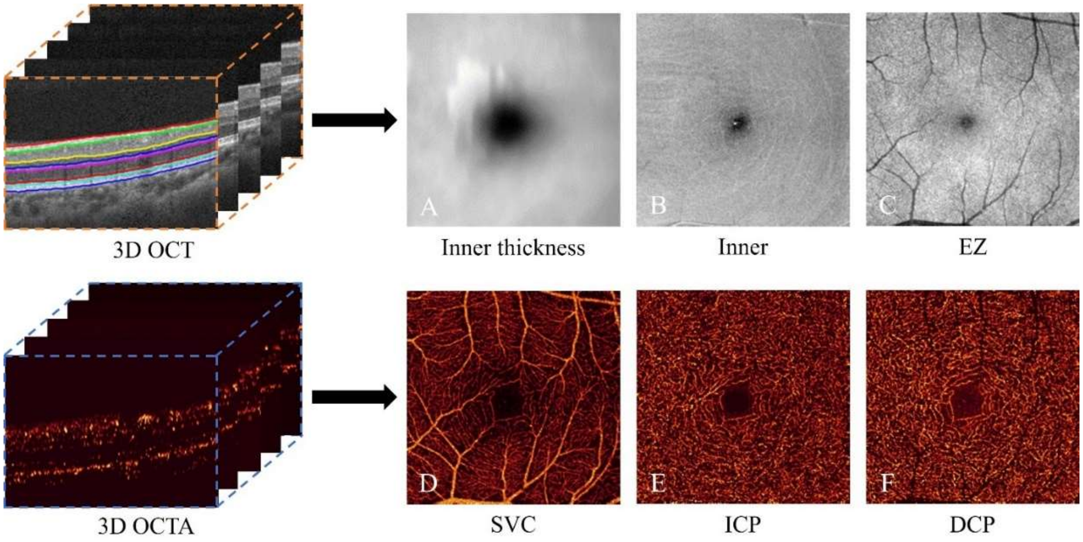


Figure 3.2: The six input channels based on the OCT and OCTA data scanned from a moderate NPDR participant. (A) Inner retinal thickness map. (B) Inner retinal *en face* average projection. (C) Ellipsoid zone (EZ) *en face* average projection. (D) Superficial vascular complex (SVC) *en face* maximum projection. (E) Intermediate capillary plexus (ICP) *en face* maximum projection. (F) Deep capillary plexus (DCP) *en face* maximum projection.

Three classification levels of each input data were built based on the ETDRS grades as scored by three ophthalmologists (Fig. 3.3). The first label was for 2 classes: nrDR and rDR. The second label was for 3 classes: no DR, NPDR and PDR. The last label was for 4 classes: no DR, mild and moderate NPDR, severe NPDR and PDR. Mild and moderate NPDR were not separated due to a lack of measurements on eyes with NPDR from which to procure make a balanced dataset. For each level, follow up scans (scanned after a one-year gap) that did not have a class change were removed from the dataset for corresponding level to avoid correlation. Therefore, number of scans for each classification level was different (Table 3.1).

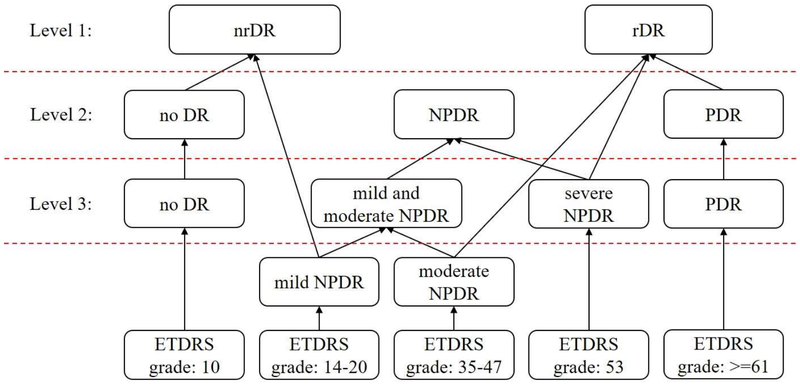


Figure 3.3: The relations between the ETDRS grades and three levels of DR classifications.

Table 3.1: Data distribution of three classification levels

Classifications	Number of scans	Whole data size
nrDR	95	294
rDR	199	
no DR	85	298
NPDR	128	
PDR	85	
no DR	85	302
mild and moderate NPDR	82	
severe NPDR	50	
PDR	85	

3.5 Methods

The architecture of the *DcardNet* is shown in Fig. 3.4. The main feature of this architecture is that the input tensor for each bottleneck block was the concatenation of the output tensors from at most the *C* previous bottleneck blocks with adaptive dropout rates. The dropout rate [79] of each bottleneck was adaptively adjusted based on the distance between the depths of this block and the block to be calculated next. In addition, the size (height and width) of the output tensor was halved *M* times through transfer blocks to perform down-sampling. Detailed information for this method is described below.

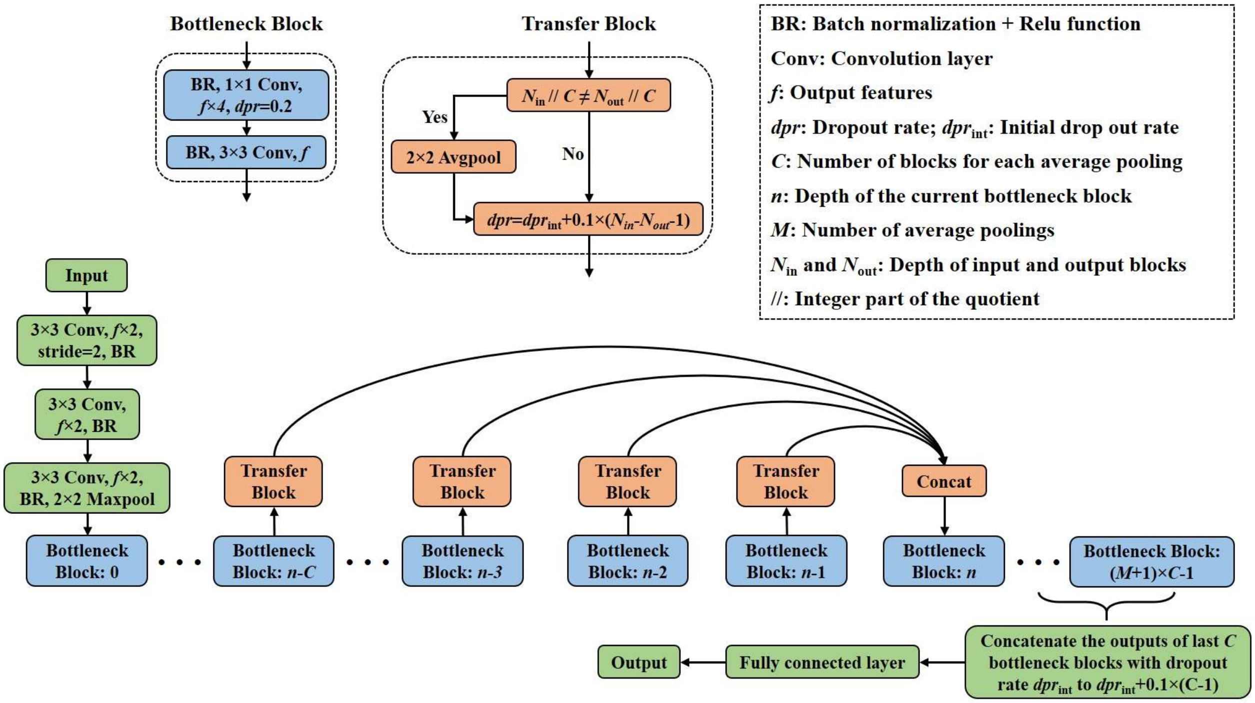

Bottleneck Block
BR, 1×1 Conv, f×4, dpr=0.2
BR, 3×3 Conv, f
Transfer Block
N_{in} // $C \neq N_{out}$ // C
Yes
No
2×2 Avgpool
$dpr=dpr_{int}+0.1\times(N_{in}-N_{out}-1)$
BR: Batch normalization + Relu function
Conv: Convolution layer
f: Output features
dpr: Dropout rate; dpr_{int}: Initial drop out rate
C: Number of blocks for each average pooling
n: Depth of the current bottleneck block
M: Number of average poolings
N_{in} and N_{out}: Depth of input and output blocks
//: Integer part of the quotient
Input
3×3 Conv, f×2, stride=2, BR
3×3 Conv, f×2, BR
3×3 Conv, f×2, BR, 2×2 Maxpool
Bottleneck Block: 0
Transfer Block
Bottleneck Block: n-C
Bottleneck Block: n-3
Bottleneck Block: n-2
Bottleneck Block: n-1
Concat
Bottleneck Block: n
Bottleneck Block: (M+1)×C-1
Concatenate the outputs of last C bottleneck blocks with dropout rate dpr_{int} to dpr_{int}+0.1×(C-1)
Fully connected layer
Output

Figure 3.4: The network architecture of the proposed *DcardNet*.

3.5.1 Bottleneck block

A 1×1 convolution is widely used as a bottleneck layer before 3×3 convolutions to improve the computational efficiency by reducing the number of input features [80]. Our network uses two convolutional layers in the bottleneck block. A 1×1 convolution layer with *f×4* output features and 0.2 dropout rate [79] was used as the first convolutional layer. The second convolutional layer in the bottleneck block is a 3×3 convolution with *f* output features. In addition, a batch normalization [63] and rectified linear unit (ReLU) activation function [81, 82] were used before each convolutional layer.

3.5.2 Transfer block

Before the concatenation of the output tensors from at most the last *C* bottleneck blocks, a transfer block was used to perform the adaptive rate dropout. The dropout rate (*dpr*) of the output tensor from each bottleneck block was calculated as

$$dpr = dpr_{\text{int}} + 0.1 \times \left(N_{in} - N_{out} - 1 \right) \tag{7}$$

where dpr_{int} is the initial dropout rate, N_{out} is the depth of each bottleneck block which is to be concatenated, and N_{in} is the depth of the bottleneck block that will use the concatenated tensor as input. In order to fulfill the down-sampling, the size of the tensor is halved before dropout using 2 × 2 average pooling if the integer part of the quotients between N_{out} / C and N_{in} / C were not equal.

3.5.3 Dense and continuous connection with adaptive dropout

Dense connectivity has been proposed by G. Huang *et al.* [80] and used in *DenseNet* to improve information flow. However, the dense connection was only used within each dense block, not the whole network. In the *DcardNet*, the dense connection was continuously used in the whole network to further improve the information flow. In addition, the size and weight of each concatenated bottleneck block was adaptively adjusted using the transfer block to fulfill down-sampling and differentiate the importance of the information in different bottleneck blocks. The input tensor to each bottleneck block was

$$x_n^{in} = \text{concat}\left[T\left(x_{n-1}^{out} \right),\ T\left(x_{n-2}^{out} \right),\ \ldots,\ T\left(x_{\max(0,\ n-C)}^{out} \right) \right] \tag{8}$$

where x_n^{in} and x_n^{out} are the input and output tensors of the n_{th} bottleneck block, concat[●] is the concatenation operation, and T(●) is the transfer block.

3.5.4 Adaptive label smoothing and data augmentation

The goal of training the network is high overall classification accuracy, defined as

$$Acc = \frac{1}{Num} \times \sum_{i=1}^{Num} a_i$$
$$a_i = \begin{cases} 1 & \arg\max\left(g_i\right) = \arg\max\left(p_i\right) \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where g_i and p_i are the i^{th} ground truth and predicted labels at a given classification level, respectively, and Num is the number of scans in the dataset. However, network parameters were optimized by minimizing the negative cross entropy loss

$$loss\left(g, p\right) = -\sum_{i=1}^{K} p_i \times \log g_i \tag{10}$$

where K was the number of classes. According to (9), the prediction will always be right as long as the location of the largest value in the predicted label is the same as the ground truth label. Once this has been achieved, continuing to reduce the negative cross entropy loss only marginally improves the overall classification accuracy, and may lead to overfitting [83, 84]. Therefore, in this study, each ground truth label was gradually smoothed by an amount s based on the class differences between the true class and false classes. Since class labels were sorted along a scale of DR severity, the smoothed class labels respect the decreasing likelihood that the label was misidentified. The labels at all three levels were smoothed according to

$$g_i = \begin{cases} 1.0 - s_i & \text{true class} \\ s_i \times \dfrac{\dfrac{1}{\left|t_j - t_i\right|}}{\displaystyle\sum_{j=1}^{K-1} \dfrac{1}{\left|t_j - t_i\right|}} & \text{other classes} \end{cases} \tag{11}$$

where s_i is the reduction in the value of true class, and t_j and t_i respectively were the indexes of each incorrect class and the true class in i^{th} label.

Variation between different OCTA data sets is intrinsically high. Some inputs converge well in a short time, but the convergence of other inputs might change significantly and repeatedly. According to the gradient of the weight variables in the network (12), the weights w will converge to an input faster when the difference between the predication and corresponding ground truth label gets larger, and slower when the difference is smaller:

$$\frac{\partial loss}{\partial w} = \frac{1}{Num}\sum_{i=1}^{Num} x_i \left(p_i - g_i\right) \tag{12}$$

where x_i is the i^{th} input, p_i and g_i are the corresponding prediction and ground truth. In order to further increase the rate of convergence on the mispredicted inputs and decrease the rate of convergence on the correctly predicted inputs, the label smoothing value s for each label was adaptively adjusted based on the prediction results during each training step according to

$$s_i = \begin{cases} \min\left(s_i + d,\ s_{\max}\right) & \arg\max\left(g_i\right) = \arg\max\left(p_i\right) \\ \max\left(s_i - d,\ 0.0\right) & \text{otherwise} \end{cases} \tag{13}$$

where s_i is the smoothing value for the i^{th} label, and d is an adjustment for each s_i and $s_{\max}$ was the upper limit of the smoothing value. Based on (13), the convergence rate of the inputs which were correctly predicted during each training iteration would be much lower than the other inputs.

In addition, no class weight balancing was used in training because adaptive label smoothing can achieve the same effect. Class weight balancing can tell the model to pay more attention to samples from an under-represented class by appropriately weighting the loss function to compensate for data deficiencies during training. Alternatively, the same effect could be achieved by smoothing the ground truth labels while maintaining the loss function (since classes with small label differences will contribute less to the loss). This is the approach taken in adaptive label smoothing, which has the additional advantage of allowing the smoothing function to updated during training to expedite balanced convergence.

Data augmentation is another method used for improving the performance generalization of a trained network. In this study, the number of training datasets was increased by a factor of 8 by including combinations of 90° rotations and horizontal and vertical flips (there is a grand total of 7 unique combinations of these transformations available). In order to make sure the selected inputs in each training batch were based on different cases, only one of the data augmented patterns (including the original inputs) was randomly chosen for each input during each training batch selection.

3.5.5 Implementation details

The maximum number of the concatenated bottleneck blocks *C* was set to 4. The number of output features *f* after each bottleneck block was set to 24. *M* was set to 3 which meant overall 16 bottleneck blocks were used in this architecture. This specific architecture is called *DcardNet*-36 which means overall 35 convolutional layers and 1 fully connected layer were used in the whole network, which yields 9264960 trainable parameters. In addition, for the 2-class, 3-class and 4-class DR classifications, the initial label smoothing value s_i were set to 0.05, 0.005 and 0.005, adjusting steps *d* were empirically chosen as 0.001, 0.0001 and 0.0001, and upper limits $s_{\max}$ were set to 0.1, 0.01 and 0.01, respectively.

In order to ensure the credibility of the overall accuracy, 10-fold cross-validation was used on the DR classification at each level. In each fold, 10% of the data (with the same class distribution as the overall data set) was split on a patient-wise basis (scans from same patient only included in one set) and used exclusively for testing. The parameters were optimized by a stochastic gradient descent optimizer with Nesterov momentum (momentum = 0.9). During the training process, a batch size of 10 was empirically chosen and the total training steps for the three-level DR classification were set to 8000. In addition, an initial learning rate $lr_{init} = 0.01$ with cosine decay was used in this study [85]:

$$\begin{aligned} lr_{curr} &= lr_{init} \times \left(0.97 \times d + 0.03\right) \\ d &= \frac{1}{2}\left[1 + \cos\left(\pi \times step_{curr} / step_{stop}\right)\right] \end{aligned} \tag{14}$$

where lr_{curr} was the current learning rate, $step_{curr}$ was the current training step and $step_{stop}$ was the step at which the learning rate ceased to decline. In this study, the $step_{stop}$ was empirically chosen as 6000.

Both training and testing were implemented in Tensorflow version 1.13 on Windows 10 (64 Bit) platform. The workstation used in this study has an Intel (R) Core (TM) i7-8700K CPU @ 3.70GHz, 64.0 GB RAM and NVIDIA RTX 2080 GPU. The training time was 7 minutes for each training process (70 minutes for 10-fold cross-validation) and the inference time for a new case was 8 seconds.

3.6 Experiments

The overall prediction accuracy (the number of correctly predicted case divided by the number of whole data set) and corresponding 95% confidence interval (95% CI) varied across the three classification levels (Table 3.2). In addition, the 10 models trained during the 10-fold cross validation were also used to predict on a balanced external dataset with 30 scans to further demonstrate the generalization of our DR classification framework. The overall accuracies of 2-class, 3-class, and 4-calss DR classification on the external dataset are 93.3% ± 2.4%, 82.7% ± 2.8%, and 68.7% ± 3.8%, respectively. Though the accuracies on the external dataset are about 2% - 3% lower than the accuracies on our local testing dataset, the results still show that our DR classification framework has a strong generalization on external dataset.

Table 3.2: DR classification accuracy at multiple levels

	2-class	3-class	4-class
10-fold Accuracy (mean ± std)	95.7% ± 3.9%	85.0% ± 3.6%	71.0% ± 4.8%
95% CI	93.3% - 98.1%	82.8% - 87.2%	68.0% - 74.0%

The sensitivity and specificity for each severity class in all three DR classification levels also varied and is shown in Table 3.3. The classification sensitivity of the severe NPDR was much lower than other classes. This is because the differences between adjacent levels of severity are much smaller than the variations between no DR, NPDR and PDR. In addition, the number of severe NPDR cases was also much smaller than other classes.

Table 3.3: Sensitivity and specificity of each class in three classification levels

Levels	DR severities	Sensitivities (mean, 95% CI)	Specificities (mean, 95% CI)
2-class	nrDR	91.0%, 86.4% - 95.6%	98.0%, 96.4% - 99.6%
	rDR	98.0%, 96.4% - 99.6%	91.0%, 86.4% - 95.6%
3-class	no DR	86.7%, 81.3% - 92.1%	93.3%, 91.8% - 94.8%
	NPDR	85.4%, 83.9% - 86.9%	89.4%, 87.1% - 91.7%
	PDR	82.5%, 78.5% - 86.5%	93.7%, 91.7% - 95.7%
4-class	no DR	86.3%, 83.9% - 88.7%	87.8%, 85.9% - 89.7%
	mild and moderate NPDR	81.3%, 77.2% - 85.4%	84.6%, 82.6% - 86.6%
	severe NPDR	12.0%, 2.0% - 22.0%	100.0%, 100.0% - 100.0%
	PDR	87.8%, 85.6% - 90.0%	87.1%, 85.1% - 89.1%

We also produced CAMs of inputs with different DR classes (Fig. 3.5), indicating the network's attention within the different DR classes. The macular regions with high positive values in the CAMs indicate they have high positive influences on the classification for the true class. On the contrary, the regions with nearly zero values in the CAMs have no or negative influence on the classification. In CAMs of cases without DR and cases with PDR regions close to the fovea had the highest positive influences on the classification. However, the vasculature around the fovea had the highest positive influences on the classification of NPDR cases. This difference may be caused by the appearance of features like fluids or non-perfusion areas. Overall, the areas with higher values (yellow to red) in the CAM were the regions the network used for decision making. By considering the CAMs, a doctor could judge the reasonableness of the automated DR classification and pay more attention on the high-value-areas during the diagnosis.

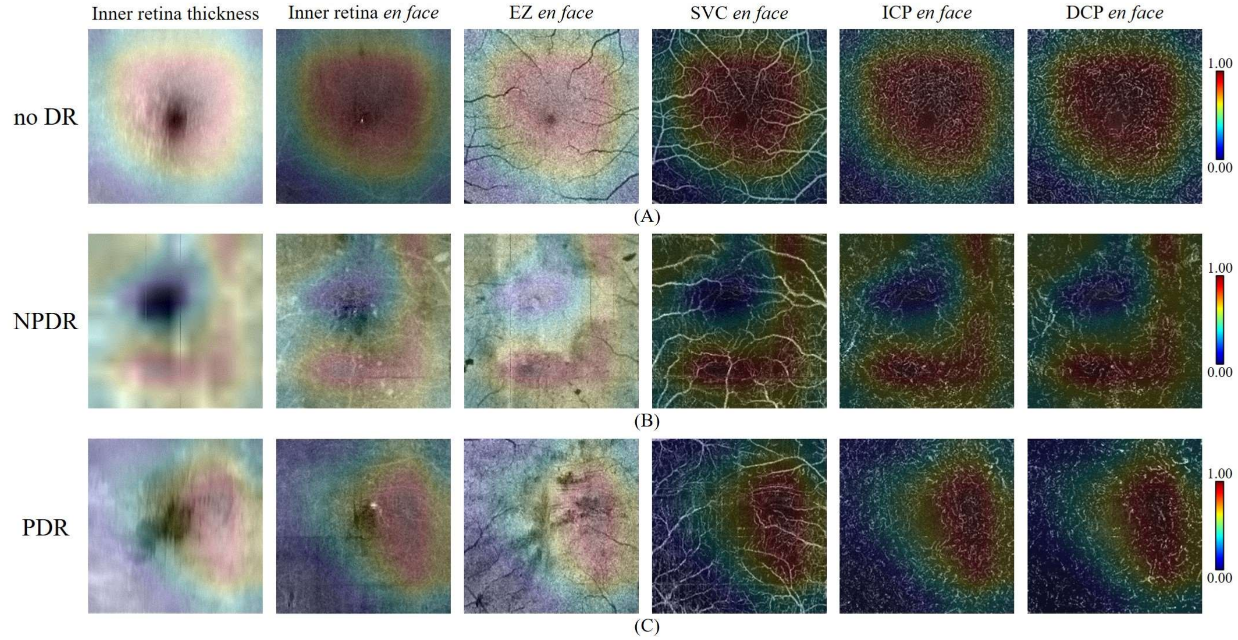


Figure 3.5: The CAMs of three correctly predicted cases with different DR classes. In each row, the inner retina thickness map, inner retinal *en face* OCT, EZ *en face* OCT, SVC *en face* OCTA, ICP *en face* OCTA, and DCP *en face* OCTA were overlaid by the corresponding CAMs. In addition, the color bar of each CAM was on the right side of each row. (A) CAMs of case without DR. (B) CAMs of a case with NPDR. (C) CAMs of a case with PDR.

To further quantitatively analyze the proposed method, we performed five comparisons on our local dataset to investigate the accuracy and stability of the proposed DR classification framework. First, we compared the performance of the network trained on combined OCTA and OCT structural data inputs to the network trained on either structural OCT or OCTA data separately. Second, we compared the performances of our network with no dropout, standard dropout (0.2 dropout rate), and proposed adaptive dropout. Third, we compared the performances of our network with traditional class weight balancing and proposed adaptive label smoothing. Fourth, we compared the performances of different network architectures (*ResNet* [77], *DenseNet* [80], *EfficientNet* [86], VGG16 [87], VGG19 [87], ResNet-v2 [88], Inception-v4 [89] and the proposed *DcardNet*) with or without the adaptive label smoothing. Finally, we compared the performances of our method with a previously proposed ensemble network [75] on the 2-class DR classification. In addition, all the results (including ours) in the comparisons below (sections *A*, *B*, *C*, *D* and *E*) were based on 5-fold cross-validation with 20% exclusively reserved for testing.

3.6.1 Comparison between the three input patterns

The inputs had six channels obtained from both OCT and OCTA data. In order to verify the necessity of this input design, comparison of classification accuracies between the OCT-based inputs, OCTA-based inputs, and OCT+OCTA-based inputs were performed. The network used a set of 6 *enface* images as input. From structural OCT-based we included an inner retina thickness map, an inner retina average projection, and an EZ average projection. The OCTA-based inputs are *enface* maximum projection of the SVC, ICP, and DCP. Table 3.4 shows the overall accuracies of the three levels of DR classification based on three different input patterns. Compared to the OCT-based input, the proposed input design greatly increased (≈ 10%) the overall accuracies of 3 and 4-class DR classification. Compared to the OCTA-based input, the overall accuracies also increased for 3-class DR classification. For the 4-class DR classification, though the overall accuracy of OCT+OCTA-based was the same as only OCTA-based, the sensitivities of OCT+OCTA-based shown in Table VI were more balanced than only OCTA-based. For the 2-class DR classification, which has the same accuracy based on three different input patterns, the CAMs only based on OCT and OCTA were both calculated to study the different influences from OCT and OCTA (Fig. 3.6). Through first row, we can see the CAMs only based on OCT were both convex polygons centered on the fovea of nrDR and rDR eyes. On the contrary, the two CAMs only based on OCTA were quite different and have more complicated shapes. This comparison shows that more detailed information was used in the DR classification only based on OCTA.

Table 3.4: Comparison of the DR classification accuracies at multiple levels between three different input patterns

Inputs patterns	2-class (mean, 95% CI)	3-class (mean, 95% CI)	4-class (mean, 95% CI)
OCT-based	94.2%, 91.1% - 97.3%	63.7%, 60.4% - 67.0%	54.7%, 52.1% - 57.3%
OCTA-based	94.2%, 90.5% - 97.9%	74.0%, 69.7% - 78.3%	64.7%, 61.5% - 67.9%
OCT+OCTA-based	94.2%, 91.9% - 96.5%	76.7%, 73.4% - 80.0%	64.7%, 61.5% - 67.9%

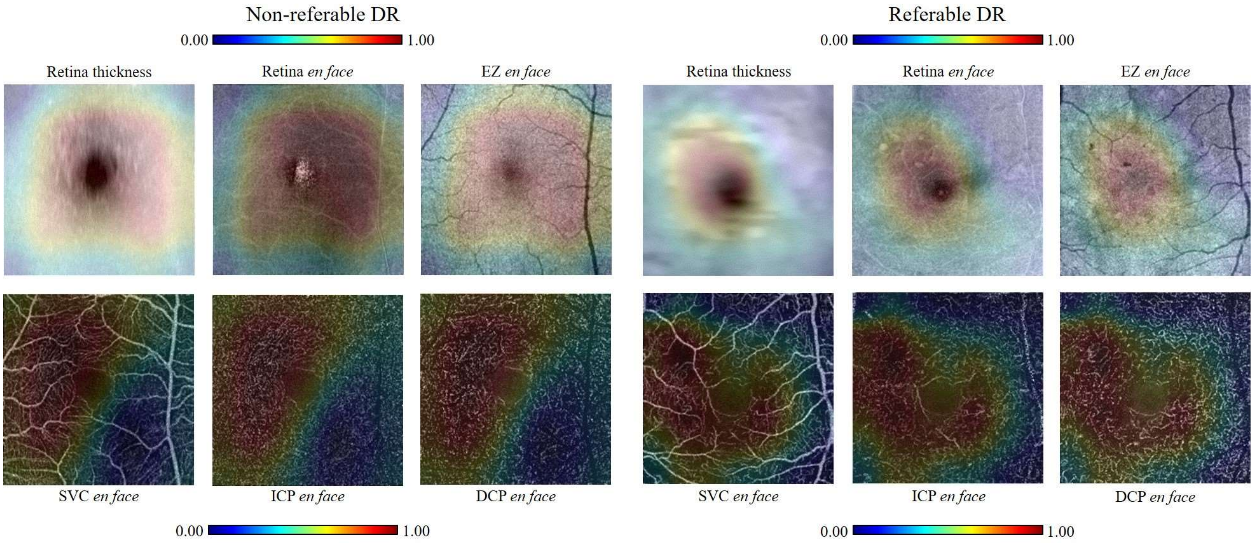


Figure 3.6: Comparison between CAMs generated from the two-class DR classification only based on OCT or OCTA. First row: CAMs from the OCT-only network overlaid on the three *en face* OCT layers scanned from nrDR and rDR eyes. Second row: CAMs from the OCTA-only network overlaid on the corresponding OCTA.

Table 3.5 summarizes the comparison of the sensitivities and specificities between the three input patterns and 4 different DR classes. The combined input design improved the sensitivities of two intermediate severity classes. While the overall accuracies of OCTA-based input and OCT+OCTA-based input were the same, using OCT+OCTA based input reduced the variation of sensitivities between different DR severities.

Table 3.5: Comparison of the sensitivities and specificities of four DR severities between three different inputs patterns

DR severities		OCT-based (mean, 95% CI)	OCTA-based (mean, 95% CI)	OCT+OCTA-based (mean, 95% CI)
no DR	Sensitivity	80.0%, 75.4% - 84.6%	84.7%, 80.1% - 89.3%	82.4%, 77.3% - 87.5%
	Specificity	77.2%, 75.5% - 78.9%	84.2%, 82.5% - 85.9%	85.1%, 82.8% - 87.4%
mild and moderate NPDR	Sensitivity	36.3%, 31.7% - 40.9%	63.8%, 59.2% - 68.4%	66.2%, 63.2% - 69.2%
	Specificity	80.5%, 78.2% - 82.8%	82.3%, 79.7% - 84.9%	81.8%, 78.6% - 85.0%
severe NPDR	Sensitivity	0.0%, 0.0% -0.0%	2.0%, 0.0% - 5.9%	4.0%, 0.0% - 8.8%
	Specificity	100.0%, 100.0% - 100.0%	100.0%, 100.0% - 100.0%	100.0%, 100.0% - 100.0%
PDR	Sensitivity	78.8%, 76.0% - 81.6%	82.4%, 77.3% - 87.5%	81.2%, 76.9% - 85.5%
	Specificity	81.4%, 80.0% - 82.8%	85.1%, 82.8% - 87.4%	85.6%, 83.9% - 87.3%

3.6.2 Comparison between different dropout strategies

The performances comparison between our network with three different dropout strategies were shown in Table 3.6. Proposed network with adaptive dropout shown the highest accuracies in all three DR classification levels. The accuracy increasing based on adaptive dropout was most obvious in the 3-class DR classification.

Table 3.6: Comparison of the overall accuracy between three different dropout strategies

Dropout strategies	2-class (mean, 95% CI)	3-class (mean, 95% CI)	4-class (mean, 95% CI)
no dropout	93.6%, 91.7% - 95.5%	73.3%, 71.9% - 74.7%	64.3%, 62.6% - 66.0%
Standard dropout (0.2)	94.2%, 90.5% - 97.9%	75.3%, 73.4% - 77.2%	64.3%, 62.6% - 66.0%
Adaptive dropout	94.2%, 91.9% - 96.5%	76.7%, 73.4% - 80.0%	64.7%, 61.5% - 67.9%

3.6.3 Comparison between class weight balancing and adaptive label smoothing

To gauge the ability of adaptive label smoothing to compensate for the unbalanced classes in our data set, we compared the performance of our network with class weight balancing, adaptive label smoothing, or both (Table 3.7). At each classification level, the network trained with adaptive label smoothing outperformed both class weight balancing and the network using both class weight and adaptive label smoothing.

Table 3.7: Comparison of the overall accuracy between three different weight balancing strategies

Weight balancing strategies	2-class (mean, 95% CI)	3-class (mean, 95% CI)	4-class (mean, 95% CI)
Class weight balancing	93.6%, 91.7% - 95.5%	75.3%, 72.7% - 77.9%	64.3%, 61.7% - 66.9%
Adaptive label smoothing	94.2%, 91.9% - 96.5%	76.7%, 73.4% - 80.0%	64.7%, 61.5% - 67.9%
Both strategies	94.2%, 1.9% - 96.5%	76.0%, 74.2% - 77.8%	63.9%, 61.3% - 66.5%

3.6.4 Comparison between different network architectures

We also compared the performances of *ResNet*-18, *EfficientNet*-B0, and *DenseNet*-53, *VGG16*, *VGG19*, *ResNet-v2*-50, *Inception-v4* and proposed *DcardNet*-36 with or without adaptive label smoothing for the DR classification at multiple levels on the same dataset. Among them, *DenseNet*-53 is a modified *DenseNet* architecture with 53 layers (52 convolution and 1 dense layers) which achieved the highest accuracy compared to other *DenseNet* architectures. In addition, no transfer learning was used in the training of all the networks above and all the final models were trained from scratch with empirically selected optimal hyper-parameters. Table 3.8 shows the overall accuracies of the three levels of DR classification based on all eight network architectures. Our network architecture with or without adaptive label smoothing achieved the highest accuracies on both 2-class and 3-class DR classifications. Only the 4-class DR classification accuracies of *VGG16* and *ResNet-v2*-50 were about 1% higher than ours. In addition, the use of the proposed adaptive label smoothing improved the classification accuracies of all architectures.

Table 3.8: Comparison of the overall accuracies between different architectures with or without adaptive label smoothing

Architectures	Label pattern	2-class (mean, 95% CI)	3-class (mean, 95% CI)	4-class (mean, 95% CI)
ResNet-18 [18]	Normal label	92.9%, 91.7% - 94.1%	71.7%, 69.9% - 73.5%	64.0%, 61.3% - 66.7%
	Adaptive label	93.6%, 90.9% - 96.3%	75.3%, 73.4% - 77.2%	64.3%, 62.1% - 66.5%
DenseNet-53 [29]	Normal label	91.5%, 90.4% - 92.6%	72.0%, 70.8% - 73.2%	64.3%, 62.1% - 66.5%
	Adaptive label	91.9%, 90.7% - 93.1%	73.3%, 72.3% - 74.3%	64.3%, 62.6% - 66.0%
EfficientNet-B0 [36]	Normal label	91.9%, 90.0% - 93.8%	70.3%, 68.4% - 72.2%	60.7%, 59.0% - 62.4%
	Adaptive label	92.9%, 91.7% - 94.1%	73.7%, 72.5% - 74.9%	61.7%, 60.3% - 63.1%
VGG16 [37]	Normal label	87.1%, 86.7% - 88.9%	71.0%, 68.3% - 73.7%	64.4%, 62.4% - 66.2%
	Adaptive label	89.5%, 86.1% - 92.9%	71.7%, 67.9% - 75.5%	66.2%, 61.4% - 71.1%
VGG19 [37]	Normal label	89.8%, 88.2% - 91.5%	72.7%, 67.8% - 77.5%	61.6%, 59.0% - 64.3%
	Adaptive label	90.8%, 87.5% - 94.2%	74.7%, 69.3% - 80.0%	63.9%, 59.4% - 68.5%
ResNet-v2-50 [38]	Normal label	89.8%, 88.5% - 91.2%	74.0%, 71.6% - 76.4%	64.6%, 62.4% - 66.7%
	Adaptive label	90.5%, 89.0% - 92.0%	76.0%, 73.5% - 78.5%	65.9%, 63.0% - 68.8%
Inception-v4 [39]	Normal label	89.2%, 86.6% - 91.7%	68.7%, 64.3% - 73.0%	57.7%, 54.9% - 60.5%
	Adaptive label	90.2%, 86.5% - 93.9%	72.7%, 69.0% - 76.3%	62.0%, 60.1% - 63.9%
DcardNet-36	Normal label	93.6%, 91.7% - 95.5%	74.7%, 73.5% - 75.9%	64.3%, 62.6% - 66.0%
	Adaptive label	94.2%, 91.9% - 96.5%	76.7%, 73.4% - 80.0%	64.7%, 61.5% - 67.9%

To further analyze the improvement in generalization by the adaptive label smoothing, we measured the losses and accuracies based on the proposed *DcardNet*-36 and *ResNet*-18 with or without adaptive label smoothing on the 3-class dataset with 20% data exclusively used as testing dataset (Fig. 3.7). The testing losses and accuracies were obtained after each 10 training steps and both smoothed by an average filter with length 50. The training accuracies were smoothed by an average filter with length 100. In Fig. 3.7(A) and 3.7(C), we can see the testing losses with adaptive label smoothing were lower than the losses without adaptive label smoothing during the entire training process. Though the training accuracies with and without adaptive label smoothing were almost the same, the testing accuracies with adaptive label smoothing were always higher than the accuracies without adaptive label smoothing (Fig. 3.7(B) and 3.7(D)). In addition, the testing accuracy with adaptive label smoothing increased more smoothly and monotonically than the accuracy without adaptive label smoothing. By comparing two rows, we can also intuitively see that *DcardNet*-36 has better generalization performance and lower overfitting than the *ResNet*-18. And as noted, the adaptive label smoothing has higher improvement on *ResNet*-18 than *DcardNet*-36.

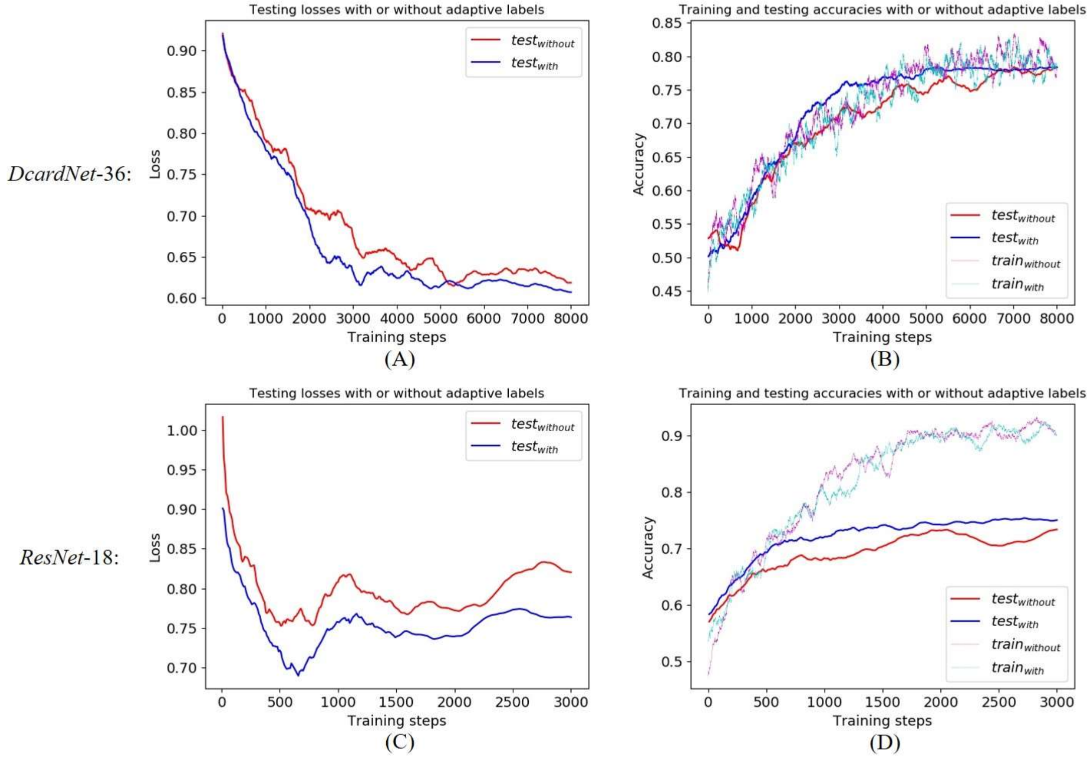


Figure 3.7: Comparisons of the losses and accuracies based on proposed *DcardNet*-36 and *ResNet*-18 with or without adaptive label smoothing on the 3-class dataset with 20% of the data as the testing dataset. (A) Comparisons of the testing losses based on *DcardNet*-36. (B) Comparisons of the training (dotted lines) and testing (solid lines) accuracies based on *DcardNet*-36. (C) Comparisons of the testing losses based on Res*Net*-18. (D) Comparisons of the training (dotted lines) and testing (solid lines) accuracies based on *ResNet*-18.

3.6.5 Comparison with ensemble networks based on enface OCT and OCTA

We also compared the performances on 2-class DR classification between our method and a previously proposed ensemble network [75] which also uses *enface* OCT and OCTA as inputs. The ensemble network consisted of four VGG19 [87] with pre-trained ImageNet parameters. The inputs of the ensemble network were SVC and DCP *enface* images respectively generated from OCT and OCTA. Based on the same implementation details, the results of the ensemble network were shown in Table 3.9. The overall accuracy, sensitivities and specificities of our method are all better than the ensemble network.

Table 3.9: Comparison of the 2-class DR classification performance between our method and the ensemble network

Methods	Accuracy (mean, 95% CI)	Sensitivity (mean, 95% CI)	Specificity (mean, 95% CI)
Ensemble network	86.8%, 85.3% - 88.2%	90.5%, 84.8% - 92.6%	78.9%, 73.1% - 88.4%
Our method	94.2%, 91.9% - 96.5%	96.0%, 94.2% - 97.8%	90.5%, 87.1% - 94.0%

3.7 Discussion

We proposed a new convolutional neural network architecture based on dense and continuous connection with adaptive rate dropout (*DcardNet*) for automated DR classification based on OCT and OCTA data. To our knowledge this is the first study to report DR classification across multiple levels based on OCT and OCTA data. A classification scheme like this is desirable for several reasons. OCT and OCTA are already an extremely common procedures in ophthalmology [90]. An automated DR classification framework could further extend the applications of these technologies. If OCT/OCTA can deliver the same diagnostic value as other modalities, the number of procedures an individual would require for accurate diagnosis would be reduced, which will ultimately lower clinical burden and healthcare costs. Furthermore, OCT/OCTA provide a unique set of features (three-dimensionality combined with high-resolutions) that may prove to have complimentary or superior diagnostic value for some diseases; however, the sheer size of OCT/OCTA data sets inhibits detailed analysis. By providing tools for automation, we can begin to acquire data that can help identify new biomarkers or other features useful for DR staging.

Our network design incorporated several ideas that enabled rapid training and accurate results. We found that, compared to the residual structure, the dense connected structure was much more resistant to overfitting. However, the dense connection also had a lower convergence rate than the residual structure (*ResNet*). In order to increase the convergence rate and keep overfitting low, the dense and continuous connection was proposed and used in this study. In the new architecture, a dense connection was continuously used within a sliding window from the first bottleneck block to the last one. Compared to use of dense connections within each block (*DenseNet*), the new structure was able to deliver useful features with lower losses. In addition, the use of dropout with adaptive rate kept overfitting low. Sixteen bottleneck blocks with 24 output features were finally chosen in this study based on the classification complexity and size of the dataset. For more classes and larger datasets (like those seen in ImageNet), more bottleneck blocks with more output features may be needed.

Adaptive label smoothing was proposed and used to reduce overfitting in this study. The labels of each of the training steps were adaptively smoothed based on their prediction histories. Because of the adaptively smoothed labels, the convergence of the network could be more focused on the mispredicted data, rather than the data that was already correctly predicted. The only concern for this technique is the inaccuracy introduced from data which have an ambiguous ground truth. Therefore, this technique is more suitable to well-labeled datasets. Another technique we used to reduce the overfitting was data augmentation, which has been widely used in medical image classification. In addition to improving data diversity, the data augmentation we used in this study also fits with practical diagnosis, where the doctors' diagnosis is not influenced by the angle of the *en face* vasculature.

For practical and historical reasons, layer segmentation has become a necessary step for most analytic pipelines using OCT and OCTA. The enface images based on segmented layers are not only used to automated DR classification but also necessary for OCT-based routine diagnosis. From a machine learning perspective, this is a mixed blessing. Dimensionality reduction enables swifter training (since 3D data sets are much sparser), but simultaneously suppresses otherwise learnable information. Our network was trained on datasets segmented using manually corrected software [78, 91-94], which introduces both a manual step into our data pipeline and some idiosyncrasy into ground truth. State of the art layer segmentation now requires less manual correction [32, 54, 95, 96], and we believe will continue to do so. However, the accuracy of our results is, unfortunately, probably negatively impacted by these limitations in the ground truth used for training. OCTA networks are also unfortunately limited by a relative paucity of data compared to other medical imaging datasets. As more OCTA data is acquired, training on 3D data volumes may become practicable, mitigating this concern.

The overall accuracies based on OCT-based inputs, OCTA-based inputs, and OCT+OCTA-based were the same of 2-class and 4-class DR classification. However, we still think the OCT+OCTA-based input is a better option. First, this input strategy still improved the overall accuracy of 3-class DR classification and also balanced the sensitivities of 4-class DR classification. Second, some DR or DME related biomarkers such as fluid could be easier detected in OCT. At last, the OCT *enface* generation is not time-consuming after the retinal layers are segmented, and this segmentation is also needed for OCTA *enface* generation. Therefore, the designed OCT+OCTA-based input pattern is still preferable for the DR classification.

The overall accuracy of the 4-class DR classification was much lower than other two classification levels. In addition, the sensitivity of severe NPDR classification was much lower than the other classes. These two issues are caused by the small differences between the two NPDR classes, which are much smaller than the differences between no DR, NPDR and PDR. Another reason for this relatively low performance is that the number of severe NPDR cases was much smaller than other classes. Therefore, the network could hardly identify the differences between two NPDR severities before overfitting sets in. In future work, we will focus on overcoming these problems by using a larger and more balanced dataset and adding some extra manually selected biomarkers to the inputs. In addition, according to the difference between accuracies based on 5-fold and 10-fold cross-validations, using "leave-one-subject-out" experiments could also help increase the final accuracy and sensitivity.

Compared to CFP-based DR classifications [7-10], the overall accuracy of our 2-class DR classification was slightly lower. One reason was that the CFP-based DR classifications had about 100 times as much data as we did. Though we have satisfied accuracies on 2-class and 3-class DR classifications based on our relatively small dataset, a huge dataset like those available from CFP could further improve our DR classification to state-of-art performance. Furthermore, the current classification used for training our algorithm, which is based on grading from color fundus photography, may not be optimal for OCTA classification. The current gold standard for DR diagnosis is based on color fundus photograph which is a considerably different modality from OCT/OCTA. Features used to distinguish some DR classifications using the ETDRS scheme may be missing from OCT/OCTA datasets, which could hurt the accuracy of our algorithm.

Furthermore, there are currently trade-offs between CFP and OCTA. CFP provides a larger field of view, but at lower resolution and the cost of a dimension of information when compared to OCTA. Both provide visualization of a unique set of pathological features. Currently, CFP can provide some information that is inaccessible to OCTA, though complimentary features of the same pathology may be visible to OCTA [97, 98]. However, we do not conceive of this work solely as a means to automatize through OCTA grading what can already also be automatized through CFP. Instead, we believe that this work demonstrates that the feature set that can be extracted through OCTA images of the macular region is sufficient to diagnose DR at a level similar to CFP, without relying on the specific features (microaneurysms, bleeding) provided by CFP. We think this is innovative of its own accord because it adds value to an existing technology.

We note additionally that the amount of data procured from structural OCT in conjunction with OCTA is much larger than that from CFP, by virtue of being high-resolution and three-dimensional. Features like microaneurysms that are currently used to stage DR may not end up being essential to DR staging, as our work shows. Close parity with ETDRS grading of CFP data indicates significant potential for OCTA staging as OCTA hardware continues to improve.

3.8 Conclusion

In conclusion, we proposed a densely and continuously connected convolutional neural network with adaptive rate dropout to perform a DR classification based on OCT and OCTA data. Among our architecture designs, the dense and continuous connections improved the convergence speed and adaptive rate dropout reduced overfitting. Three classification levels were finally performed to fulfill requests from clinical diagnosis. In addition, adaptive label smoothing was proposed and used in this study. With the addition of adaptive label smoothing, the convergence of the network could be more focused on the mispredicted data, rather than the data that was already be correctly predicted. In the end, the trained model focused more on the common features of the whole dataset, which also reduced overfitting. Classifying DR at three levels and generating CAMs could both help clinicians improve diagnosis and treatment.

4 A Diabetic Retinopathy Classification Framework based on Deep-learning Analysis of Angiographic Optical Coherence Tomography

4.1 Abstract

Reliable classification of referable and vision threatening diabetic retinopathy (DR) is essential for diabetic patients to prevent blindness. Optical coherence tomography (OCT) and its angiography (OCTA) have several advantages over fundus photographs. We evaluated a deep-learning-aided DR classification framework using volumetric OCT and OCTA. 456 OCT and OCTA volumes were scanned from eyes of 50 healthy participants and 305 patients with diabetes. Retina specialists labeled the eyes as non-referable (nrDR), referable (rDR) or vision threatening DR (vtDR). Each eye underwent a 3x3-mm scan using a commercial 70-kHz spectral-domain OCT system. We developed a DR classification framework and trained it using volumetric OCT and OCTA to classify eyes into rDR and vtDR. For the scans identified as rDR or vtDR, 3D class activation maps were generated to highlight the subregions which were considered important by the framework for DR classification. For rDR classification, the framework achieved a 0.96 ± 0.01 area under the receiver operating characteristic curve (AUC) and 0.83 ± 0.04 quadratic-weighted kappa. For vtDR classification, the framework achieved a 0.92 ± 0.02 AUC and 0.73 ± 0.04 quadratic-weighted kappa. In addition, the multiple DR classification (non-rDR, rDR but non-vtDR, or vtDR) achieved a 0.83 ± 0.03 quadratic-weighted kappa. A deep learning framework only based on OCT and OCTA can provide specialist-level DR classification using only a single imaging modality. The proposed framework can be used to develop clinically valuable automated DR diagnosis system because of the specialist-level performance showed in this study.

4.2 Introduction

Diabetic retinopathy (DR) is a leading cause of preventable blindness globally [1]. Currently, DR classification uses fundus photographs or clinical examination to identify referable DR (rDR) and vision-threatening DR (vtDR). Eyes with worse than mild nonproliferative DR (NPDR) on the International Diabetic Retinopathy Severity Scale are considered rDR, and eyes with severe NPDR, proliferative DR (PDR), or those with diabetic macular edema (DME) are considered vtDR [3]. An efficient and reliable classification system is essential in identifying patients who can benefit from treatment without an undue burden to the clinic. Eyes with rDR but without vtDR can be observed closely without referral to an ophthalmologist, helping preserve scarce resources for patients that require treatment. To do this safely requires an accurate stratification of patients into these categories [4, 5].

Deep learning has enabled multiple reliable automated systems that classify DR from fundus photographs [7-10]. However, fundus photographs have a low sensitivity (60-73%) and specificity (67-79%) for detecting diabetic macular edema (DME), which accounts for the majority of vision loss in DR [14, 15]. This means that even when a network performs very well against a ground truth generated from fundus photographs, patients with DME may still frequently be misdiagnosed. Supplementing fundus photography with OCT, which is the current gold standard for diagnosing macular edema, can avoid this problem [13, 99-107]. However, reliance on multiple imaging modalities is undesirable as it increases logistic challenges and cost.

Our group and others have demonstrated that OCT angiography (OCTA) can be used to stage DR according to fundus photography-derived DR severity scales using various biomarkers linked to capillary changes in DR [17-23, 108]. Because OCTA can be simultaneously acquired with structural OCT scans used for DME diagnosis, an automated system based on OCTA volume scans can potentially use a single imaging modality to accurately classify DR while avoiding low DME detection sensitivities and associated misdiagnoses that occur in systems based on just fundus photographs.

Despite this advantage, OCTA-based analyses require improvements. Previous methods for classifying DR using OCTA relied on accurate retinal layer segmentation and *en face* visualization of the 3-D volume to visualize or measure biomarkers [33, 72-76]. However, with advanced pathology, retinal layer segmentation can become unreliable. This lowers OCTA yield rate, and may also lead to misclassification through segmentation errors. In addition, quantifying only specific biomarkers fails to make use of the information in the latent feature space of the OCT/OCTA volumes, which may be helpful for DR classification [109].

In this study, we propose an automated convolutional neural network (CNN) [24] that uses the volume-rendered OCT/OCTA to directly classify eyes as either non-rDR (nrDR) or rDR, and as vtDR or non-vtDR (nvtDR). We also include a multiclass classification that classifies eyes as nrDR, rDR/nvtDR, (eyes with referable but not vision-threatening DR) or vtDR. To demonstrate which features the framework relies on to make the classification, the network also generates 3D class activation maps (CAMs) [71]. Visualizations such as these are essential features of direct classification systems, since they allow graders to verify algorithm outputs. To the best of our knowledge, this is the first study to propose an automated multiclass DR severity-level classification framework based directly on OCT and OCTA volumes.

4.3 Methods

4.3.1 Data acquisition

We recruited and examined 50 healthy participants and 305 patients with diabetes at the Casey Eye Institute, Oregon Health & Science University in the United States (50 healthy participants and 234 patients); Shanxi Eye Hospital in China (60 patients); and the Department of Ophthalmology, Aichi Medical University in Japan (11 patients). We included diabetic patients with the full spectrum of disease from no clinically evident retinopathy to proliferative diabetic retinopathy. One or both eyes of each participant underwent 7-field color fundus photography and an OCTA scan using a commercial 70-kHz spectral-domain OCT system (RTVue-XR Avanti, Visionix Inc) with 840-nm central wavelength. The scan depth was 1.6 mm in a 3.0 × 3.0 mm region (640 × 304 × 304 pixels) centered on the fovea. Two repeated B-frames were captured at each line-scan location. The structural images were obtained by averaging the two repeated and registered B-frames. Blood flow was detected using the split-spectrum amplitude-decorrelation angiography algorithm [23, 57]. For each volumetric OCT/OCTA, two continuously acquired volumetric raster scans (one x-fast scan and one y-fast scan) were registered and merged through an orthogonal registration algorithm to reduce motion artifacts [58]. In addition, the projection-resolved OCTA algorithm was applied to all OCTA scans to remove flow projection artifacts in the deeper layers [68, 69]. Scans with a signal strength index (SSI) lower than 50 were excluded. The data characteristics are shown below (Table 4.1). When the classes in our data set weren't balanced class weights were adjusted to prevent performance loss. Based on the data distribution showed in Table 4.1, the class weights for nrDR, r/nvtDR and vtDR were 0.76, 1.87 and 0.87 respectively.

Table 4.1: Data for DR classification

Characteristics	rDR classification		vtDR classification		Multiclass DR classification		
Severity	nrDR	rDR	nvtDR	vtDR	nrDR	r/nvtDR	vtDR
Number of eyes/scans	199	257	280	176	199	81	176
Age, mean (SD), y	48.8 (14.6)	58.4 (12.1)	52.2 (14.7)	57.5 (12.3)	48.8 (14.6)	60.4 (14.7)	57.5 (12.3)
Female, %	50.8%	49.0%	50.0%	49.4%	50.8%	48.2%	49.4%
No DR, %	83.4%	1.6%	59.3%	2.2%	83.4%	0.0%	2.2%
Mild NPDR, %	16.6%	0.0%	11.8%	0.0%	16.6%	0.0%	0.0%
Moderate NPDR, %	0.0%	44.4%	28.9%	18.8%	0.0%	100.0%	18.8%
Severe NPDR, %	0.0%	19.8%	0.0%	29.0%	0.0%	0.0%	29.0%
PDR, %	0.0%	34.2%	0.0%	50.0%	0.0%	0.0%	50.0%
DME, %	0.0%	32.3%	0.0%	47.2%	0.0%	0.0%	47.2%

DR = diabetic retinopathy; rDR = referable DR; vtDR = vision threatening DR; r/nvtDR = referable but not vision threatening DR; NPDR = nonproliferative DR; PDR = proliferative DR; DME = diabetic macular edema

A masked trained retina specialist (TSH) graded 7-field color fundus photographs based on Early Treatment of Diabetic Retinopathy Study (ETDRS) scale [11, 12]. The presence of DME was determined using the central subfield thickness from structural OCT based on the Diabetic Retinopathy Clinical Research Network (DRCR.net) standard [4]. We defined nrDR as ETDRS level better than 35 and without DME (also included healthy eyes); referrable DR as ETDRS level 35 or worse, or any DR with DME; r/nvtDR as ETDRS levels 35-47 without DME; and vtDR as ETDRS level 53 or worse or any stage of DR with DME [3]. The participants were enrolled after an informed consent in accordance with an Institutional Review Board approved protocol. The study complied with the Declaration of Helsinki and the Health Insurance Portability and Accountability Act.

4.3.2 Data inputs

Optical coherence tomography and OCTA generate detailed depth-resolved structural and microvascular information from the fundus (Fig. 4.1). Extracting DR-related features using neural networks can, however, be more challenging and time consuming from 3D volumes such as those produced by OCTA than from 2D sources like fundus photography. To improve the computational and space efficiency of the framework, each volumetric OCT and OCTA were resized to 160 × 224 × 224 voxels and normalized to voxel values between 0 and 1. The input was the combination of each pair of resized-volumes, giving final input dimensions of 160 × 224 × 224 × 2 pixels (Fig. 4.1).

4.3.3 DR classification framework

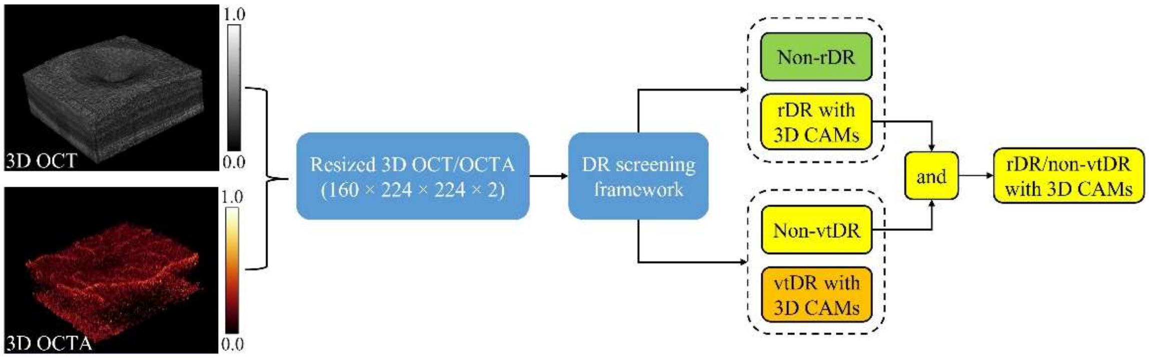


Figure 4.1: Automated DR classification framework using volumetric OCT and OCTA data as inputs. Inputs are first resized to 160 × 224 × 224 × 2 pixels (two channels 3D input with a 160 × 224 × 224 structural and a 160 × 224 × 224 angiographic volume). These inputs are fed into a DR screening framework based on a 3D CNN architecture. The network produces two outputs: a non-referable (nrDR) or referable (rDR) DR classification, and a non-vision-threatening (nvtDR) or vision threatening (vtDR) DR classification. The multiclass DR classification result is defined based on the rDR and vtDR classification results. Class activation maps (CAMs) are also output for each classification result.

A novel 3D CNN architecture (Fig. 4.1) with 16 convolutional layers was designed and used as the core classifier in the DR classification framework (Fig. 4.2). Five convolutional layers with stride 2 were used to downsample the input data. To avoid losing small but important DR-related features, diminishing convolutional kernel sizes were used in the five downsampling layers. We used batch normalization [63] after each 3D convolutional layer to increase convergence speed. In order to improve the computational efficiency while ensuring the resolution of the features, most of the 3D convolutional layers were used with the middle size inputs (after the first downsampling, but before the last). A global average pooling layer was used after the last 3D convolutional layer to generate the 1D input for the output layers.

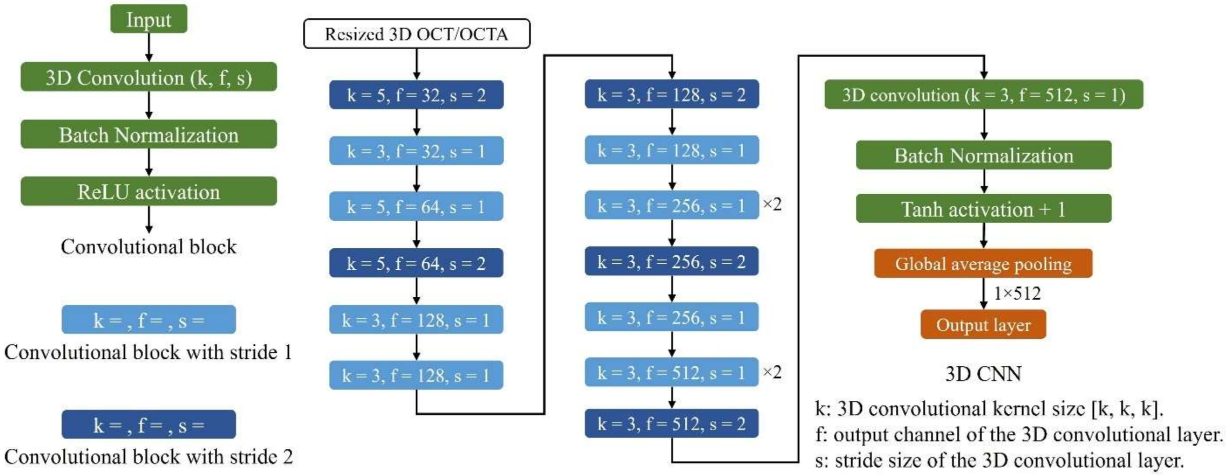


Figure 4.2: Detailed architecture of the novel 3D convolutional neural network (CNN). Sixteen convolutional blocks were used in this 3D CNN. Each convolutional block was constructed as 3D convolutional layer with batch normalization and ReLU activation. Five convolutional blocks with diminishing kernel size (5 to 3) were used to downsample the inputs.

One subtlety in our approach for multiclass classification is the need to correctly identify rDR/nvtDR eyes. Familiar frameworks for image classification like those used to diagnose medical conditions rely on the positive identification features associated with the malady. In our framework, rDR and vtDR classification works similarly by using ReLU activations in the last convolutional layer and weight parameters of all the fully connected layers to guarantee positive-definite prediction values (Fig. 4.3) [81, 82]. However, the identification of r/nvtDR does not depend on just the presence of rDR associated features, but also the *absence* of vtDR-associated features. To solve this issue, two parallel output layers were respectively used to detect rDR and vtDR at the same time (Fig. 4.1). Each output layer was constructed by a fully connected layer with a softmax function (Fig. 4.3). The inputs data can be then classified as nrDR, r/nvtDR, or vtDR based on rDR and vtDR classification outputs.

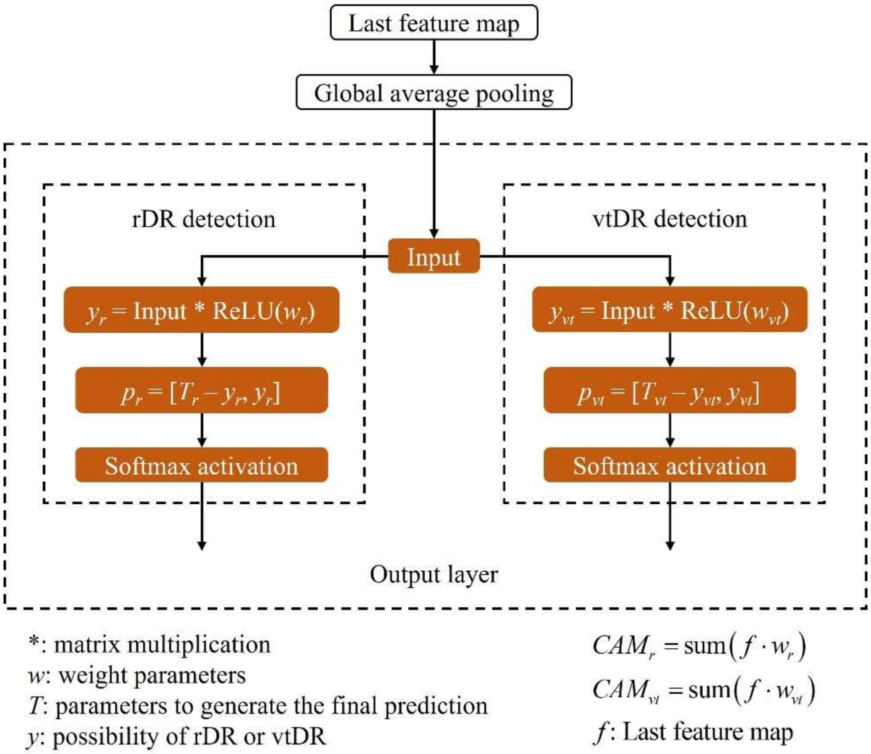

Last feature map
Global average pooling
rDR detection
vtDR detection
Input
y_r = Input * ReLU(w_r)
y_{vt} = Input * ReLU(w_{vt})
$p_r = [T_r - y_r, y_r]$
$p_{vt} = [T_{vt} - y_{vt}, y_{vt}]$
Softmax activation
Softmax activation
Output layer
*: matrix multiplication
w: weight parameters
T: parameters to generate the final prediction
y: possibility of rDR or vtDR
$CAM_r = \text{sum}(f \cdot w_r)$
$CAM_{vt} = \text{sum}(f \cdot w_{vt})$
f: Last feature map

Figure 4.3: Detailed design of the output layer. Two paratactic layers were used to detect referable DR (rDR) and vision threatening DR (vtDR), respectively. The class activation maps (CAMs) for rDR and vtDR were generated according to the weighted sum of the last feature map.

4.3.4 Evaluation and statistical analysis

Overall accuracy, quadratic-weighted Cohen's kappa [110], and area under the receiver operating characteristic curve (AUC) were used to evaluate the DR classification performance of our framework. Among these evaluation metrics, the AUCs were used as the primary metrics for rDR and vtDR classifications. For the multiclass DR classification, the quadratic-weighted kappa was used as the primary metric. Five-fold cross-validation was used in each case to explore robustness. From the whole data set, 60%, 20%, and 20% of the data were split for training, validation, and testing, respectively. Care was taken to ensure data from the same patients were only included in one of either the training, validation, or testing data sets. The parameters and hyperparameters in our framework were trained and optimized only using the training and validation data set. In addition, adaptive label smoothing was used during training to reduce the overfitting [33].

4.3.5 Comparison with a two-dimensional input approach

In contrast to the method proposed in this work, most OCT/OCTA-based DR classification algorithms operate on 2D *en face* images [33, 75, 76, 48]. *En face* projections are popular input choices because (1) they correspond to data the representation most familiar to graders and (2) they typically reduce the size of the input data set relative to the full OCT/OCTA data volume by more than an order of magnitude, which simplifies network training. The trade-off with this data reduction is that networks analyzing *en face* images cannot learn all of the features latent in the full data volume, since many of these features will be removed during projection. Furthermore, *en face* images are vulnerable to segmentation artifacts, which require time consuming review to correct [111]. For these reasons models capable of analyzing OCT/OCTA volumes are desirable, but to be useful such models should reach performance parity with approaches using 2D inputs. To investigate, we compared our model with our previous approach, which was a CNN designed around dense and continuous connection with adaptive rate dropout (*DcardNet*) [33]. This 2D model was trained, validated, and evaluated based on the same data sets of our 3D model.

4.3.6 Three-dimensional class activation maps and evaluation

For the detected rDR and vtDR cases, the 3D CAMs were generated by projecting the weight parameters from corresponding output layer back to the feature maps of the last 3D convolutional layer before global average pooling (Fig. 4.3). To assess whether or not the framework can correctly identify pathological regions, 3D CAMs were overlaid on *en face* or cross-sectional OCT and OCTA images. In order to generate the *en face* projections, an automated algorithm (commercial software provided by Visionix Inc) segmented the following retinal layers (Fig. 3.2): inner limiting membrane (ILM), nerve fiber layer (NFL), ganglion cell layer (GCL), inner plexiform layer (IPL), inner nuclear layer (INL), outer plexiform layer (OPL), outer nuclear layer (ONL), ellipsoid zone (EZ), retinal pigment epithelium (RPE), and Bruch's membrane (BM). For the cases with severe pathologies, trained graders manually corrected the layer segmentation when necessary, using our custom designed COOL-ART grading software [78]. From OCT volumes, we generated the inner retinal (the slab between the Vitreous/ILM and OPL/ONL) thickness map, *en face* mean projection of OCT reflectance, and EZ *en face* mean projection (ONL/EZ to EZ/RPE). From OCTA volumes, we generated the superficial vascular complex (SVC), intermediate capillary plexus (ICP), and deep capillary plexus (DCP) angiograms [18, 41, 70]. The SVC was defined as the inner 80% of the ganglion cell complex (GCC), which included all structures between the ILM and IPL/INL border. The ICP was defined as the outer 20% of the GCC and the inner 50% of the INL. The DCP was defined as the remaining slab internal to the outer boundary of the OPL. The segmentation step and projection maps were just for evaluating the usefulness of 3D CAMs, not as input to the classification framework.

4.4 Results

Model performance was the best for rDR classification, followed by vtDR, then multi-level DR classification (Table 4.2, Fig. 4.4). For the multiclass DR classification, which classifies each case as nrDR, r/nvtDR, or vtDR, we achieved a quadratic-weighted kappa 0.83, which is on par with the performance of ophthalmologists and retinal specialists (0.80 to 0.91) [112]. The network was notably better at classifying rDR and vtDR compared to r/nvtDR (Table 4.2). Most false positive r/nvtDR eyes were classified as vtDR (66.67%) instead of nrDR (33.33%).

Table 4.2: Automated DR classification performances

Metric	rDR classification	vtDR classification	Multiclass DR classification
Overall accuracy	91.52% ± 1.87%	87.39% ± 2.02%	81.52% ± 1.19%
Sensitivity	90.77% ± 4.28%	82.22% ± 2.83%	
Specificity	92.50% ± 3.16%	90.71% ± 3.46%	
AUC (mean ± std)	0.96 ± 0.01	0.92 ± 0.02	
Quadratic-Weighted Kappa	0.83 ± 0.04	0.73 ± 0.04	0.83 ± 0.03

DR = diabetic retinopathy; rDR = referable diabetic retinopathy; vtDR = vision threatening diabetic retinopathy; AUC = area under the receiver operating characteristic curve.

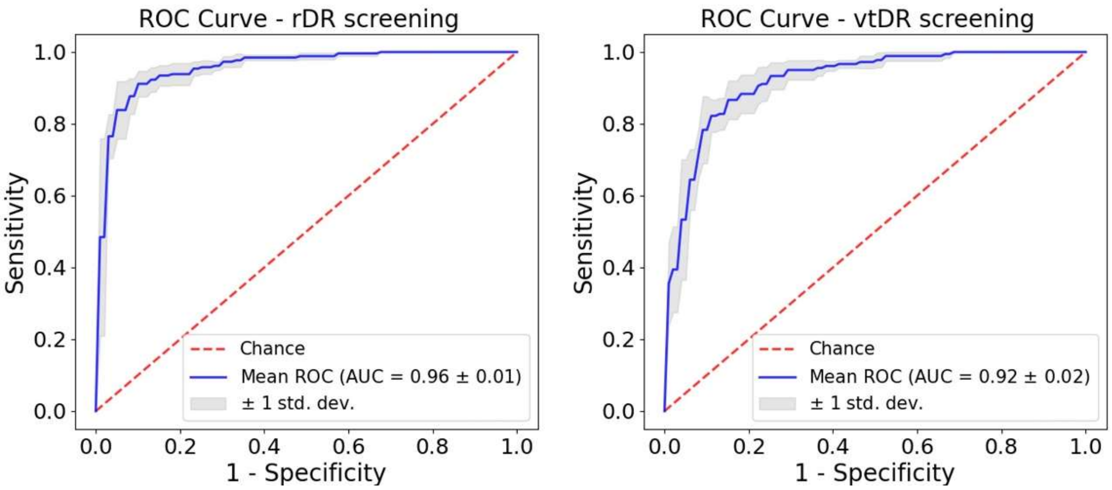
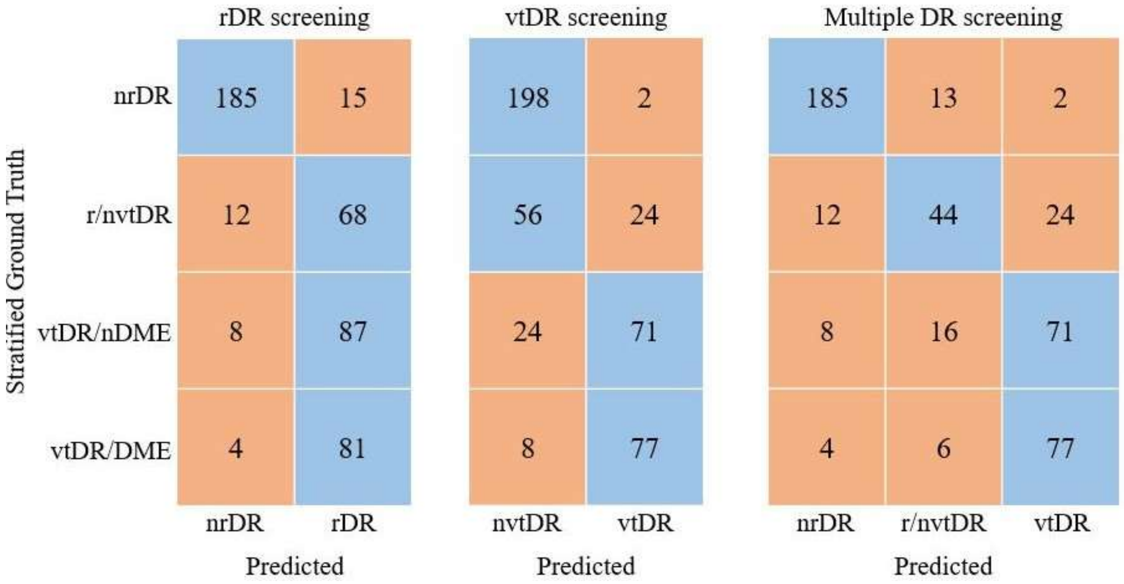


Figure 4.4: The mean receiver operating characteristic (ROC) curve derived from the 5-fold cross-validation for rDR (right) and vtDR (left) classifications based on our DR classification framework. The models achieve an AUC of 0.96 ± 0.01 on rDR classification and AUC of 0.92 ± 0.02 on vtDR classification.

The 3D model performed slightly better for rDR classification, and slightly worse for vtDR classification, than our previous 2D model (Table 4.3). These mixed results indicate that our current model using volumetric data as input was able to train successfully enough to achieve parity with a 2D-image-based approach.

Table 4.3: Comparison between our 3D model and previous 2D model

Models	rDR classification		vtDR classification	
	Overall accuracy	AUC	Overall accuracy	AUC
2D model	89.67% ± 2.50%	0.95 ± 0.02	88.99% ± 0.84%	0.94 ± 0.02
3D model	91.52% ± 1.87%	0.96 ± 0.01	87.39% ± 2.02%	0.92 ± 0.02



rDR screening

Stratified Ground Truth	nrDR	rDR
nrDR	185	15
r/nvtDR	12	68
vtDR/nDME	8	87
vtDR/DME	4	81

Predicted

vtDR screening

Stratified Ground Truth	nvtDR	vtDR
nrDR	198	2
r/nvtDR	56	24
vtDR/nDME	24	71
vtDR/DME	8	77

Predicted

Multiple DR screening

Stratified Ground Truth	nrDR	r/nvtDR	vtDR
nrDR	185	13	2
r/nvtDR	12	44	24
vtDR/nDME	8	16	71
vtDR/DME	4	6	77

Predicted

Figure 4.5: Three confusion matrices for referable DR (rDR) classification, vision threatening DR (vtDR) classification, and multiclass DR classification based on the overall 5-fold cross-validation results. The vtDR was split as non-DME (nDME) and DME in the matrices. The correctly and incorrectly classified cases are shaded blue and orange, respectively.

To demonstrate the deep-learning performance more explicitly, we compared the stratified ground truth with the network prediction with confusion matrices using the overall values from 5-fold cross-validation (Fig. 4.5). In the three confusion matrices, the vtDR cases were separated into non-DME (nDME) and DME to investigate whether the presence of DME can affect rDR and vtDR classification accuracy. In the rDR classification task, we found the classification accuracies of vtDR/nDME and vtDR/DME to be similar (87/95 and 81/85). For vtDR classification, the network identified cases with DME (77/85) with a greater accuracy than nDME cases (71/95), which may imply DME features were likely influential for decision making. In the multi-level classification, the network misclassified 16/95 vtDR/nDME cases as r/nvtDR. In addition, most of the r/nvtDR cases with false-positive results were classified as vtDR. Only 2 nrDR cases were misidentified as vtDR.

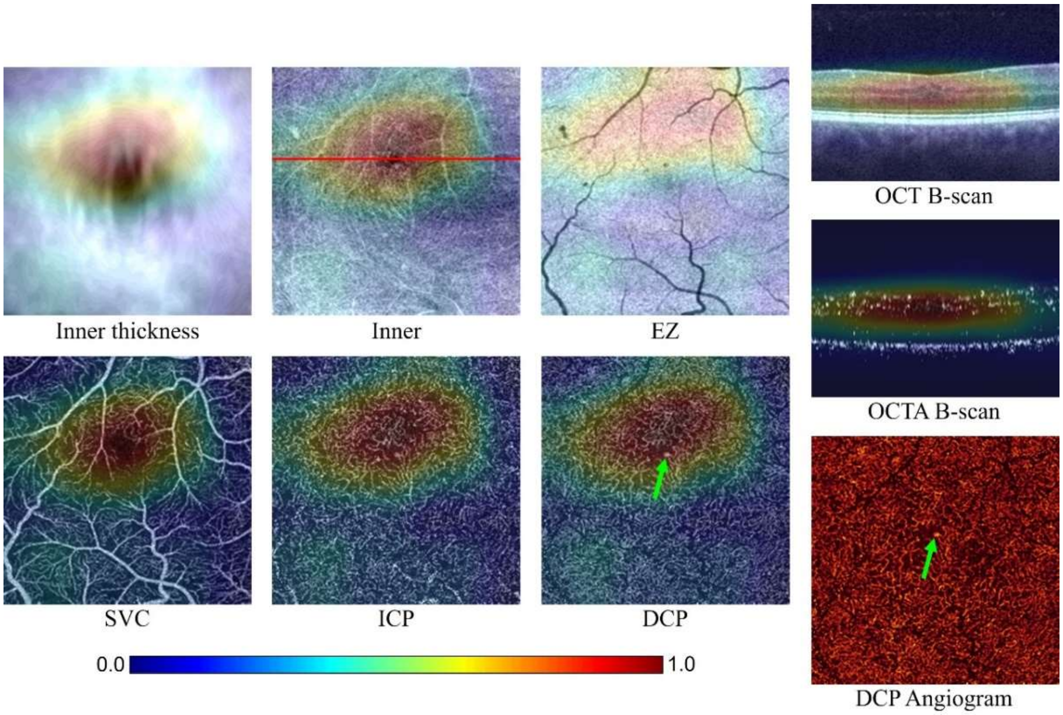


Figure 4.6: Class activation maps (CAMs) based on the referable DR (rDR) output layer of our framework for data from an eye with rDR without vision threatening DR (vtDR). Six *en face* projections covered with the corresponding projections of the 3D CAMs are shown. Extracted CAMs for an OCT and OCTA B-scans (red line in the inner retina *en face* projection) are also shown. The deep capillary plexus (DCP) angiogram without a CAM is shown so that the pathology highlighted by the corresponding CAM can be more easily identified. The green arrows indicate an abnormal vessel in the DCP. For descriptions of the regions projected over to produce the *en face* images, see the caption to Fig. 3.2.

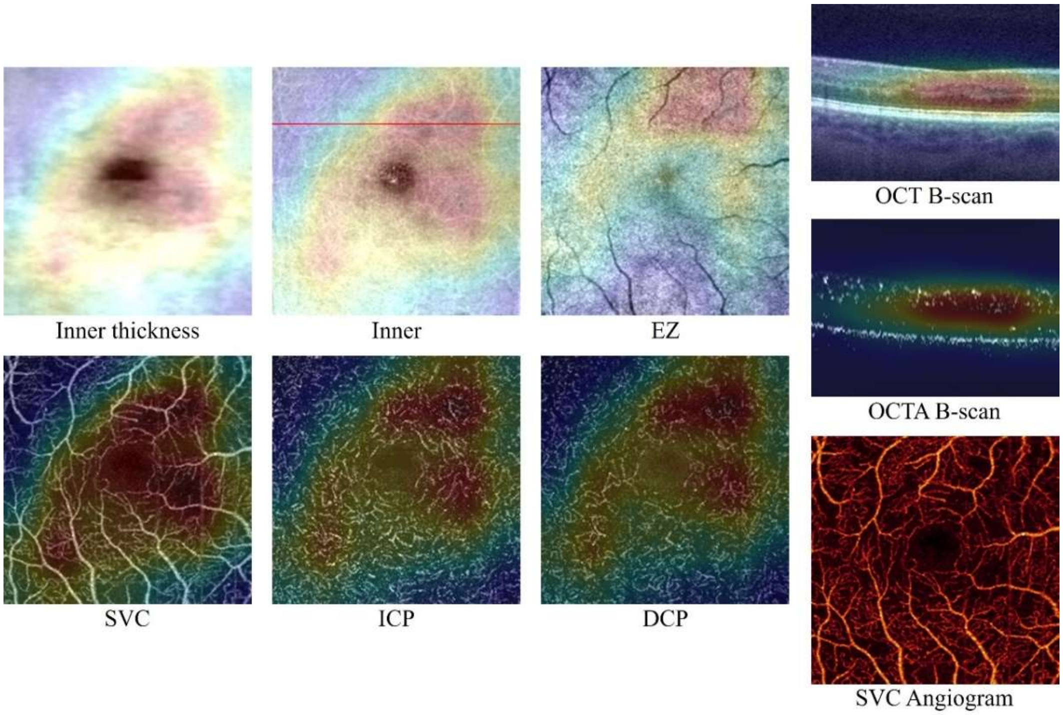


Figure 4.7: Class activation maps (CAMs) based on the vision threatening DR (vtDR) output layer of our framework for data from an eye with vtDR but without DME. Six *en face* projections covered with corresponding projections of the 3D CAMs are shown. Extracted CAMs for an OCT and OCTA B-scan (red line in the inner retina *en face* projection) are also shown. A SVC angiogram without a CAM is also shown to help identify pathological features for comparison. The SVC CAM indicates that the framework learned to identify non-perfusion areas, which are known biomarkers for DR diagnosis. For descriptions of the regions projected over to produce the *en face* images, see the caption to Fig. 3.2.

To better understand network decision making we produced CAMs for some example cases. The CAM output of a r/nvtDR case points to dilated vessels in the DCP and a perifoveal area of decreased vessel density (Fig. 4.6). Meanwhile, in a vtDR case without DME, the CAMs have a larger area of high attention (Fig. 4.7), indicating that the DR pathology is more pervasive throughout the volume. In addition to pointing to areas of decreased vessel density, the CAM overlaid on a structural OCT B-scan points to an area with abnormal curvature of the retinal layers. Finally, for a vtDR case with DME, the CAM pointed to areas with intraretinal cysts and abnormal curvature of the retinal layers on structural OCT, as well as decreased vessel density and abnormally dilated vessels on OCTA (Fig. 4.8). This is clearly an improvement over our previous 2D CAM output (Fig. 4.9) [33], which identified changes in the perifoveal region, but missed other pathologies, such as intraretinal cysts and abnormally dilated vessels. Based on the distribution of the highlighted regions from all the 3D CAMs, we found the non-perfusion areas near fovea and most fluids were preferentially selected by our framework for decision making. In addition, the non-perfusion areas at the boundary of the inputs were barely selected by our framework.

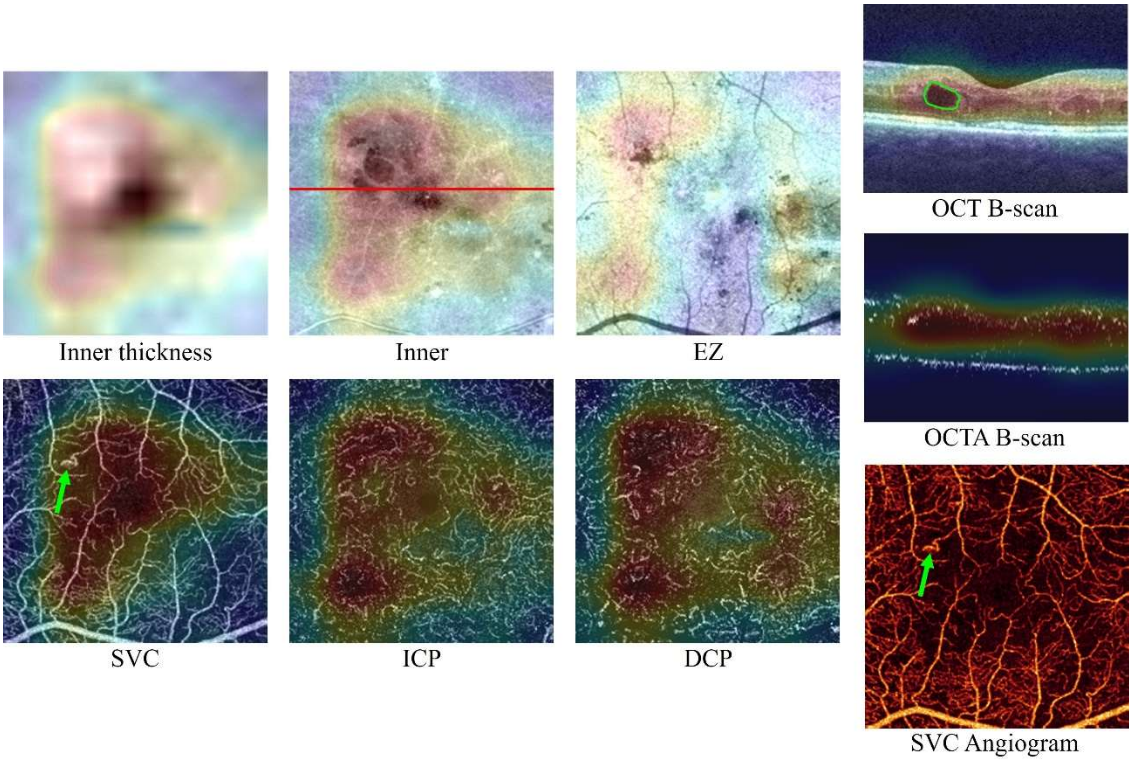
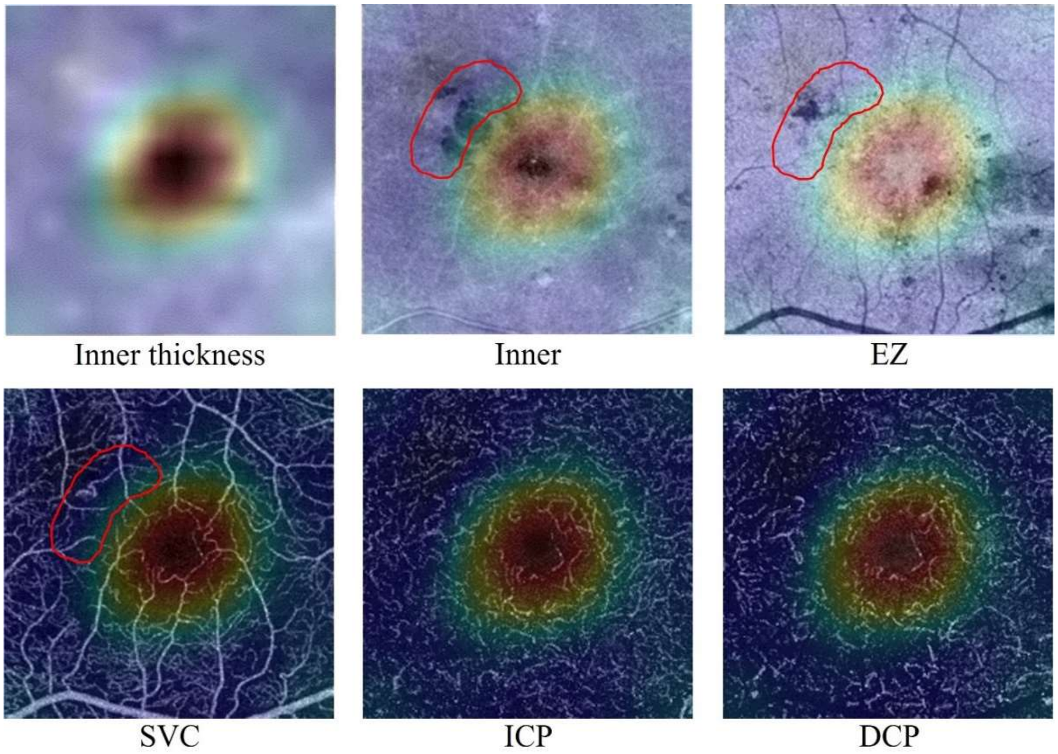


Figure 4.8: Class activation maps (CAMs) based on vision threatening DR (vtDR) output layer of our framework for data from an eye with vtDR and DME. Six *en face* projections covered with the corresponding projections of 3D CAMs are shown. Extracted CAMs for an OCT and OCTA B-scan (red line in the inner retina *en face* projection) are also shown. The SVC angiogram without a CAM is shown to more readily observe pathology. The green arrow in the SVC CAM shows an abnormal vessel, which can also be seen in the angiogram. Central macular fluid is marked by green circle on the OCT B-scan. The CAM allocated high weights to both of these regions. For descriptions of the regions projected over to produce the *en face* images, see the caption to Fig. 3.2.



Inner thickness

Inner

EZ

SVC

ICP

DCP

Figure 4.9: Two-dimensional class activation maps (CAMs) generated by our previous study for data from an eye with vtDR and DME. Six *en face* projections (see Fig. 3.2 for details) covered with the same 2D CAMs are shown. The abnormal vessels and central macular fluid, which were highlighted regions in the 3D CAMs in Fig. 4.8, were not weighted highly by the 2D CAM algorithm (red circles in the inner and EZ CAMs).

4.5 Discussion

In this study, we proposed a CNN-based automated DR classification framework that operates directly on volumetric OCT/OCTA data without requiring retinal layer segmentation. This framework classified cases into clinically actionable categories (nrDR, r/nvtDR, and vtDR) using a single imaging modality. For multiclass DR classification, the framework achieved a quadratic-weighted kappa of 0.83 ± 0.03, which is on par with the performance of human ophthalmologists and retinal specialists (0.80 to 0.91) [112]. The network also demonstrated robust performance on both rDR and vtDR classification (AUC = 0.96 ± 0.01; 0.92 ± 0.02, respectively).

The framework used feature-rich structural OCT and OCTA volumes as inputs and a deep-learning model as the core classifier to achieve a high level of performance. The majority of DR classification algorithms to date have been based on fundus photographs [7-10]. However, fundus photographs detect DME with only about a 70% accuracy relative to structural OCT, while DME accounts for the majority of vision loss in DR [14, 15]. Our method, on the other hand, actually performs better in the presence of DME (Fig. 4.5).

Our image labels appealed to structural OCT to detect DME, and so did not adhere exactly to the ETDRS scale (the current gold standard for DR grading), which uses only seven field fundus photographs. This prevented our model from learning to misdiagnose eyes based on the presence of DME not detected by fundus photography. However, at the same time OCTA may not recapitulate every feature in fundus photography used for staging DR on the ETDRS scale. For example, OCTA does not detect intraretinal hemorrhages and may not detect all microaneurysms [23]. Achieving comparable performance to fundus photographs-based automated classification frameworks indicates that these disadvantages were surmounted by our approach.

Another important feature in our framework design is the use of a deep-learning model for the classifier. Compared to previously published OCT/OCTA-based DR classification algorithms, the proposed framework has several innovations. One advantage is the use of the volume-rendered OCT/OCTA, instead of pre-selected features from segmented *en face* images. This means that correlations or structures within the data volume that may be difficult for a human to identify can still be incorporated into the decision making in our framework. Two dimensional approaches may miss important features without access to cross-sectional information, as happens with color fundus photography and DME [76]. As a corollary, our framework may then also have a greater capacity to improve with more training data since no data is removed by projection. Moreover, accurate retinal layer segmentation is required to generate the *en face* images. In severely diseased eyes, automated layer segmentations often fail. Mis-segmented layers can introduce artifacts into *en face* images unless they are manually corrected, a labor-intensive task that may not be clinically practical. By using volumetric data, our framework avoids this issue entirely. Another advantage built into our framework is the ability to detect both rDR and vtDR. This higher level of granularity makes a more efficient use of resources possible compared to solutions that only identify rDR [7, 9, 72-76].

A final significant advantage in our framework is the inclusion of 3D CAMs. While independent of model performance, generating CAMs allow clinicians to interpret the classification results and ensure model outputs are correct. This is important since, outside of visualizations such as CAMs, users cannot in general ascertain how deep learning algorithms arrive at a classification decision. However, in medical imaging it is essential to be able to verify and understand these classification decisions since doing so could prevent misdiagnosis. Black-box algorithms such as deep learning algorithms may hide important biases that could prove to be disadvantageous for certain groups. This risk can be lowered when the results are interpretable. With our framework this is possible. The CAMs in this work were generated volumetrically. Compared to 2D CAMs, the current framework using 3D OCT/OCTA as inputs can identify and learn relevant features (Fig. 4.8 and 4.9). The resulting CAMs consistently highlighted macular fluid (Fig. 4.8), demonstrating that the model did indeed learn relevant features since central macular fluid is the most important biomarker for detecting DME [109]. We also found our 3D CAMs pointed to other key features such as lower vessel density and dilated capillaries (Fig. 4.6 and 4.7). Although the 3D CAM did not identify all DR features (e.g. certain regions with lower vessel density were ignored), it found many key features, indicating that our framework has successfully learned relevant features and that 3D CAMs could be useful in clinical review. In addition, the purpose of generating 3D CAMs is not necessarily to find all DR biomarkers, but simply to highlight the features used by the network to make decisions. That the network ignored some known DR-associated features is interesting, since it implies that these features were not critical for diagnosing DR at a given severity.

There are aspects of our framework that could be improved in future work. The sensitivity for r/nvtDR classification (55.00% ± 15.51%) was lower than the other two grades (92.50% ± 3.16% for nrDR and 81.11% ± 2.08% for vtDR). Larger data sets with more r/nvtDR cases could help mitigate this performance gap, and it is worth noting that most r/nvtDR misclassifications resulted in vtDR classifications. While this is obviously not optimal, this outcome at least spares patients with referable DR from failing to receive needed clinical attention. In addition, the r/nvtDR is a middle DR severity which also makes the classification more challenge than other two grades. The classification performance for rDR (AUC = 0.96 ± 0.01) also outperforms vtDR (AUC = 0.92 ± 0.02). Our model relied on a small scan region (3.0 × 3.0 mm) at the central macula [17, 19, 113]. However, a larger scan area with appropriate sampling density (e.g. not lower than 10 um/pixel) could still improve the DR classification performance, as there are key DR features such as neovascularization and venous beading that are typically outside the 3x3mm region. Since these features are associated with more advanced stages of DR, exploring models that use larger fields of view may preferentially improve vtDR diagnosis. Therefore, in the future, we hope to improve DR classification performance with larger datasets and scans with larger field of view. In addition, to improve the reliability of our evaluation results, we also hope to test our framework on external data set based on federated learning.

4.6 Conclusion

We proposed a fully automated DR classification framework using 3D OCT and OCTA as inputs. Our framework achieved reliable performance on multiclass DR classification (nrDR, rDR/nvtDR, and vtDR), and produces 3D CAMs that can be used to interpret the model's decision making. By using our framework, the number of imaging modalities required for DR classification was reduced from fundus photographs and OCT to an OCTA procedure alone. This accuracy of the model output in this study also suggests the combination of OCT/OCTA and deep learning could perform well in a clinical setting.

5 Deep-Learning-Aided Diagnosis of Diabetic Retinopathy, Age-Related Macular Degeneration, and Glaucoma based on Structural and Angiographic Optical Coherence Tomography

5.1 Abstract

Timely diagnosis of eye disease is paramount to obtaining the best treatment outcomes. Optical coherence tomography (OCT) and its angiography (OCTA) have several advantages that lend themselves to the early detection of ocular pathology, including that the techniques produce large, feature-rich data volumes. However, the full clinical potential of both OCT and OCTA is stymied when the complex data they acquire must be manually processed. Here we propose an automated diagnostic framework based on structural OCT and OCTA data volumes that could substantially support the clinical application of these technologies. Five hundred and twenty-six OCT and OCTA volumes were scanned from the eyes of 91 healthy participants, 161 patients with diabetic retinopathy (DR), 95 patients with age-related macular degeneration (AMD), and 108 patients with glaucoma. The diagnosis framework was constructed based on semi-sequential 3D convolutional neural networks. The trained framework classifies a combined structural OCT and OCTA scan as normal, DR, AMD, or glaucoma. Five-fold cross-validation was performed, with 60% of the data reserved for training, 20% for validation, and 20% for testing. Training, validation, and testing data sets were independent, with no shared patients. For the scans diagnosed as DR, AMD, or glaucoma, 3D class activation maps were generated to highlight the subregions which were considered important by the framework for the automated diagnosis. Area under the curve (AUC) of the receiver operating characteristic curve and quadratic-weighted kappa were used to quantify the diagnostic performance of the framework. For the diagnosis of DR, the framework achieved a 0.95 ± 0.01 AUC. For the diagnosis of AMD, the framework achieved a 0.98 ± 0.01 AUC. For the diagnosis of glaucoma, the framework achieved a 0.91 ± 0.02 AUC. A deep learning framework can provide reliable, sensitive, interpretable, and fully automated eye disease diagnosis.

5.2 Introduction

Diabetic retinopathy (DR), age-related macular degeneration (AMD), and glaucoma each represent a leading cause of blindness [114-118]. While the pathophysiologic processes behind vision loss in each of these diseases are unique, they share qualities that make early diagnosis essential. Each is usually asymptomatic during early development [115, 118, 119]. In the case of DR and glaucoma, treatment during the early stages is effective for slowing disease progression and preventing otherwise incurable vision loss [3, 120]. In AMD, conversion to the exudative form of the disease can also lead to rapid, catastrophic vision loss; diagnosis of wet AMD is consequently a major treatment indicator [119]. For DR, AMD, and glaucoma, then, effective screening and early diagnosis are key to preventing poor visual outcomes. However, current diagnostic protocols face important challenges. Among these is a reliance on qualitative traits that may instill subjectivity into diagnoses. Also problematic are protocols that recommend multiple imaging modalities (for example, fundus photography supplemented with optical coherence tomography (OCT) to confirm the presence of edema or exudation) [3], which increases screening cost, requires more training for instrument technicians and can encourage patient non-compliance with clinician recommendations [108].

These issues can be alleviated through the use of OCT and OCT angiography (OCTA), which together provide depth-resolved (3D), micrometer-scale-resolution structural and vascular images of the retina [13, 21-23]. Numerous studies from multiple investigators have confirmed the ability of combined OCT and OCTA imaging to diagnose and detect pathology related to DR, AMD, and glaucoma using quantitative measurements [20, 121-127]. Additionally, combined structural OCT and OCTA have several advantages as a screening technology. Since 2014 OCT has been the most common procedure in ophthalmic practice and is cost-effective relative to allied modalities such as color fundus photography or dye-based angiography [128]. Furthermore, since OCTA can be acquired from structural OCT through alternative data processing, only one procedure is required to obtain both structural and vascular information [129]. And finally, since procedures are non-invasive, combined OCT and OCTA imaging can be performed at will.

Despite these advantages for diagnosing DR, AMD, and glaucoma, a diagnostic platform based on combined structural OCT and OCTA imaging will still require innovation before it can be translated into the clinic. Combined structural OCT and OCTA datasets are large, and manual review of these datasets can be prohibitively time-consuming. Manual review is nonetheless often required, particularly in analytic frameworks that rely on *en face* images since retinal slab segmentation errors (which are common in more pathologic retinas) can introduce artifacts [111]. In underserved areas, the clinical infrastructure to meet these image analysis demands may not be available [130]. To resolve these issues automated image analysis is required. Deep learning is a data-driven technique that is currently the most powerful tool for medical image classification tasks [24]. Previously, diagnostic deep learning networks have been proposed for DR [33, 34, 72-76], AMD [131, 132], and glaucoma [133-136]. However, none of these methods can be used for the automated diagnosis of all three of these diseases simultaneously, which means each must be applied sequentially. This has the net effect of undermining generality and will require technicians to be familiar with several algorithms. Here, we instead present a deep-learning-based platform using combined structural OCT and OCTA data volumes as inputs capable of simultaneously diagnosing DR, AMD, and glaucoma. By relying on data volumes this platform avoids mis-segmentation artifacts in *en face* images (which are difficult to correct). Providing a unified diagnostic framework also ensures that each of these important diseases will be screened for, and also saves computational resources by checking for each disease type simultaneously. In addition, the network outputs 3D class activation maps (CAMs) to highlight the disease-related biomarkers which are helpful for treatment decisions and management, as well as for verifying the algorithm's predictions.

5.3 Methods

5.3.1 Data acquisition

In this study, 102 eyes of 91 healthy participants, 161 eyes of 161 DR patients, 142 eyes of 95 AMD patients and 121 eyes of 108 glaucoma patients were examined at the Casey Eye Institute, Oregon Health & Science University, USA. Each patient had one or both eyes scanned; the entire data set used in this study included 526 volumetric scans. For each eye, the macular region was scanned using a commercial 70-kHz spectral-domain OCT system (Avanti RTVue-XR, Optovue Inc) with an 840-nm central wavelength. The scan depth was 1.6 mm in a 6.0 × 6.0 mm region (640 × 400 × 400 pixels) centered on the fovea. Blood flow was detected using the split-spectrum amplitude-decorrelation angiography algorithm based on the speckle variation between two repeated B-frames [23]. The OCT structural images were obtained by averaging two repeated B-frames. For each data set, two volumetric raster scans (one x-fast scan and one y-fast scan) were registered and merged through an orthogonal registration algorithm to reduce motion artifacts [58]. In addition, the projection-resolved OCTA algorithm was applied to all OCTA scans to remove flow projection artifacts in the deeper plexuses [68, 69]. According to manufacturer recommendations and our experience, OCT/OCTA scans with a signal strength index lower than 50 are generally low quality and were excluded.

A masked trained retina specialist (TSH) graded 7-field color fundus photographs based on the Early Treatment of Diabetic Retinopathy Study (ETDRS) scale [11, 12] to generate the positive ground truth labels for the DR data volumes. Diabetic macular edema (DME) was identified using the central subfield thickness from structural OCT based on the Diabetic Retinopathy Clinical Research Network standard [4]. The eyes with an ETDRS score of 14 or worse or any stage with DME were graded as DR cases. Another masked trained retina specialist (STB) generated the positive AMD ground truth labels by grading 7-field color fundus photographs based on the Age-Related Eye Disease

Study (AREDS) scale [137]. The eyes with AREDS simplified score of 1 or worse were graded as AMD cases. Glaucomatous eyes were determined through clinical diagnosis, and the inclusion criteria for this study were an optic disc rim defect (thinning or notching) or nerve fiber layer defect visible on slit-lamp biomicroscopy (DH). Participants were enrolled after informed consent in accordance with an Institutional Review Board approved protocol, and this study was conducted in compliance with the Declaration of Helsinki and Health Insurance Portability and Accountability Act.

5.3.2 Data inputs

While 3D OCT and OCTA scans can provide much more detailed information than 2D data projections, it is also much more challenging to train a network to extract the relevant information from data volumes than images. This difficulty is compounded in our work by the need to extract relevant features for three different diseases. In order to improve the computational and space efficiency of the framework, each volumetric OCT and OCTA is resized to 160 × 224 × 224 voxels and normalized to voxel values between 0 and 1. Combining the structural OCT and OCTA volumes, the final input dimensions were 160 × 224 × 224 × 2 pixels (Fig. 5.1).

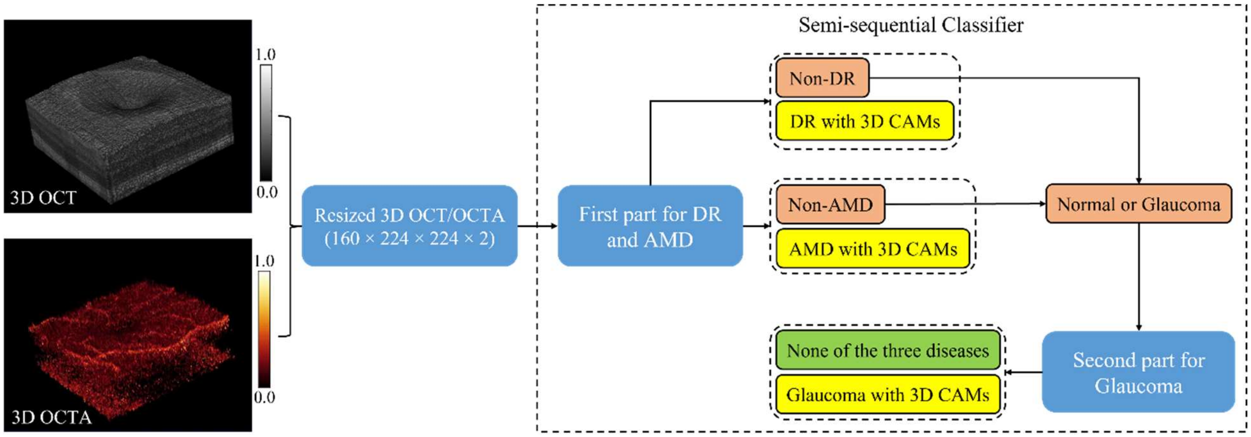


Figure 5.1: Automated combined DR, AMD, and glaucoma diagnostic framework using volumetric OCT and OCTA data as inputs. Structural OCT and OCTA data volumes are resampled and combined to form the input to a semi-sequential classifier. The first part of the classifier then diagnoses DR and AMD. Data not diagnosed as DR and AMD by the first part is fed to the second part for glaucoma. Eyes not diagnosed with DR, AMD or glaucoma can be considered normal or other diseases. For any disease diagnosis, the network also outputs 3D CAMs.

5.3.3 Diagnostic framework for the three eye diseases

The proposed automated DR, AMD, and glaucoma diagnostic framework use a semi-sequential classifier which includes two parts (Fig. 5.1). The first part is a classifier used to diagnose DR and AMD in parallel. This part was trained based on the whole data set with a ground truth label of three classes (DR, AMD, and neither). The second part diagnoses glaucoma from data that was not diagnosed as DR or AMD by the first part, which means the glaucoma was sequentially diagnosed after the DR and AMD diagnosis. Therefore, the combination of these two parts was named as semi-sequential classifier since it contained both parallel and sequential diagnoses. The reason for using a semi-sequential structure to diagnose glaucoma is that the difference between normal and glaucoma in macular OCT/OCTA is much smaller than the difference between normal and DR or AMD. Glaucoma cannot be accurately detected if only one part is used for the diagnosis of DR, AMD, and glaucoma at the same time (see results). To ensure the second part only focused on the difference between normal and glaucoma, it was trained only based on the normal and glaucoma data with two-class labels. Therefore, the two parts were trained separately. Data not diagnosed as DR, AMD, or glaucoma could be considered as normal during our training framework, which relied on healthy eyes being distinguished from these three diseases. However, we note that other eye diseases could still be present in a clinical context. The classifier of each part uses a customized 3D convolutional neural network architecture with 16 convolutional layers (Fig. 5.2). For the first part, two output layers were designed for DR and AMD diagnosis, respectively. For the second part, only one output layer was used to classify each input as normal or glaucoma. Each output layer is a fully-connected layer with a softmax function. For the scans diagnosed as DR, AMD, or glaucoma by the full semi-sequential classifier, 3D CAMs are generated by projecting the weight parameters from the corresponding output layer back to the feature maps of the last convolutional layer before global average pooling.

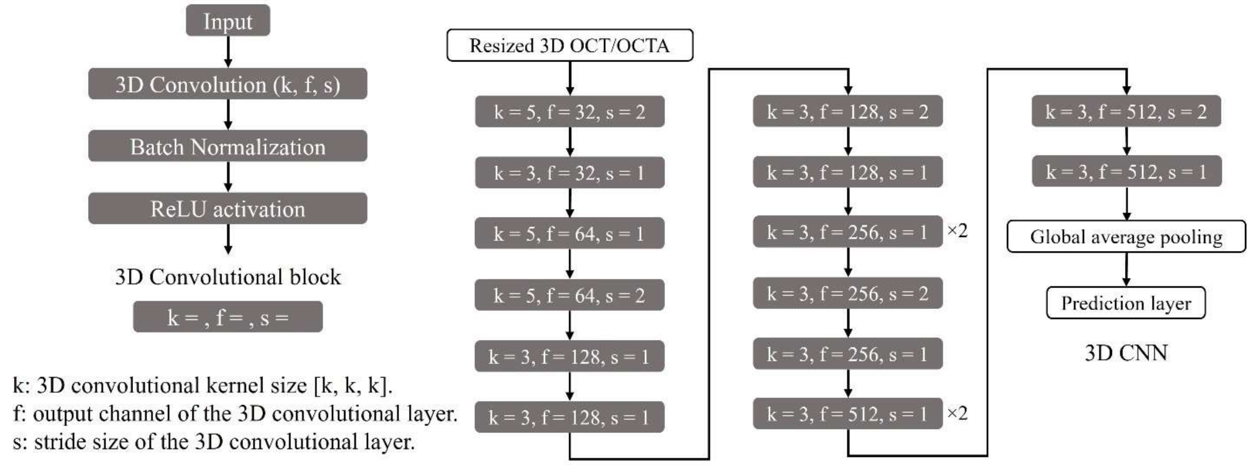


Figure 5.2: The detailed architecture of the three-dimensional (3D) convolutional neural network (CNN) used in this study. Each part of the classifier was established based on this architecture.

5.3.4 Evaluation and statistical analysis

The area under the curve (AUC) for receiver operating characteristic (ROC) and precision-recall curves were used as the primary evaluation metrics to quantify diagnostic accuracy for each disease. Quadratic-weighted Cohen's kappa [110] was used as the metric to evaluate multiple disease diagnostic performance. In addition, the overall accuracy, sensitivity, and specificity were also calculated. Five-fold cross-validation with a 60/20/20 training/validation/testing data distribution was used to assess performance reliability. Data from a single participant was included in only one of either the training, validation or testing data sets. The parameters and hyperparameters in our framework were trained and optimized only using the training and validation data set. The test data set was used exclusively for evaluation to guarantee performance was not biased. In addition, adaptive label smoothing was used during training to reduce overfitting [33].

To evaluate the performance improvement brought by the semi-sequential structure, a parallel classifier with three output layers was constructed to classify each input as normal, DR, AMD, or glaucoma. The parallel classifier was trained, validated, and evaluated based on the same data set as the semi-sequential classifier. But unlike the semi-sequential classifier, glaucoma would be parallelly classified with DR and AMD by the parallel classifier.

5.4 Results

The framework achieved reliable performance as indicated by the AUCs of ROC curves on the test data set, which exceeded 0.9 for each disease in this study (Table 5.1, Fig. 5.3). For the precision-recall curves, both DR and AMD diagnoses achieved high AUCs (above 0.9). Though a separate part in the semi-sequential classifier was used to diagnose glaucoma, the AUCs of both ROC and precision-recall curves for glaucoma diagnosis were still lower than the other two eye diseases (Fig. 5.3). The overall accuracy of the multiple eye disease diagnosis (normal, DR, AMD, and glaucoma) was about 80%.

Table 5.1: Automated disease diagnosis performance

Metric	DR diagnosis	AMD diagnosis	Glaucoma diagnosis	Eye diseases diagnosis
Overall accuracy	90.19% ± 2.03%	94.53% ± 0.71%	89.25% ± 1.75%	79.43% ± 2.01%
Sensitivity	90.00% ± 2.34%	88.28% ± 5.60%	71.67% ± 4.08%	
Specificity	90.27% ± 1.99%	96.88% ± 1.76%	94.39% ± 1.98%	
AUC of ROC	0.95 ± 0.01	0.98 ± 0.01	0.91 ± 0.02	
AUC of precision-recall	0.91	0.95	0.71	
Quadratic-Weighted Kappa	0.78 ± 0.05	0.86 ± 0.02	0.68 ± 0.05	0.57 ± 0.05

AMD = age-related macular degeneration; AUC = area under the curve; DR = diabetic retinopathy; ROC = receiver operating characteristic.

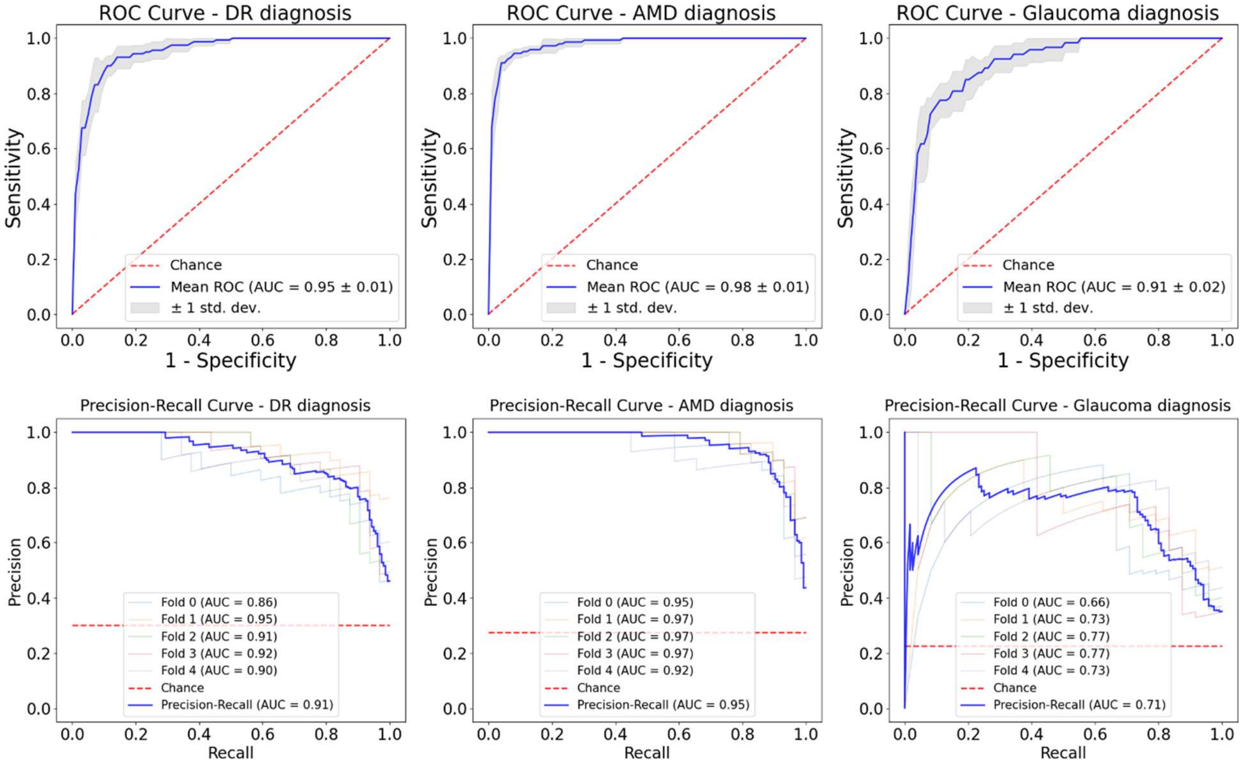


Figure 5.3: Receiver operating characteristic (ROC) and precision-recall curves derived from 5-fold cross-validation for the diagnosis of DR, AMD, and glaucoma based on the full framework. The area under the curve (AUC) was calculate for both curves. The models achieve AUCs of 0.95 ± 0.01 and 0.91 for ROC and precision-recall on DR diagnosis, AUCs of 0.98 ± 0.01 and 0.95 for ROC and precision-recall on AMD diagnosis, and AUCs of 0.91 ± 0.02 and 0.71 for ROC and precision-recall on glaucoma diagnosis. In addition, the glaucoma precision-recall curve look different from the other two curves since the glaucoma prediction was the combination of two parts in the semi-sequential classifier.

We also constructed two confusion matrices (for the first part of the semi-sequential classifier and the full semi-sequential classifier) using the overall results from a 5-fold cross-validation (Fig. 5.4). In the first part that only diagnosis DR and AMD, most misdiagnoses were between normal/glaucoma and DR. In the full semi-sequential classifier (which also includes glaucoma and normal diagnoses), normal eyes were most often misdiagnosed, and when diseased eyes were misdiagnosed, it was most often as normal eyes.

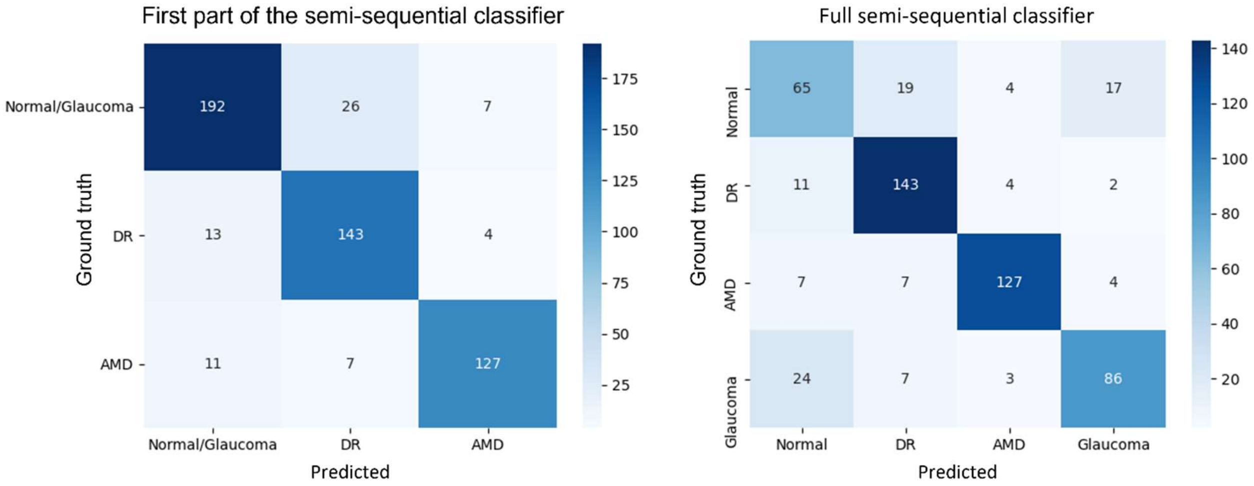


Figure 5.4: Confusion matrices for the first part of the semi-sequential classifier (left) and the full semi-sequential classifier (right) based on the overall results of 5-fold cross-validation.

To quantify the performance improvement brought by the semi-sequential structure, a comparison between the glaucoma classification performances of semi-sequential and parallel classifiers was performed (Table 5.2). The comparison was performed only based on the normal and glaucoma testing data. With the semi-sequential classifier, the sensitivity, specificity, and AUC of the ROC curve were respectively improved by 12.00%, 6%, and 0.1. This improvement was because the semi-sequential diagnosis makes more sense than the parallel diagnosis of multiple diseases in this context, given the fact that the difference between normal and glaucoma data is much smaller than the differences between normal and AMD or DR data. In the training, the parallel classifier mostly focused on the learning of unique features of DR and AMD and ignored the glaucoma features. The improvement brought by the semi-sequential structure was critical for the glaucoma diagnosis performance of the proposed diagnosis framework.

Table 5.2: Comparison between the glaucoma classification performances of the semi-sequential and parallel classifiers

	Overall accuracy	Sensitivity	Specificity	AUC of ROC
Semi-sequential classifier	77.33% ± 3.82%	78.33% ± 5.53%	76.19% ± 6.73%	0.78 ± 0.03
Parallel classifier	63.11% ± 4.35%	56.67% ± 6.24%	70.48% ± 3.56%	0.68 ± 0.03

AUC = area under the curve; ROC = receiver operating characteristic.

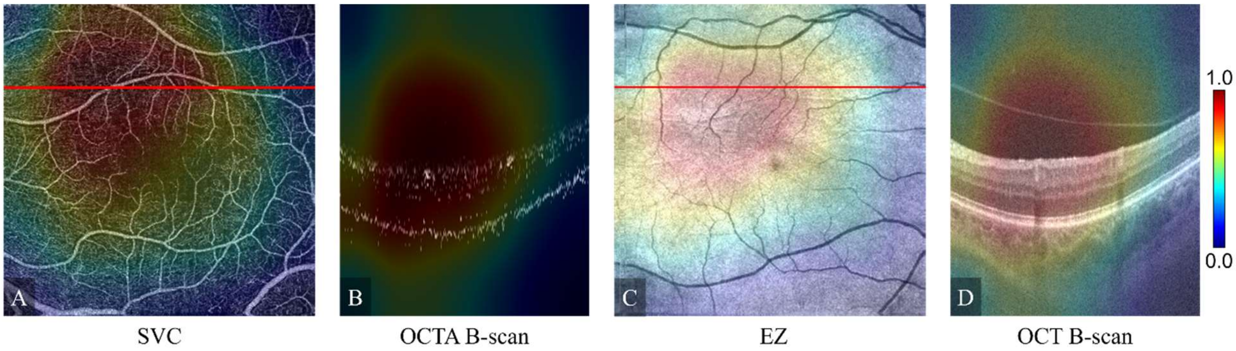


Figure 5.5: Class activation map (CAM) based on the DR output layer of the semi-sequential classifier for a correctly classified DR eye. (A) OCTA *en face* projection of the superficial vascular complex (SVC; inner 80% of the ganglion cell complex). The non-perfusion and low-perfusion areas were highlighted by the CAM. (B) Corresponding B-scan at the position of the red line in (A). (C) Structural OCT *en face* image of the ellipsoid zone (EZ; outer nuclear layer / ellipsoid zone boundary to ellipsoid zone / retinal pigment epithelium boundary). (D) Corresponding B-scan at the location of the red line in (C).

In order to aid confirmation of the model's outputs and interpret its decision-making, our framework also produces 3D class activation maps (CAM) (Fig. 5.5 and 5.6). We found that the CAMs frequently highlight pathology that is known to be associated with the diseases in this study, for example, non-perfusion and low-perfusion areas around the fovea were highly weighted for decision making of DR (highlighted regions in Fig. 5.5(A)). In AMD data, the CAMs highlighted most of the drusen areas (Fig. 5.6(C) and (D)).

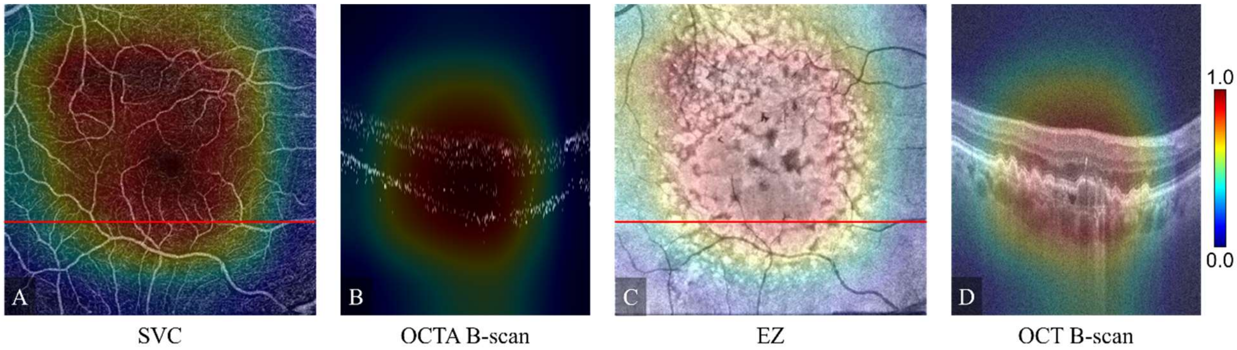


Figure 5.6: Class activation map (CAM) based on the AMD output layer of the semi-sequential classifier for a correctly classified AMD eye. (A) OCTA *en face* projection of the superficial vascular complex (SVC; inner 80% of the ganglion cell complex). (B) Corresponding B-scan at the position of the red line in (A). (C) Structural OCT *en face* image of the ellipsoid zone (Outer nuclear layer / ellipsoid zone boundary to ellipsoid zone / retinal pigment epithelium boundary). (D) Corresponding B-scan at the location of the red line in (C). The drusen area was highlighted by the CAM.

From glaucoma classification (Fig. 5.7), we can see that the semi-sequential classifier was mostly focused on the vanished nerve fiber layer, which is consistent with known glaucoma pathophysiology [138, 139] (Fig. 5.7(D)). In addition, the low perfusion area was also highlighted by the CAM (Fig. 5.7(A)). These attention maps offer many opportunities for us to validate the performance of deep learning frameworks and discover new potential biomarkers for disease understanding and diagnosis.

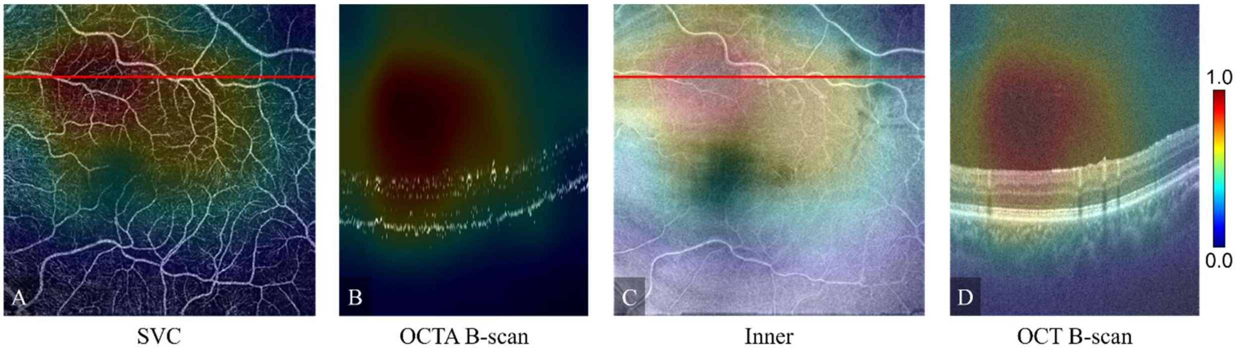


Figure 5.7: Class activation map (CAM) based on the glaucoma output layer of the semi-sequential classifier for a correctly classified glaucoma eye. (A) OCTA *en face* projection of the superficial vascular complex (SVC; inner 80% of the ganglion cell complex). The low perfusion area was highlighted (B) Corresponding B-scan at the position of the red line in (A). (C) Structural OCT *en face* image of the inner retina (Vitreous / inner limiting membrane boundary to outer plexiform layer / outer nuclear layer boundary). (D) Corresponding B-scan at the location of the red line in (C). The region of the vanished nerve fiber layer was highlighted.

5.5 Discussion

In this study, we proposed an automated diagnostic framework based on volumetric OCT/OCTA data that diagnoses DR, AMD, and glaucoma. The framework uses a semi-sequential classier which consists of two parts with identical architecture, one of which diagnoses DR and AMD, and the other of which diagnoses glaucoma. We found that this semi-sequential structure, which uses separate parts (classifiers) for AMD/DR and glaucoma, outperformed a single parallel classifier that learns to diagnose all three diseases. The framework achieved an AUC of ROC curve over 0.9 for the diagnosis of each disease. These results indicate that our automated framework achieved reliable DR, AMD, and glaucoma diagnosis performance using only a single ophthalmic imaging modality.

Compared to current deep-learning-aided eye disease diagnosis methods based on OCT/OCTA, our framework also includes several advantages. The first advantage is our framework can be used to diagnose DR, AMD, and glaucoma simultaneously, which could reduce the time and financial costs of screening. In addition, ophthalmologists could have a more comprehensive understanding of the eye condition of the referred patients based on our diagnosis results. The second advantage is the use of the whole 3D volume. Other approaches which rely on *en face* images are prone to segmentation errors and may miss important features without access to cross-sectional information (such as small drusen or retinal fluid). And traditional frameworks that perform diagnosis based on the presence or absence of known pathologic features may not account for undiscovered relevant features or information, and fail to utilize all of the information available in combined structural OCT and OCTA data volumes. In the contrast, our approach is biomarker/feature-agnostic, which means that correlations or structures within the data volume that may be difficult for a human to identify can still be incorporated into decision-making. As a corollary, our framework may also have a greater capacity to improve with more training data since a full data volume contains far more information than an image formed by projection.

Another significant advantage of our framework is the inclusion of 3D CAMs. Deep learning algorithms are often likened to "black-box", since their decision-making is difficult to interpret. This is problematic, since opaque decision-making may hide important biases that could prove to be disadvantageous for certain groups. The interpretability provided by the 3D CAMs would allow clinicians to verify and understand the diagnosis decisions and ensure they are correct, an essential requirement in any diagnostic framework. Compared to 2D CAMs, 3D CAMs indicate which retinal layer in each B-scan is relevant for each diagnosis. We verified that the CAM output by our model highlighted features known to be associated with each of the diseases examined in this study: non-perfusion areas in DR (Fig. 5.5(A)), drusen in AMD (Fig. 5.6(D)), and nerve fiber layers with abnormal structure in glaucoma diagnosis (Fig. 5.7(D)). Although the 3D CAMs did not demonstrate all features used for diagnosing eye diseases, they found many key features, indicating that our framework has successfully learned relevant features and that 3D CAMs could be useful in clinical review and sanity checks. In addition, the 3D CAMs were used to highlight the biomarkers which were selected by our framework, but not all the biomarkers were selected. That only some of the biomarkers were highlighted means, these biomarkers were already sufficient for our framework to make the diagnosis decision.

There are three aspects of the diagnosis performance of our framework that could be improved in future work. Firstly, our data set only contained healthy eyes or eyes that had one of the three diseases (DR, AMD, or glaucoma), whereas in clinical practice an eye may suffer from different condition (e.g., branch retinal vein occlusion) or even multiple diseases simultaneously (e.g., AMD with DR or AMD with retinitis pigmentosa). This limitation may lead to performance loss in our model if it were attempted on an eye with conditions that were not included in our data set. Secondly, the use of a semi-sequential structure increased the glaucoma diagnosis accuracy but also limited the framework for diagnosing eyes with both glaucoma and DR or glaucoma and AMD. This framework solved the multiple classification problem for a single diagnosis among three eye diseases, but future work will need to generalize our strategy in order to make multiple simultaneous diagnoses for more diseases or eyes with multiple diseases. Finally, some of the design choices that led to the second limitation were to improve glaucoma diagnostic performance (Table

5.2), but even so, the sensitivity for glaucoma diagnosis (71.67% ± 4.08%) was lower than the other two grades (90.00% ± 2.34% for DR and 88.28% ± 5.60% for AMD). Because only scans on macula were used in this study information from the optic disc, where glaucoma pathology is more prominent [140], was unavailable for decision making. Training on a larger dataset with cases of multiple diseases would likely improve performance for not only glaucoma but the other diseases in this study as well. In particular, the accuracy of the parallel classifier could probably be similar to the semi-sequential classifier in the main module if more glaucoma data for training was available. The framework limitation could therefore be solved by using a better-trained parallel classifier.

In addition to diagnosis performance, there are also limitations if we use our framework in real-world clinical applications right now. Our framework can only be used in clinics with both OCT and OCTA available. But this limitation will gradually disappear as OCTA applications become more widespread. In addition, the data set used in this study were all scanned by Avanti RTVue-XR in Casey Eye Institute, Oregon Health & Science University, and only scans with signal strength index above 50 were preserved. The diagnosis performance may be lower on external or lower quality data, or data scanned on other OCT devices. Therefore, to improve the clinical utility of our framework, data without these limitations will also be included in the future.

5.6 Conclusion

We proposed a deep-learning-aided DR, AMD, and glaucoma diagnostic framework that takes combined 3D structural OCT and OCTA data as inputs. Our framework achieved reliable performance on the diagnosis of each disease for which it was designed, and produces 3D CAMs that can be used to interpret the model's decision-making. By using our framework, the number of scanning procedures and eye exams required for the diagnosis of the three different eye diseases was reduced to just a single OCT/OCTA procedure. In addition, by using 3D data as inputs, our framework can totally avoid the influences from unstable retinal layer segmentation. At last, our results show that the biomarker-agnostic framework based on 3D OCT and OCTA could be beneficial for clinical practice.

6 Interpretable Diabetic Retinopathy Diagnosis based on Biomarker Activation Map

6.1 Abstract

Deep learning classifiers provide the most accurate means of automatically diagnosing diabetic retinopathy (DR) based on optical coherence tomography (OCT) and its angiography (OCTA). The power of these models is attributable in part to the inclusion of hidden layers that provide the complexity required to achieve a desired task. However, hidden layers also render algorithm outputs difficult to interpret. Here we introduce a novel biomarker activation map (BAM) framework based on generative adversarial learning that allows clinicians to verify and understand classifiers' decision-making. A data set including 456 macular scans were graded as non-referable or referable DR based on current clinical standards. A DR classifier that was used to evaluate our BAM was first trained based on this data set. The BAM generation framework was designed by combing two U-shaped generators to provide meaningful interpretability to this classifier. The main generator was trained to take referable scans as input and produce an output that would be classified by the classifier as non-referable. The BAM is then constructed as the difference image between the output and input of the main generator. To ensure that the BAM only highlights classifier-utilized biomarkers an assistant generator was trained to do the opposite, producing scans that would be classified as referable by the classifier from non-referable scans. The generated BAMs highlighted known pathologic features including nonperfusion area and retinal fluid. A fully interpretable classifier based on these highlights could help clinicians better utilize and verify automated DR diagnosis.

6.2 Introduction

Deep learning classifiers have achieved excellent performance in several automated eye disease diagnosis tasks [24-28]. Among these tasks, diabetic retinopathy (DR) diagnosis based on optical coherence tomography (OCT) and its angiography (OCTA) play an important role in ophthalmology since DR is a leading cause of preventable blindness globally and may be asymptomatic even in the referable stages [1-5]. Therefore, an efficient and reliable diagnosis system is essential in identifying DR patients at an early stage, when the disease has the best prognosis and visual loss can be delayed or deferred [4, 5]. In addition, OCT and OCTA (which can be acquired by the same device at the same time) can provide accurate DR diagnosis based on both the standard fundus photography-derived DR severity scales [13, 21-23] and the detection of diabetic macular edema (DME, an important DR pathology that cannot be accurately detected by fundus photography [14, 15]).

However, the high performance of current OCT/OCTA-based DR diagnosis often comes at the cost of inscrutable outputs [33, 34, 75, 76, 141]. The presence of hidden layers in classifier architectures renders a straightforward account of the classifier's action on inputs inaccessible and makes deep learning classifier outputs difficult to verify. In the absence of heuristic devices, deep-learning-aided DR diagnosis cannot be confirmed outside of manual grading, which largely defeats the purpose of automation. The poor interpretability can also obfuscate potential bias that could negatively affect performance in external data sets: a classifier trained only based on one data set may be biased when evaluated on the data from others if non-clinical biomarkers were utilized. These issues present a major hurdle for translating deep-learning-aided DR classifiers into the clinic [29-31].

Contemporary interpretability methods that have been used for deep-learning-aided DR diagnosis can be summarized in two categories. The first is methods which mainly focus on correlations between manually selected biomarkers and DR diagnostics [72-74, 142]. The selected biomarkers are first segmented from OCT and OCTA and then used to train the DR classifier. However, these methods have limited DR diagnosis performance since the classifiers could not learn from the much richer feature space latent in the entire OCT/OCTA data. The second and more common interpretability methods are attention maps. They indicate the relative importance of regions of an image for classifier decision making and indicate which features were useful for the DR diagnosis. However, these methods are originally developed for non-medical image recognition tasks (e.g., dog vs. cat classification, Fig. 6.1) [143-145], which is distinct from DR diagnosis in several regards and consequently poses many challenges in presenting clinically meaningful attention maps. In particular, unlike non-medical image classification where classes are typically identified by unique features not shared between the separate classes, in DR diagnosis different classes actually share most features (Fig. 6.1). Instead of containing unique identifying features, OCT/OCTA scans of non-referable DR lack features of referable cases. The identification of a non-referable DR case is therefore based on the absence, rather than presence, of specific pathologies.

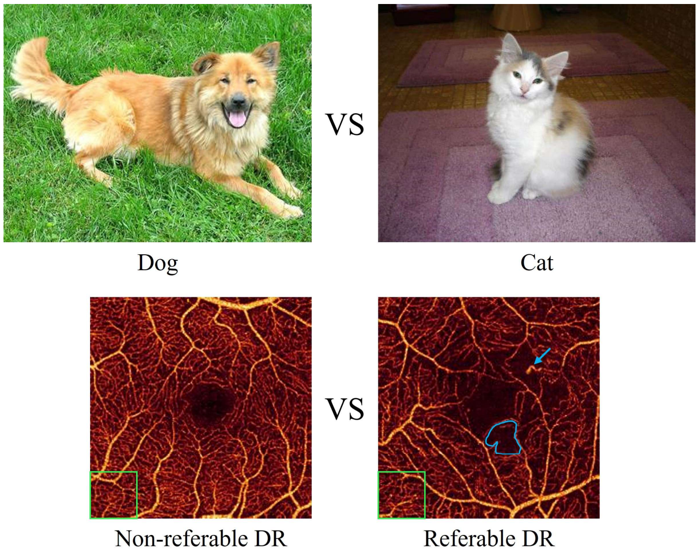


Figure 6.1: Comparison between non-medical and optical coherence tomography angiography (OCTA) images. Top row: the background in dog and cat pictures can be totally different because a classifier can learn to classify each by appealing to obvious, unique features (for example, the pets' faces). This contrasts with OCTA (bottom row) of diabetic retinopathy (DR), where features are largely shared between classes (here, a non-referable and referable DR case, respectively). Additionally, only the referable DR case has unique features (DR-related biomarkers). Features found in the healthy image, for example the large vessels with surrounding small capillary are also present in the referable DR image (green rectangles). The healthy image must therefore be identified based on a lack of features associated with the referable DR image (non-perfusion area and abnormal vessels marked by blue line and arrow, respectively).

To provide sufficient clinically meaningful interpretability for an OCT/OCTA-based DR classifier we propose a novel biomarker activation map (BAM) generation framework. The BAM generation framework is trained based on generative adversarial learning [146-149] to highlight the unique classifier-utilized biomarkers that only present in OCT/OCTA scans of referable DR. The main contributions of the present work are:

- We proposed the first interpretability method which was specifically designed for DR diagnosis based on both OCT and OCTA.
- Our design recognizes that a DR classifier should be highlight the unique biomarkers which only belong to the referable DR cases.
- We used generative adversarial learning to interpret a DR classifier instead of generating the pseudo-healthy or feature attribution maps based on the ground truth classes.
- We demonstrate that a generative adversarial learning approach can be used to provide interpretability to a DR classifier that achieved state-of-the-art performance.

6.3 Related work

To achieve interpretable DR diagnosis based on OCT and OCTA, several methods were proposed to illustrate which biomarkers were utilized in the decision-making of the classifier. Most of these methods belong to one of two categories: biomarker preselection methods and attention map methods.

6.3.1 Biomarker preselection methods

Biomarker preselection methods achieved interpretable DR diagnosis by using preselected and segmented biomarkers to train the classifier [72-74, 142]. Examples include H. S. Sandhu et al. and M. Alam et al. which proposed two computer-assisted diagnostic systems for DR diagnosis based on quantified features from OCTA [73, 74]. Several DR-related biomarkers, such as foveal avascular zone size and blood vessel tortuosity and density were extracted from OCTA images to train a DR classifier. In addition, Deep Mind proposed a retinal disease (including DR) diagnostic system based on several pre-segmented biomarkers from OCT [142]. The interpretability issue was well addressed in these DR diagnostic tasks since clinicians can clearly know which biomarkers were used to train the classifier. However, only some biomarkers were used in these methods, which means the classifiers could not learn from the much richer feature space latent in the full OCT/OCTA data volumes.

6.3.2 Attention map methods

Attention heatmaps are frequently used approaches for interpreting deep-learning-aided DR classifiers. Within this category, the gradient-based, class activation map (CAM)-based, and propagation-based methods are the most important techniques.

Gradient-based methods generate heatmaps based on the gradients of different convolutional layers with respect to the input [150-153]. Among these methods, Integrated Gradients which are based on multiplication between the average gradient and a linear interpolation of the input were evaluated on a DR classifier based on fundus photography [150]. To consider a more complete gradient, the FullGrad method also includes the gradient with respect to the bias term [153]. In practice, the outputs from these gradient-based methods are class-agnostic, resulting in heatmaps that are similar between different classes [154]. However, in DR diagnosis, non-referable and referable images share features, making it difficult for a gradient-based method to meaningfully distinguish pathologies from healthy tissues.

CAM-based methods are class-specific and are widely used in studies of deep learning DR diagnosis [71, 155-157]. The basic CAM method combines the class-specific weight and the output of the last convolutional layer before global average pooling to produce the attention map [71]. Grad-CAM introduces the gradients of target convolutional layers to the basic CAM [155]. Grad-CAMs have been widely used in deep-learning-aided DR diagnostic studies to provide interpretability to classifiers because it is easy to implement [33, 34, 75, 141]. However, most CAM-based methods only use the top convolutional layer, which generates low-resolution heatmaps [153]. In addition, the CAMs generated on lower convolutional layers are hard to interpret due to scattered features. Clinicians still need to manually discern the biomarkers inside the coarsely highlighted regions, which is time consuming and not clinically practical.

Propagation-based methods [154, 158-165] mostly rely on the deep Taylor decomposition (DTD) framework [158]. In these methods, the attention map is generated by tracing the contribution of the output back to the input using back propagation through the classifier based on the DTD principle. S. Bach et al. proposed the Layer-wise Relevance Propagation (LRP) method, which calculates the contribution of each element in the input back propagated from the output using the DTD principle [159]. However, some of these methods are class-agnostic in practical applications [154]. To solve this issue, class-specific propagation-based methods were proposed [164, 165]. J. Gu et al. proposed the contrastive-LRP method in which the contributions based on non-target classes are removed on average from the heatmap [164]. B. K. Iwana et al. proposed the softmax-gradient-LRP in which the contribution of each non-target class is removed based on their own probability value after softmax [165]. Compared to CAMs, these class-specific LRP methods can generate higher resolution attention maps but have many fewer applications in DR diagnosis due to accuracy concerns and implementation difficulties.

Several methods which do not belong to these three attention map categories have also been proposed to interpret deep learning classifiers. These include input-modification-based methods [166-171], saliency-based methods [172-175], an activation maximization method [176], an excitation backprop method [177], and perturbation methods [178, 179]. However, these methods do not in general achieve the accuracy of gradient-, CAM-, or propagation-based methods [154].

6.4 Materials

In this study, we included healthy and diabetic participants representing the full spectrum of diseases, from no clinically evident retinopathy to proliferative diabetic retinopathy. One or both eyes of each participant underwent 7-

field color fundus photography and an OCTA scan using a commercial 70-kHz spectral-domain OCT system (RTVue-XR Avanti, Optovue Inc) with 840-nm central wavelength. The scan depth was 1.6 mm in a 3.0 × 3.0 mm region (640 × 304 × 304 pixels) centered on the fovea. Two repeated B-frames were captured at each line-scan location. Blood flow was detected using the split-spectrum amplitude-decorrelation angiography algorithm [23, 57]. The OCT structural images were obtained by averaging two repeated and registered B-frames. Two continuously acquired volumetric raster scans (one x-fast scan and one y-fast scan) were registered and merged through an orthogonal registration algorithm to reduce motion artifacts [58]. In addition, the projection-resolved algorithm was applied to all data volumes to remove flow projection artifacts in posterior layers [68, 69]. Scans with a signal strength index (SSI) lower than 50 were excluded.

A masked trained retina specialist (Thomas S. Hwang) graded the photographs based on the Early Treatment of Diabetic Retinopathy Study (ETDRS) scale [11, 12] using 7-field color fundus photographs. The presence of diabetic macular edema (DME) was determined using the central subfield thickness from structural OCT based on the DRCR.net standard [4]. We defined non-referrable DR as ETDRS level better than 35 and without DME, and referrable DR as ETDRS level 35 or worse, or any ETDRS score with DME [5]. The participants were enrolled after informed consent in accordance with an Institutional Review Board (IRB # 16932) approved protocol. The study complied with the Declaration of Helsinki and the Health Insurance Portability and Accountability Act.

To generate the input data set for the referable DR classifier, the following retinal layer boundaries were automatically segmented using commercial software in the spectral-domain OCT system (Avanti RTVue-XR, Optovue Inc): the vitreous / inner limiting membrane (ILM), inner plexiform layer (IPL) / inner nuclear layer (INL), and the outer plexiform layer (OPL) / outer nuclear layer (ONL) (Fig. 6.2). In addition, the automated layer segmentation was manually corrected by graders using custom COOL-ART software in cases where pathology caused segmentation errors in the commercial software [78].

For each case, a two-channel input was generated based on the segmented boundaries (Fig. 6.2). The first input consists of the *en face* OCTA image was generated using maximum projection of the superficial vascular complex (SVC), defined as the inner 80% of the ganglion cell complex (GCC), which included all structures between the ILM and IPL/INL border [18, 41] (Fig. 6.2) [70]. *En face* structural OCT images were generated through average projection (Vitreous/ILM to OPL/ONL) and used as the 2nd channel of each input (Fig. 6.2).

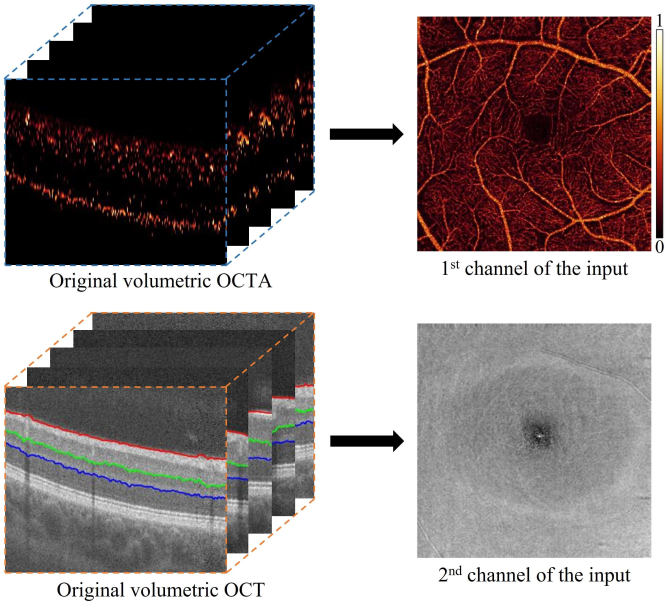


Figure 6.2: Input generation. Superficial vascular complex (inner 80% from vitreous to inner plexiform layer) *en face* maximum projections were generated from the volumetric OCTA and used as the 1st channel of the input [18, 41]. An *en face* mean projection of the inner retina (vitreous through the outer plexiform layer) was generated from volumetric OCT data and used as the 2nd channel of the input. Three boundaries- vitreous / inner limiting membrane (red), inner plexiform / inner nuclear layers (green), and outer plexiform / outer nuclear layers (blue)- were segmented for the generation process in both the first and second channels.

6.5 Methods

To provide meaningful interpretability to a DR classifier, our BAM generation framework was trained to learn which biomarkers were important for classifier decision making. In training, we combined two generators (a main and assistant) to learn the necessary changes (classifier-utilized biomarkers) which would change classifier decision making. In this study, the positive and negative decision making of the classifier corresponded to referable and non-referable DR, respectively. The main generator was trained to forge a negative output by adding changes to a positive input. To reduce unnecessary changes made by the main generator, inspired by cycle-consistency generative adversarial learning [149], the assistant generator was trained to do the opposite. However, the BAM was generated as the difference image between the output and input of the main generator, only. All the biomarkers highlighted in the BAM were the classifier-utilized biomarkers, which could be textures, artifacts, shadows, etc. that were learned and utilized by the classifier in the decision making. To be clear, the classifier-utilized biomarkers could be different from the clinical biomarkers which mainly were the pathologies correlated to the selected disease (referable DR in this study).

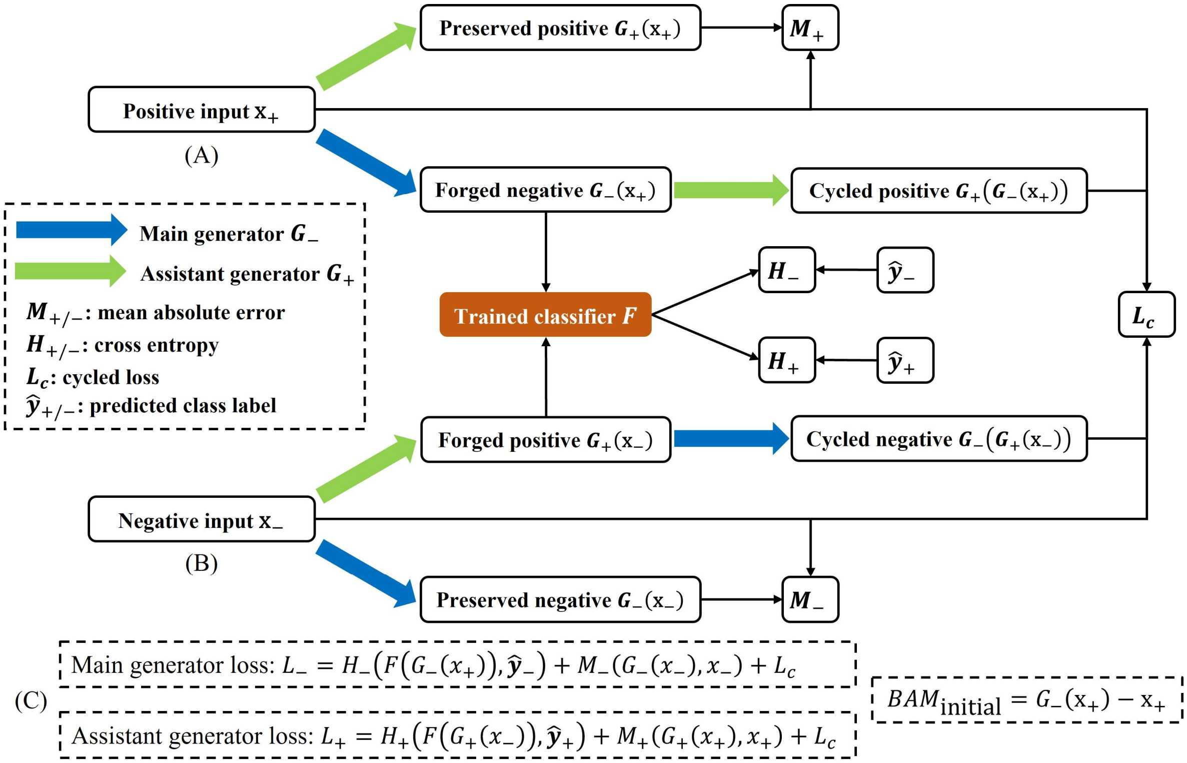

Preserved positive $G_+(x_+)$
M_+
Positive input x_+
(A)
Forged negative $G_-(x_+)$
Cycled positive $G_+(G_-(x_+))$
Main generator G_-
Assistant generator G_+
$M_{+/-}$: mean absolute error
$H_{+/-}$: cross entropy
L_c: cycled loss
$\hat{y}_{+/-}$: predicted class label
H_-
$\hat{y}_-$
Trained classifier F
L_c
H_+
$\hat{y}_+$
Forged positive $G_+(x_-)$
Cycled negative $G_-(G_+(x_-))$
Negative input x_-
(B)
Preserved negative $G_-(x_-)$
M_-
(C)
Main generator loss: $L_- = H_-(F(G_-(x_+)), \hat{y}_-) + M_-(G_-(x_-), x_-) + L_c$
Assistant generator loss: $L_+ = H_+(F(G_+(x_-)), \hat{y}_+) + M_+(G_+(x_+), x_+) + L_c$
$BAM_{\text{initial}} = G_-(x_+) - x_+$

Figure 6.3: BAM generation framework architecture and training process. Two generators were trained to produce the biomarker activation maps (BAMs). (A) Positive class inputs are acted on by the main and assistant generators (blue and green arrows, respectively). The main generator was trained to produce output images that would be classified as the negative class by the classifier. The assistant generator performs the inverse task, producing outputs that the classifier would diagnose as the positive class. (B) Training from negative inputs is symmetric, with the labels switching roles. (C) Training of the main and assistant generators occurred simultaneously. This scheme prevents the main generator from overfitting the negative class inputs by producing unnecessary changes to the inputs.

6.5.1 Training

We consider a DR classifier F trained to predict positive and negative class labels $\hat{y}$ from input data $\mathbf{x}$ according to $\hat{y} = F(\mathbf{x})$. The positive and negative class y_+ and y_- indicated referable and non-referable DR, respectively. We note that, in general, the predicted class label $\hat{y}$ is not identical to the true class labels, y, since most classifiers are not perfect; however, in this work we are primarily concerned with classifier outputs, not the ground truth classifications. Accordingly, we define two input classes $\mathbf{x}_+$and $\mathbf{x}_-$ according to $F(\mathbf{x}_{+/-}) = \hat{y}_{+/-}$, *i.e.* $\mathbf{x}_+$ corresponds to data that was predicted by the classifier to belong to the positive class (i.e. referable DR), and $\mathbf{x}_-$ to the negative class. In the BAM framework we seek to train a main generator to transform data so that it is always classified as negative by the classifier (Fig. 6.3(A)); that is, we seek a generator G_- such that $F(G_-(\mathbf{x})) = \hat{y}_-$. In the case that the input data was originally classified as positive, i.e. $\mathbf{x} = \mathbf{x}_+$, this creates "forged data", in which the target output classification $\hat{y}_-$ differs from the classification of the input for which $F(\mathbf{x}_+) = \hat{y}_+$. If, alternatively, G_- operates on data $\mathbf{x}_-$ already predicted to belong to the negative class, the desired output classification matches the input, i.e. $F(G_{-}(\boldsymbol{x}_{-})) = \hat{y}_{-}$, creating "preserved data." Both forged and preserved data were used during the training process to calculate loss by comparing to their corresponding ground truths. Specifically, the cross-entropy loss H_{-} between the classifier prediction on the forged data $F\big(G_{-}(\mathbf{x}_{+})\big)$ and the desired prediction $\hat{y}_{-}$ was used to train the generator to produce data resembling the desired class. In addition, to prevent large changes to the main generator output, the mean absolute error loss M_{-} between the raw input $\mathbf{x}_{-}$and the preserved data $G_{-}(\mathbf{x}_{-})$ was included in the loss.

However, simply optimizing over forged and preserved data can lead to overfitting, in which the main generator learns to modify features that were not utilized by the classifier F (e.g., shared features between $\mathbf{x}_{+}$and $\mathbf{x}_{-}$) in order to achieve the desired output label $\hat{y}_{-}$. To ensure that the main generator only learns to remove relevant features we also simultaneously trained an assistant generator G_{+} that performs the inverse task; that is, we desire the trained assistant generator to produce $F\big(G_{+}(\mathbf{x})\big) = \hat{y}_{+}$ (Fig. 6.3(B)). Note that, like the main generator, this produces both preserved and forged data, since the assistant generator also acts on both $\mathbf{x}_{+}$ and $\mathbf{x}_{-}$. The assistant generator is used in conjunction with the main generator to produce "cycled data" $G_{-}(G_{+}(\mathbf{x}_{-}))$ and $G_{+}\big(G_{-}(\mathbf{x}_{+})\big)$ created by allowing the main and assistant generator to operate on data forged by the other. The cycled loss, defined as the mean absolute errors between the original and cycled data

$$L_C = \frac{1}{N_{-}} \sum_i \left|\mathbf{x}_{-,i} - G_{-}\left(G_{+}(\mathbf{x}_{-,i})\right)\right| + \frac{1}{N_{+}} \sum_j \left|\mathbf{x}_{+,j} - G_{+}\left(G_{-}(\mathbf{x}_{+,j})\right)\right| , \tag{15}$$

where $N_{+}$and N_{-} are the pixel number of positively and negatively classified images, respectively, can then be included in the overall loss function in order to ensure that only features utilized by the classifier F are modified. The overall loss for each generator is then given by the sum of the cross-entropy loss between forged labels and predicted labels, the mean absolute error loss between preserved data and input data, and the cycled loss:

$$\begin{aligned} L_{-} &= H_{-}\big(F\big(G_{-}(x_{+})\big), \hat{\boldsymbol{y}}_{-}\big) + M_{-}(G_{-}(x_{-}),\ x_{-}) + L_c \\ L_{+} &= H_{+}\big(F\big(G_{+}(x_{-})\big), \hat{\boldsymbol{y}}_{+}\big) + M_{+}(G_{+}(x_{+}),\ x_{+}) + L_c \end{aligned} \tag{16}$$

6.5.2 Generator architecture

Both the main and assistant generators were constructed based on a customized U-shape residual convolutional neural network [60] (Fig. 6.4). The output is calculated as the sum of input and Tanh output since both generators are trained to only change necessary biomarkers that are utilized by the classifier. To ensure the generator output has the same value range as the input, clipping was used after the summation (the values higher or lower than the original maximum or minimum of the input will be set to the maximum or minimum values, respectively). In addition, zero initialization is used for the last convolutional layer to ensure the initial BAM was a zero matrix before the training. This initialization strategy could avoid changes to biomarkers which were not utilized by the classifier in the beginning of the training.

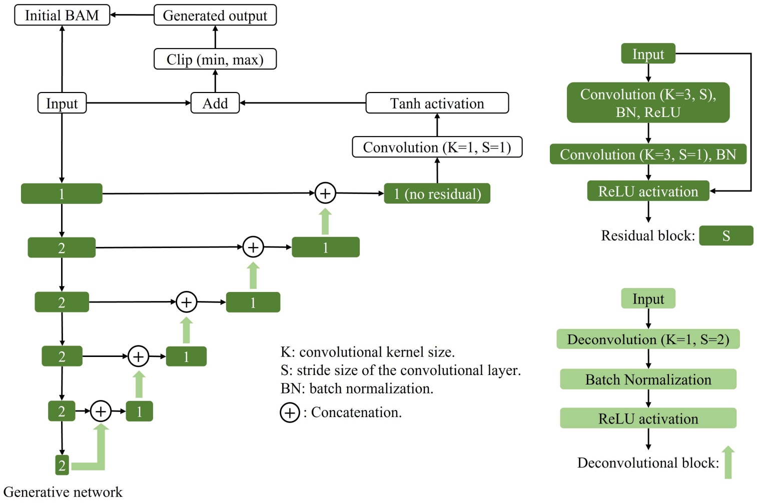


Figure 6.4: Detailed architecture of the main generator. The dark green patches and pale green arrows represent the residual and deconvolutional block, respectively. The number in the dark green patch is the stride size. The number of blocks can be adjusted based on the input. The architecture of the assistant generator is identical.

6.5.3 Model selection and biomarker activation map generation

After the training, the final model (including both trainable parameters and hyper-parameters) for BAM generation was selected based on the validation loss $L\_$ of the main generator. Loss from the assistant generator was not considered because, unlike the main generator which only needs to remove the unique biomarkers (pathologies) at specific locations, there is no *a priori* reason for the assistant generator to add pathology-like features at a particular location. This resulted in high variability in assistant generator output, which would make model selection based on assistant generator loss unreliable.

The initial BAM is calculated between the output and input of the main generator. Since this is a difference image it can have both positive and negative pixel values. The absolute difference between these values represents the overall contribution each biomarker made to the classification. Alternatively, positive/negative values indicate regions in which pixel values in the output were increased/reduced relative to the input in order for the classifier to produce a negative (non-referable DR) classification. By keeping track of these sign differences, we can understand more about how a DR classifier understands different biomarkers. Accordingly, the output of our framework is two processed BAMs, with the first obtained by measuring the absolute value of all differences

$$BAM_{abs} = f_g(|G_-(\mathbf{x}_+) - \mathbf{x}_+|), \tag{17}$$

while the second is generated by separating positive and negative values

$$\begin{aligned} BAM_{+/-} = f_g\big(\text{ReLU}(G_-(\mathbf{x}_+) - \mathbf{x}_+)\big) - \\ f_g\big(\text{ReLU}(\mathbf{x}_+ - G_-(\mathbf{x}_+))\big) \end{aligned}, \tag{18}$$

where f_g is a Gaussian filter and ReLU is the activation function which only preserves positive values [81, 82]. The BAM_{abs} then indicates the overall contribution of each biomarker to classifier decision making, while the $BAM_{+/-}$ indicates how different biomarkers were learned by the classifier.

6.5.4 Implementation details

To evaluate our BAM generation framework, a classifier that took *en face* OCT and OCTA data as inputs to diagnose referable DR was constructed based on a VGG19 architecture [87] with batch normalization and only one fully connected layer. Two classifier-utilized DR biomarkers- non-perfusion area (NPA) and fluids- were used to evaluate the OCTA and OCT channels of the generated BAMs, respectively. The DR classifier was only used to evaluate our BAMs. The development of this classifier was not a part of our BAM generation framework.

Before the evaluations, 60%, 20%, and 20% of the data were split for training, validation, and testing, respectively. Care was taken to ensure data from the same subjects are only included in one of either the training, validation, or testing data sets. The BAM framework was trained, validated, and evaluated respectively based on the same data set of the DR classifier. Two stochastic gradient descent optimizers with Nesterov momentum (momentum = 0.9) were used simultaneously to train the generators. Hyperparameters during training included a batch size of 3, 500 training epochs, and a learning rate of 0.0005 used for all the training steps. The trained main generator with lowest validation loss was selected for the final evaluation. In the evaluation, the binary maps for NPA and fluids were generated by two deep-learning-aided segmentation methods previously designed by our group, respectively [180, 181]. For qualitative analysis, we used perceptually uniform color maps to illustrate the BAMs [182]; compared to the traditional Jet colormap, perceptually uniform color maps have even color gradients that can reduce visual distortions causing feature loss or the appearance of false features [183]. For quantitative analysis, the F1-socre, intersection over union (IoU), precision, and recall were calculated between each channel of the generated BAMs and segmented DR biomarkers (NPA and fluid).

This study was implemented in TensorFlow version 2.6.0 on Ubuntu 20.04 server. The server has an Intel(R) Xeon(R) Gold 6254 CPU @ 3.10GHz ×2, 512.0 GB RAM and four NVIDIA RTX 3090 GPUs. But only one GPU was used in this study. The average training time for each epoch was 15 seconds.

6.5.5 Sanity checks

To assess if our BAM was correlated with the interpreted DR classifier two sanity checks were performed [184]. First, we performed model parameter and data randomization tests. In the model parameter randomization test DR classifier parameters were divided into six parts based on the five max pooling layers. The parameters in each of the six parts were randomized in two ways. In cascading randomization we randomized the parameters from the top part of the trained DR classifier (after last max pooling) successively all the way to the bottom part (before first max pooling). In the independent randomization, the parameters in each part were randomized independently. All the parameter randomizations created 11 different models. In the data randomization test, a model with the same architecture of the DR classifier was trained based on randomized labels. The model training was stopped after the training accuracy reached 95%. The generated BAMs of these 12 models were compared with the BAMs generated based on the original DR classifier. For quantitative comparison, we calculated the spearman rank correlation, the structural similarity index (SSIM), and the Pearson correlation of the histogram of gradients (HOGs) between the $BAM_{+/-}$ generated on these models and the original classifier.

6.6 Results

We recruited and examined 50 healthy participants and 305 patients with diabetes. After the DR severity grading, 199 non-referable and 257 referable DR inputs were used to train, validate, and evaluate the DR classifier and our framework (Table 6.1). In the evaluation, the classifier achieved an area under the receiver operating characteristic curve (AUC) of 0.97 and a quadratic-weighted kappa of 0.85, which is on par with the performance of ophthalmologists and therefore adequate to evaluate our BAM framework [112].

Table 6.1: Data distribution

Severity	Number	Age, mean (SD), years	Female, %
Non-referable DR	199	48.8 (14.6)	50.8%
Referable DR	257	58.4 (12.1)	49.0%

6.6.1 Sanity checks

To firstly demonstrate that our BAM could correctly learn the interpretability of the DR classifier, two tests were performed based on randomized parameters and labels, respectively [184]. Most models in the parameter randomization test predicted all the data as the same class (either non-referable or referable DR), which means no BAMs were generated for these models since the training of our framework needed data predicted as both classes. In the cascading randomization, only the model with randomized parameters after the 4th max pooling layer had predictions for both classes. In the independent randomization, only the model with randomized parameters before the first max pooling layer had predictions for both classes. Therefore, these two models were used to represent the cascading and independent parameter randomizations, respectively. The three BAMs generated in both parameter and label randomization tests showed large differences compared to the original BAMs (Fig. 6.5 and Table 6.2), which shows our BAMs are sensitive to potential interpretability changes of the classifier. The two models based on randomized labels and cascading parameter randomization highlighted totally different regions compared to the original BAMs. The highlighted regions in the model based on independent randomization had some overlaps with the original BAMs. But the differences between these two BAMs were still large and clear.

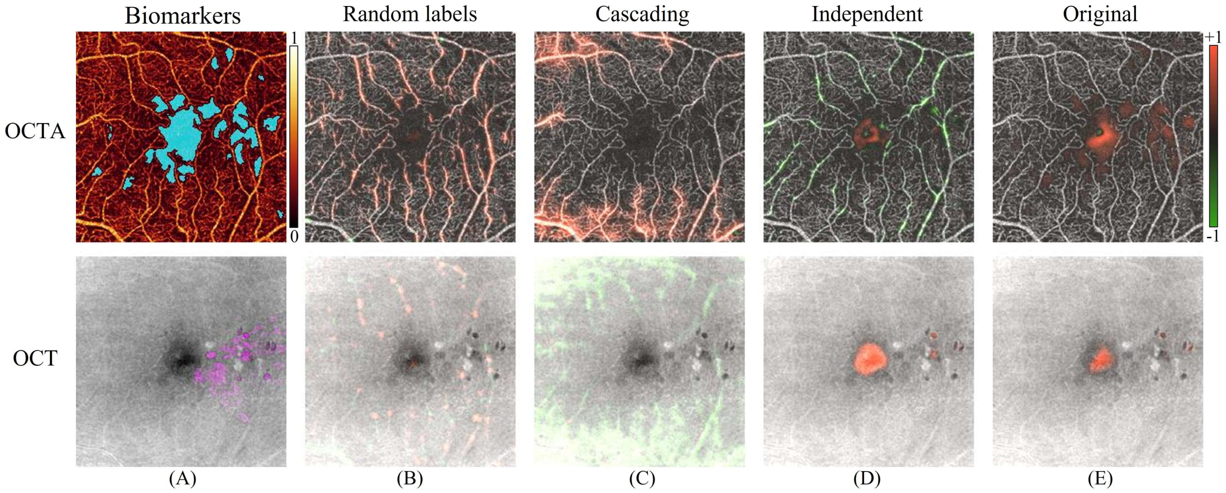


Figure 6.5: BAMs generated in the sanity checks which showed our BAM was sensitive to the interpretability changes between different randomized models. (A) Segmented non-perfusion areas and fluids. (B) The $BAM_{+/-}$ of the model based on randomized labels. (C) The $BAM_{+/-}$ of the model based on cascading parameter randomization. (D) The

$BAM_{+/-}$ of the model based on independent parameter randomization. (E) The $BAM_{+/-}$ generated based on the original DR classifier.

Table 6.2: Quantitative sanity checks

Models	Spearman rank correlation		SSIM		HOGs	
	OCTA	OCT	OCTA	OCT	OCTA	OCT
Random labels	-0.11 ± 0.17	0.22 ± 0.22	0.13 ± 0.07	0.75 ± 0.11	0.07 ± 0.06	0.01 ± 0.09
Cascading random	-0.21 ± 0.06	0.11 ± 0.03	0.18 ± 0.10	0.24 ± 0.11	0.00 ± 0.06	-0.09 ± 0.06
Independent random	0.45 ± 0.14	0.82 ± 0.11	0.30 ± 0.12	0.91 ± 0.12	0.05 ± 0.05	0.54 ± 0.15

6.6.2 Qualitative analysis

To demonstrate the utility of the proposed BAM framework we consider an eye correctly classified as referable DR by the DR classifier (Fig. 6.6). Compared to the clinical DR biomarkers (Fig. 6.6(C) and 6.6(H)), our BAMs highlighted similar regions (Fig. 6.6(D) and 6.6(I)), demonstrating that BAMs can improve interpretation of the DR classifier result. Specifically, the BAM_{abs} output indicates that the classifier focused on important pathologic features which (NPA and retinal fluid) [180, 181] in decision making. Additionally, the $BAM_{+/-}$ indicates that the pathological NPA was correctly differentiated from the foveal avascular zone, which is avascular but not pathological. In non-referable eyes, the pixel values near fovea should span a wide range corresponding to- from highest to lowest- larger vessels around fovea, small capillary structure, and the foveal avascular zone. These pixel populations are compressed in an *en face* OCTA angiograms of a referable DR eye (Fig. 6.6(A)). By adding positive values (white dots) around fovea and negative values in the fovea, the main generator expanded the range of pixel values to craft an image that looked like a non-referable eye to the classifier (Fig. 6.6(B)). The structure of the generated image shows that the classifier learned the anatomical structure near fovea in a normal eye. We also noticed that, with the exception of NPA and fluid, other DR-related biomarkers (such as microaneurysms, Fig. 6.6(I)) were ignored by the BAMs, which means these biomarkers were not utilized by the classifier in decision making.

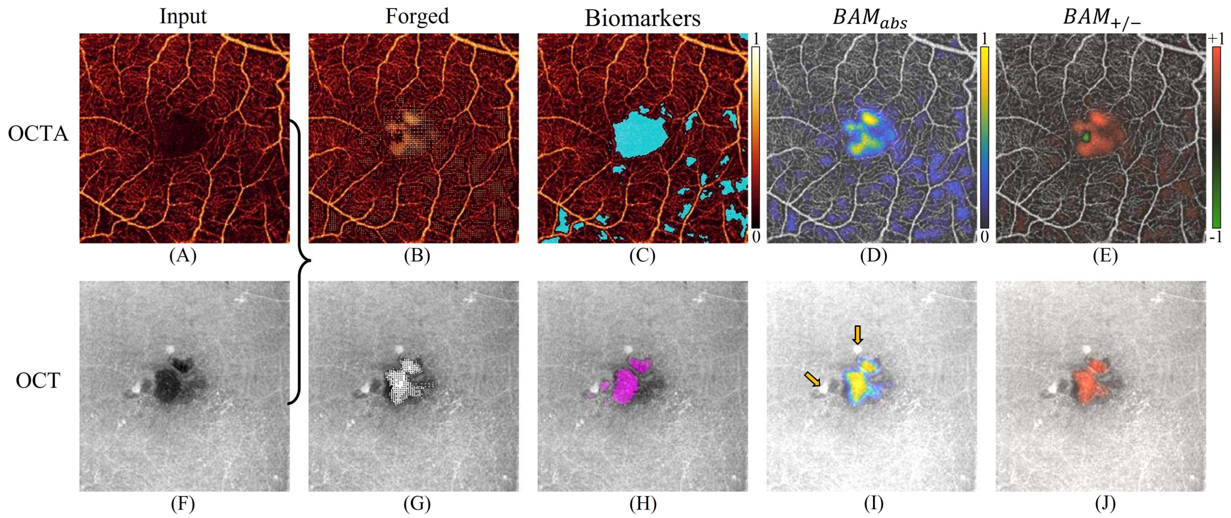


Figure 6.6: BAMs for a correctly predicted referable DR scan. (A) Superficial vascular complex *en face* maximum projection, which is the first input channel for the BAM framework. (B) The forged main generator output image for this channel, which should be classified as non-referable DR by the classifier. Compared to (A), positive white dots and negative dark regions were added by the main generator. (C) Segmented non-perfusion area (NPA), which is an important DR-related biomarker (marked by cyan), based on a previously reported deep learning method [180]. (D) The BAM_{abs} is the absolute difference between (B) and (A) after Gaussian filtering (Eq. 17). The highlighted areas are similar to the segmented NPA in (C). (E) The $BAM_{+/-}$ is the (non-absolute) differences between (B) and (A) after Gaussian filtering (Eq. 18). Red highlights pathological non-perfusion area while the foveal avascular zone (highlighted by green), which is not pathological, was identified as a separate feature by the classifier network. (F) *En face* mean projection over the inner retina, which is the second input channel to the classifier. (G) Main generator output for this channel which should be classified as non-referable DR by the classifier. Compared to (F), positive white dots were added by the main generator. (H) The inner mean projection of the segmented fluid (an important DR-related biomarker, marked by magenta) based on a previously reported deep learning method [181]. (I) The BAM_{abs} is the absolute difference between (G) and (F) after Gaussian filtering (Eq. 17). The highlighted areas resemble the fluid regions in (H). However, the microaneurysms (hyperreflective spots marked by orange arrows) were not highlighted, which means this biomarker was not utilized by the classifier. (J) The $BAM_{+/-}$ is the difference between (G) and (F) after Gaussian filtering (Eq. 18). The red highlighted areas also focus on fluids, and no green highlighted area is shown, indicating that the network did not learn separate fluid features. This is anatomically accurate, since unlike NPA (which is non-pathologic in the foveal avascular zone) all retinal fluid is pathologic.

To demonstrate how our BAM could help detect biased classifiers 5 rectangular artifacts were added to both the OCTA and OCT channels of all the referable DR scans. In the training, the classifier was forced to utilize the artifacts to make predictions, thereby introducing bias into the classifier output. Our BAM network trained on this biased classifier was able to successfully highlight the artifactual features (Fig. 6.7). The highlighted artifacts (Fig. 6.7(B) and 6.7(E)) show that the clinical biomarkers were not utilized by the biased classifier, despite it reaching 100% accuracy. The $BAM_{+/-}$ (Fig. 6.7(C) and 6.7(F)) reveals how the biased classifier utilized the artifacts. In addition, these artifacts could be much less obvious in a real-world application (such as Fig. 6.7(D)).

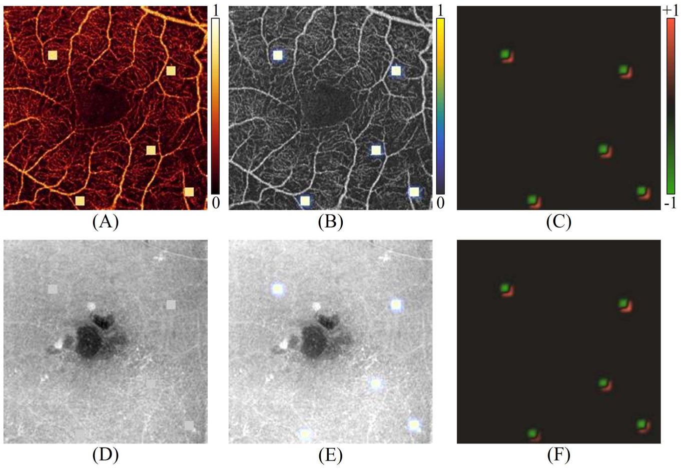


Figure 6.7: BAMs for a correctly predicted referable DR scan based on a biased classifier. (A) Superficial vascular complex *en face* maximum projection with added artifacts. (B) OCTA channel of the BAM_{abs} which only highlighted the classifier-utilized artifacts. (C) OCTA channel of the $BAM_{+/-}$ without coverage of input. (D) *En face* mean projection over the inner retina with added artifacts. (E) The OCT channel of BAM_{abs} which only highlighted the classifier-utilized artifacts. (F) OCT channel of the $BAM_{+/-}$ without coverage of input.

To demonstrate the advantages of the BAM generation framework, we compared its output to gradient-based [112], propagation-based [165], and CAM-based methods [71] (Fig. 6.8). Compared to the attention maps generated by these methods in referable DR scans, the BAMs showed sharper distinctions between significant and insignificant regions for decision making, highlighted features at a higher resolution, and indicated that the classifier could distinguish different features. In addition, the BAM framework could separately highlight the important features in *en face* OCTA and structural OCT rather than blending them together. This improves interpretability since the features the network is trying to learn do not necessarily overlap in the separate channels. For example, NPA does not always overlap with diabetic macular edema. Especially for graders reviewing the images, if structural OCT and OCTA features are not separated it may be unclear if healthy regions in one channel are being misinterpreted as pathologic, or if the pathology is in the other channel. However, gradient-based, CAM-based, and propagation-based methods all highlighted similar regions between two different channels (Fig. 6.8). Several NPAs were ignored in the OCTA channel. In addition, small retinal fluids, which were highlighted by our BAMs, were also ignored by these three attention maps (marked by orange arrows in Fig. 6.8(B)).

In non-referable DR scans our BAM still highlighted a sharp foveola because due to the classifier needing to segment non-pathological NPA in the foveal avascular zone. Since this feature exists in both referable and non-referable classes, it is also highlighted in the non-referable case by the BAM (Fig. 6.8(C)). The gradient-based and CAM-based maps of the non-referable class highlighted similar areas compared to the maps for the referable DR class, which means their highlighting was incorrect in non-referable DR class. The propagation-based method correctly highlighted the surrounding areas of foveola, but the highlighting was inaccurate since areas without highlighting were much larger than the foveola.

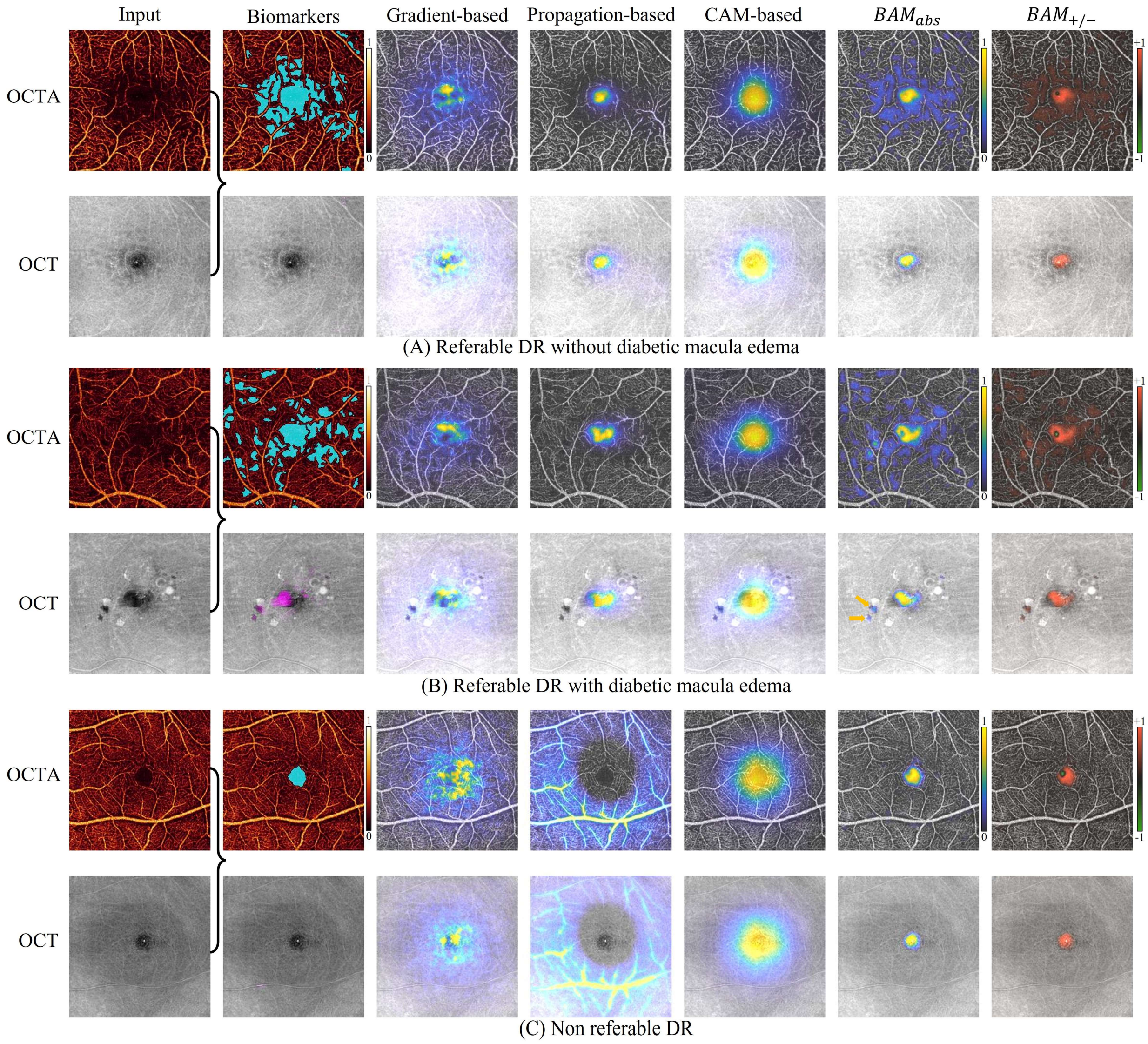


Figure 6.8: Comparison between the BAM generation framework and three other prominent attention maps (gradient, propagation, and class activation) [71, 112, 165]. The biomarkers column shows non-perfusion area (NPA, marked by cyan) in OCTA channel and retinal fluid (marked by magenta) in the structural OCT channel, respectively, both segmented using previously reported deep learning methods [180, 181]. Compared to the other three attention maps, our BAMs accurately highlight each classifier-selected biomarker at higher resolution and highlighted just the classifier-selected biomarkers. In addition, the normal tissues (such as vessels between NPAs) that were not selected by the classifier were not highlighted by our BAMs. (A) Results based on a referable DR case without diabetic macular edema (DME). (B) Results based on a referable DR case with DME. Small fluids were sharply highlighted by the BAMs (marked by orange arrows). (C) Results based on a non-referable DR case. The BAM highlighted a sharp foveola because our BAM would always highlight the areas which were learned as referable DR biomarkers by the classifier. The gradient-based and CAM-based maps of the non-referable class highlighted similar areas compared to the maps of referable DR class, which means their highlighting was incorrect in non-referable DR class. The

propagation-based method correctly highlighted the surrounding areas of foveola, but the highlighting was inaccurate since areas without highlighting were much larger than the foveola.

6.6.3 Quantitative analysis

To compare our BAM with the other three attention maps quantitatively, the F1-score, IoU, precision and recall were calculated between the segmented biomarkers and binary masks of each attention map (Table 6.3). The binary mask of each map was generated using threshold *mean* + *std* × *h* on all positive values. For each attention method, the threshold h was selected based on the highest average F1-score. The OCTA channels of each attention map were compared with segmented NPAs. On this channel, our BAM achieved significantly higher performance than the other three attention maps (Table 6.3). On the OCT channels, which were compared with segmented fluids, our BAM still achieved higher performance based on most measurements. Only recall for fluids was lower than the other methods. But the higher precision and lower recall of the BAM actually demonstrate that our method mostly focused on the part of the DR biomarkers which were utilized by the classifier while ignored the healthy tissues. The lower precision and higher recall of these established attention maps can be attributed to the fact that they highlighted a large area which included more healthy tissues than DR biomarkers, which is not clinically meaningful. In addition, all four attention maps achieved higher performance on the OCTA channel than the OCT channel, which means the classifier was more focused on the NPAs rather than fluids.

Table 6.3: Quantitative comparison

Methods		Gradient	Propagation	CAM	BAM
Inference time (s/scan)		0.10	0.07	0.05	0.07
F1-score	NPAs	0.39 ± 0.07	0.42 ± 0.08	0.44 ± 0.09	**0.63 ± 0.09**
	Fluids	0.10 ± 0.15	0.11 ± 0.17	0.11 ± 0.17	**0.13 ± 0.20**
IoU	NPAs	0.24 ± 0.05	0.27 ± 0.06	0.29 ± 0.07	**0.47 ± 0.09**
	Fluids	0.06 ± 0.10	0.07 ± 0.12	0.07 ± 0.12	**0.09 ± 0.14**
Precision	NPAs	0.40 ± 0.08	0.40 ± 0.11	0.46 ± 0.10	**0.58 ± 0.15**
	Fluids	0.07 ± 0.11	0.08 ± 0.15	0.08 ± 0.13	**0.20 ± 0.31**
Recall	NPAs	0.41 ± 0.14	0.48 ± 0.12	0.46 ± 0.15	**0.72 ± 0.07**
	Fluids	0.54 ± 0.34	**0.62 ± 0.35**	0.54 ± 0.36	0.21 ± 0.29

Based on the generated BAMs (Fig. 6.6 and 6.8, Table 6.3), the interpretability of the DR classifier can be summarized as follows. First, only parts of the NPA and fluid regions were utilized by the classifier (Fig. 6.6 and 6.8). Other DR-related biomarkers (such as microaneurysms, Fig. 6.6(I)) were not utilized by the classifier. Second, pathological NPAs were correctly differentiated from the foveal avascular zone by the classifier (Fig. 6.6(E)). Third, the classifier utilized the foveola in OCT channel even though this dark region was not caused by fluids (BAMs in

Fig. 6.8(A)). Lastly, more NPAs were utilized than fluid areas in the decision making of the classifier (Fig. 6.6 and 6.8, Table 6.3).

6.6.4 Biomarkers analysis

As sanity checks, qualitative and quantitative analyses above demonstrated that our BAM could provide sufficient interpretability to a DR classifier based on OCT/OCTA. To find all the biomarkers which could contribute to referable DR diagnosis, the BAMs were also generated for two classifiers (0.75 and 0.71 kappas) trained based only on the OCTA and OCT scans, respectively (Fig. 6.9). In the OCTA scans, compared to the DR classifier trained with two channel inputs, the BAMs also highlighted most of the NPAs but with basically equal intensity (Fig. 6.9(D)), which means prediction contributions from the NPAs outside the foveola were improved when OCTA was the only input. In addition, small parts of the vessels with higher intensities were also highlighted (Fig. 6.9(E)). In the OCT scans, compared to the OCT channel of Fig. 6.7(B), all the fluid and other hyperreflective regions were highlighted by the BAMs (Fig. 6.9(I)). In addition, hyperreflective spots which were not highlighted before were also highlighted this time (Fig. 6.9(J)) because referable DR could not be detected only based on fluid. In summary, the NPAs, fluid, and abnormal hyperreflective spots (could be exudation, calcification, and microaneurysm) could all contribute to the deep-learning-aided DR diagnosis, which is consistent with clinical findings. In addition, some vessel parts with high intensity, hypo-reflective areas without fluids, and hyperreflective spots without pathologies were also highlighted by the BAMs. The highlighting of these non-clinical biomarkers may be caused by the imperfect classifiers or potential correlations that have not been found.

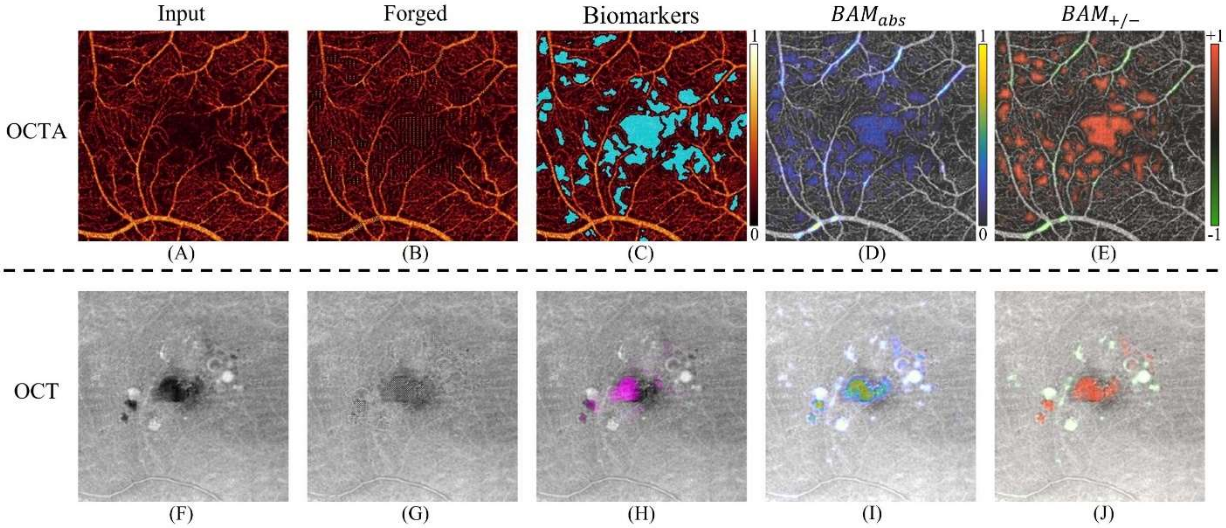


Figure 6.9: BAMs generated for two DR classifiers trained based only on the OCTA and OCT scans, respectively. (A) Superficial vascular complex *en face* maximum projection of OCTA. (B) The forged main generator output. (C) Segmented non-perfusion area (NPA) based on a previously reported deep learning method [180]. (D) The BAM_{abs} is the absolute difference between (B) and (A) after Gaussian filtering (Eq. 17). The highlighted areas are similar to the segmented NPA in (C). (E) The $BAM_{+/-}$ is the (non-absolute) differences between (B) and (A) after Gaussian filtering (Eq. 18) Except to the NPAs highlighted by red (positive values), parts of the vessels of high intensities were highlighted by green (negative values). (F) *En face* mean projection over the inner retina of OCT. (G) The forged main generator output. Hypo-reflective fluids and hyperreflective spots in (F) were both changed typical reflectivity values. (H) The inner mean projection of the segmented fluid based on a previously reported deep learning method [181]. (I) The BAM_{abs} is the absolute difference between (G) and (F) after Gaussian filtering (Eq. 17). (J) The $BAM_{+/-}$ is the difference between (G) and (F) after Gaussian filtering (Eq. 18). The red highlighted areas focus on fluids, and green areas focus on abnormal hyperreflective spots.

6.6.5 Ablation experiments

In the proposed framework, both $M_{+/-}$ and L_c losses were used to ensure that the BAM framework only highlighted the classifier-utilized biomarkers. But the use of these two losses also reduced the computational efficiency. To explore its merit, we compared BAMs generated from our proposed framework and its three variations. The first variation was trained only based on H_-, which means no non-referable DR data or the assistant generator were used. The second variation was trained based on H_- and L_c, which means no preserved output was generated. The third variation was trained based on H_- and M_-, which means no assistant generator was used. Except for the BAMs generated from our proposed framework, the BAMs of the three variations all highlighted features not related to DR pathology such as normal microvasculature and large vessels (marked by blue arrows in Fig. 6.10). In addition, the foveal avascular zone could not be distinguished from the pathological NPAs in the $BAM_{+/-}$ generated with these three training variations (Fig. 6.10).

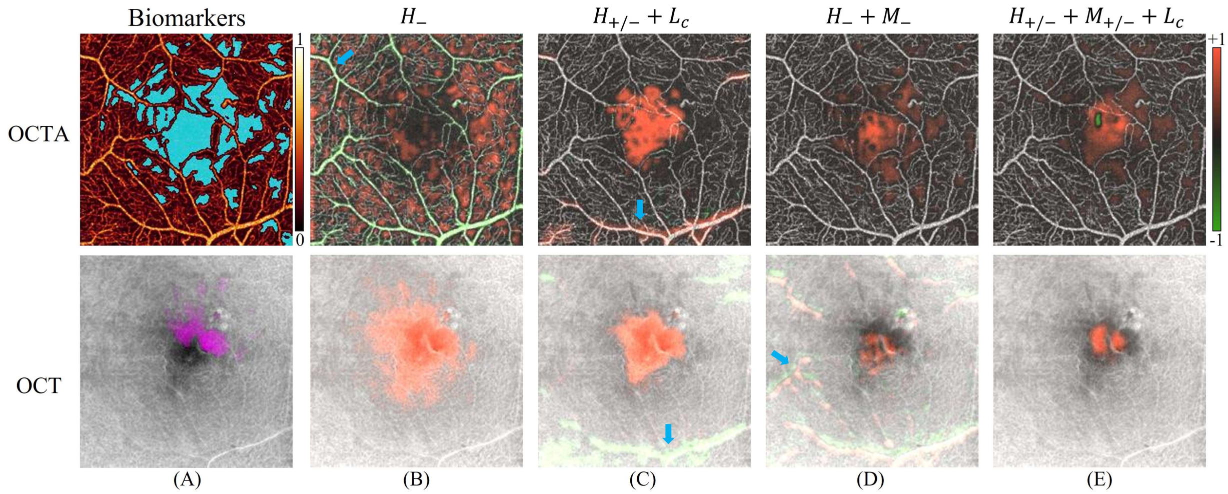


Figure 6.10: BAMs generated in the ablation experiments. Large vessels highlighted by the three variations are marked by blue arrows. (A) Segmented non-perfusion areas and fluids. (B) The $BAM_{+/-}$ generated without non-

referable DR data and assistant generator (loss: H_-). (C) The $BAM_{+/-}$ generated without preserved output (loss: $H_{+/-} + L_c$). (D) The $BAM_{+/-}$ generated without the assistant generator (loss: $H_- + M_-$). (E) The $BAM_{+/-}$ generated based on proposed framework (loss: $H_{+/-} + M_{+/-} + L_c$).

6.7 Discussion

We proposed a BAM generation framework to aid in the interpretation of deep-learning-aided DR diagnosis. The core design concept of our framework is based on the recognition of unique requirements of DR diagnosis compared to non-medical image classification. By designing around this principle, we implemented a framework that enables visualization of specific biomarkers, rather than highlighting shared features between non-referable and referable OCT/OCTA images, which are irrelevant for verifying DR classifier outputs. The framework consists of two U-shaped generators (a main and an assistant generator), and was trained using generative adversarial learning. The BAMs clearly highlight biomarkers utilized by the classifier, which facilitate identification of clinically recognized pathology. The $BAM_{+/-}$ can also distinguish multiple features in the same image. To the best of our knowledge, the proposed BAM generation framework is the first interpretability method specifically designed for deep-learning-aided DR classifiers based on both OCT and OCTA. Based on both qualitative and quantitative comparisons between the BAM and attention maps generated by other methods our framework achieved state-of-the-art performance in providing interpretability to a DR classifier based on OCT and OCTA.

Existing interpretability methods were designed for non-medical image classification and produce attention maps that are not necessarily useful for validating classifier decision making in a medical context like DR diagnosis (Fig. 6.6). A lack of interpretability in medical deep-learning classifiers could lead to ethical and legal challenges, and as such a heuristic method that can provide sufficient interpretability for deep-learning classifiers is now recognized as an urgent need [29-31]. It is difficult to investigate bias if the reasons for the classifier's decisions are unclear [29, 30]. In part to address these concerns, the European Union's General Data Protection Regulation law requires that algorithm decision-making be transparent before it can be utilized for patient care [31, 185].

To evaluate the performance of our BAM, the DR classifier was forced to learn and utilize NPA and fluids which were two important pathologies (clinical biomarkers) for DR diagnosis. Among them, NPA is a biomarker closely correlated to ischemia which is a critical consequence that can be found in the early stage of DR [3, 18, 180]. Fluid is a biomarker closely correlated to DME which is the most common cause of vision loss in DR [3, 181]. In this study, the ground truth NPA and fluid were segmented by previously proposed deep learning methods, respectively [180, 181]. However, these two segmentation methods have no correlation with the interpretability; they were designed to segment all the areas with NPA/fluids no matter whether these clinical biomarkers areas were utilized by the classifier or not. (For example, the fluid segmentation method would identify fluid outside of the macula even though this is subclinical feature for DME.) Only our BAM could provide sufficient interpretability to a DR classifier by accurately highlight the classifier-utilized biomarkers. Clinicians could then verify whether these classifier-utilized biomarkers are clinical biomarkers or not.

In addition to the commonly used attention maps with only positive values (BAM_{abs}), we also generated $BAM_{+/-}$ which separating the positive and negative highlighted biomarkers. There are two major significances of generating $BAM_{+/-}$. Firstly, we could separate the biomarkers that were differently understood by the classifier. Especially the adjacent biomarkers like foveal avascular zone (marked by green in Fig. 6.6(E)) and surrounded pathological NPAs (marked by red in Fig. 6.6(E)), and hypo- and hyper- reflective spots (marked by red and green in Fig. 6.9(J), respectively). Secondly, with $BAM_{+/-}$, we could better understand how a biomarker was learned and utilized by the classifier. The highlighted areas in Fig. 6.6(E) shown the classifier has learned what the anatomical structure near fovea in a normal eye should be. The highlighted areas shown in Fig. 6.7(C) and 6.7(F) told us the classifier not only learned the intensity of these areas in a negative case (without rectangle artifacts) should be lower (much more green than red), but also learned the intensity in these areas of a negative case should not be even (have both green and red).

Generative adversarial learning has been used in several methods to generate pseudo-healthy (corresponding to forged negative in our study) images [186-189]. However, compared to our BAM framework, the different maps generated by these methods cannot be used to provide interpretability to a DR classifier. Firstly, because these methods were trained based on the ground truth labels using several discriminators, which means their generation results only correlated to the data set and could have no correlation with the classifier. Secondly, they only focused on the detection of each pathology. But a DR classifier may not utilize all the pathologies if only a subset of was sufficient for the diagnostic task. This appears to be the case with our classifier, which largely ignored pathology like hyper-reflective foci or microaneurysms. On the contrary, our framework was trained to only remove the classifier-utilized biomarkers from input positive images by using predicted labels without any discriminator. Therefore, compared to previously proposed pseudo-healthy image generation methods, only our framework could be used to provide adequate interpretability to a DR classifier.

Technically, the cycle-consistency generative adversarial learning strategy was used to train our framework [149]. However, compared to the original architecture and training protocol, our framework had several innovations. Firstly, the two discriminators were replaced by a DR classifier that will be interpreted. This design means the framework had the ability to learn provide a heuristic for interpreting the DR classifier. Secondly, the model selection was based on the loss function of only one generator (the main) since our goal was to highlight all the classifier-utilized biomarkers that only belong to referable DR. This design allowed us to select the main generator with highest performance. Thirdly, the generator output was calculated as the sum of input and Tanh output with zero initialization. This design avoided changes of the biomarkers which were not utilized by the classifier in the beginning of the training.

Though our BAM generation framework was only evaluated on an OCT/OCTA-based DR classifier, it can be easily transferred to interpretability tasks for other disease classifiers. For a binary disease classifier, the main generator always carries inputs to forged negative outputs, and vice versa for the assistant generator. For a single disease classifier which classifies each input to S ($S \geq 3$) severities, overall $S - 1$ BAMs are needed to provide sufficient interpretability to the whole classifier (not just the diagnosis of one severity). Each BAM is generated between two adjacent severities by respectively combing all lower and higher severities as one class. Alternatively, for a multiple disease classifier (e.g. a system that diagnoses DR and age-related macular degeneration) a BAM for each disease can be generated between the selected disease and normal class.

In addition to providing interpretability to disease classifiers, our BAM could also facilitate identification of new biomarkers and assessment of pharmacological impact via medical imaging. For example, if a disease classifier were trained on an imaging modality in which the disease is not well characterized the generated BAM could indicate features that should be explored. In drug development, for a classifier trained to classify the cases before and after the treatment, the generated BAM could highlight all the changes caused by the new drug.

Unlike other interpretability methods, our BAM generation framework uses deep learning networks to interpret another deep learning network, which leads to its own questions about interpretability. For networks designed for classification or segmentation, the interpretability issue can be described as how the classification or segmentation results are acquired. In medical image analysis, the concern for interpretability can be further described as whether the clinically meaningful biomarkers are used by the network to make decisions. However, the training target of our framework is generating an output which can be classified as negative by the trained classifier from a positive input.

The interpretability issue in this context can be described as asking how the output is generated to achieve the desired classification, and asking which biomarkers correlate with classifier decision making. The BAM apparently improved interpretability by highlighting the biomarkers that were involved in changing the output of the trained classifier. Therefore, our BAM generation framework is self-interpreted and can be used to provide interpretability to deep-learning-aided disease classifiers.

Though our BAM could accurately highlight the classifier-utilized biomarkers and was sensitive to the potential interpretability changes, the anatomical structures of the forged outputs still look different from real data. In a future study, we will modify our BAM to not only highlight the classifier-utilized biomarkers but also generate the forged output with similar anatomical structure of the real data.

6.8 Conclusion

We proposed a BAM generation framework which can be used to provide interpretation of deep-learning-aided DR classifier. The BAMs demonstrated here accurately highlighted different classifier-utilized biomarkers at high resolution, which enable quick review by image graders to verify whether clinically meaningful biomarkers were used by the classifier. Our BAM generation framework can improve the clinical acceptability and real-world applications for deep-learning-aided DR classifiers.

7 Future Studies

In this dissertation, we proposed 5 systems for automated DR diagnosis based on OCT and OCTA using deep learning techniques. The first system achieved competitive performance for retinal layer segmentation at the ONH using a customized U-Net and a multi-weight graph search algorithm [32]. The second 2D multi-level DR classification system, compared with current deep learning architecture, achieved the highest DR classification performances by using the proposed *DcardNet* architecture [33]. The third 3D multi-level DR classification system achieved specialist-level performance in the detection of both referable and vision threatening DR [34]. The fourth multi-eye-disease detection system was the first deep learning classifier which could diagnose DR, AMD and glaucoma at the same time [35]. The final interpretability system was the first interpretability method specifically designed for DR classifiers and provided clinically meaningful interpretability to DR classifiers [36].

To improve the performance of our deep-learning-aided DR diagnostic system, future studies will be done mainly on the data set and network architecture aspects.

Data set. Each neural network in the DR diagnostic system will be redeveloped on larger data sets to improve generalization. The current performance of this system was limited by the depth and breadth of the development data set. To enhance the applicability of the system for complex real-world clinical scenarios, the new data set should be improved in three ways. Firstly, to generalize the system to inputs scanned on different OCT devices, the data set should include the OCT volumes scanned on other devices with different scanning protocol. Secondly, to empower the system to differentiate DR from eye diseases not currently included, the data set should include OCT volumes scanned from patients with other eye diseases. Finally, to enrich the DR-related information that can be learned by the system, the data set should include OCT volumes scanned on wider field of view which includes other regions of the retina. One of the big challenges with contemporary AI approaches lies in bridging the gap between working with small, curated datasets and functioning effectively in real-world clinical practice. Improving the performance of each neural network based on deeper and wider data sets could enable the DR diagnostic system to work on different clinical settings around the world.

Network architecture. Given the rapid evolution of AI, keeping the proposed system up to date with the latest techniques is crucial for preserving its relevance. The core network of each module will be refined with state-of-the-art AI-techniques, respectively. First, the customized U-Net in the preprocessing module will be modified based on attention mechanism which is widely used in transformer networks [190]. The convolutional layer can only capture the local information of the input and may ignore some global correlations. By using the attention mechanism, the modified U-Net can encode and decode the inputs based on both local features and global correlations between different regions (e.g., some pathologies like edema are critical at the macula but could be ignored elsewhere). Second, in addition to the attention mechanism, the classifiers in the modules (ii), (iii), and (iv) will be modified based on both Bayesian deep learning and position embedding [190, 191]. By using Bayesian deep learning techniques, a confidence score can be generated for each diagnostic result. This confidence score can improve the applicability of the DR diagnostic system in the clinics. The human verification can mainly focus on the diagnostic results with low confidence scores, which can improve both time and cost efficiencies for the whole DR screening process. By using position embedding, the classifiers can be sensitive to the location of similar features. In a pure CNN, similar features in different positions are encoded in the same way, which may negatively impact both performance and interpretability. With the position embedding used in the transformer, features from different positions can be differentiated based on weights added on different positions of the inputs. Third, the BAM generation framework in the last module will be modified to generate the forged output with similar anatomical structure as the real data (rather than the artificial disruptions such as dots that are produced by the current iteration). Except for the transformer, state-of-the-art generative AI techniques like variational autoencoder and stable diffusion will be used in the modified BAM generation framework [192-194].

8. Conclusion

This dissertation presented 5 deep learning systems to achieve automated and interpretable DR diagnosis based on OCT and OCTA. The combination of the proposed systems achieved state-of-the-art performance in multi-level DR diagnosis and provide clinical interpretability. The 5 systems respectively are: (i) a retinal layer segmentation system based on U-Net and multi-weight graph search [32], (ii) a 2D multi-level DR classification system which only needs *en face* projections of OCT and OCTA as inputs, (iii) a 3D multi-level DR classification system which can use the original OCT and OCTA volumes as inputs (iv) a multi-eye-disease detection system focused on the identification of DR from healthy eyes and other eye diseases, , and (v) an interpretability system which can highlight the biomarkers utilized in the systems (ii), (iii), and (iv) on an attention map for each input. In real-world practice, the proposed deep learning systems could reduce vision loss and lower clinical burden by providing time-efficient, cost-efficient, and clinically explainable DR diagnosis.

9. Reference

[1] C. P. Wilkinson, *et al.*, “Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales,” *Ophthalmology*, vol. 110, no. 9, pp. 1677-1682, Sep. 2003.

[2] National Eye Institute. Eye health data and statistics. Accessed September 1, 2019.

[3] T. Y. Wong, *et al.*, “Guidelines on diabetic eye care: the international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings,” *Ophthalmology*, vol. 125, no. 10, pp. 1608-1622, Oct. 2018.

[4] C. J. Flaxel, *et al.*, “Diabetic retinopathy preferred practice pattern®,” *Ophthalmology*, vol. 127, no. 1, pp. 66-145, Jan. 2020.

[5] D. A. Antonetti, R. Klein, and T. W. Gardner, “Diabetic retinopathy,” *N. Engl. J. Med.*, vol. 366, pp. 1227–1239, 2012.

[6] E. A. Lundeen, *et al.*, “Prevalence of Diabetic Retinopathy in the US in 2021.” *JAMA ophthalmology*, vol 141, no. 8, pp. 747-754, 2023.

[7] R. Gargeya and T. Leng, “Automated identification of diabetic retinopathy using deep learning,” *Ophthalmology*, vol. 124, no. 7, pp. 962-969, 2017.

[8] M. D. Abràmoff, *et al.*, “Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning,” *Investig. Ophthalmol. Vis. Sci.*, vol. 57, no. 13, pp. 5200-5206, 2016.

[9] V. Gulshan, *et al.*, “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *JAMA*, vol. 316, no. 22, pp. 2402-2410, 2016.

[10] R. Ghosh, *et al.*, “Automatic detection and classification of diabetic retinopathy stages using CNN,” in *Proc. 4th SPIN*, 2017, pp. 550-554.

[11] Early Treatment Diabetic Retinopathy Study Research Group, "Fundus photographic risk factors for progression of diabetic retinopathy: ETDRS report number 12," *Ophthalmology*, vol. 98, no. 5, pp. 823-833, May. 1991.

[12] D. Ophthalmoscopy, and E. Levels, "International clinical diabetic retinopathy disease severity scale detailed table," 2002.

[13] D. Huang, *et al.*, "Optical coherence tomography," *Science*, vol. 254, no. 5035, pp. 1178–1181, Nov. 1991.

[14] R. Lee, T. Y. Wong, and C. Sabanayagam, "Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss," *Eye Vis.*, vol. 2, no. 1, pp. 1–25, Sep. 2015.

[15] G. Prescott, *et al.*, "Improving the cost-effectiveness of photographic screening for diabetic macular oedema: a prospective, multi-centre, UK study," *Br. J. Ophthalmol.*, vol. 98, no. 8, pp. 1042–1049, Jul. 2014.

[16] L. Guariguata, *et al.*, Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes Res Clin Pract.*, vol. 103, no. 2, pp. 137-149, 2014.

[17] T. S. Hwang, *et al.*, "Visualization of 3 Distinct Retinal Plexuses by Projection-Resolved Optical Coherence Tomography Angiography in Diabetic Retinopathy," *JAMA ophthalmol.*, vol. 134, no. 12, pp. 1411-1419, 2016.

[18] M. Zhang, *et al.*, "Automated Quantification of Nonperfusion in Three Retinal Plexuses Using Projection-Resolved Optical Coherence Tomography Angiography in Diabetic Retinopathy," *Investig. Ophthalmol. Vis. Sci.*, vol. 57, no. 13, pp. 5101-5106, 2016.

[19] T. S. Hwang, *et al.*, "Automated quantification of capillary nonperfusion using optical coherence tomography angiography in diabetic retinopathy," *JAMA ophthalmol.*, vol. 134, no. 4, pp. 367-373, 2016.

[20] T. S. Hwang, *et al.*, "Optical coherence tomography angiography features of diabetic retinopathy," *Retina*, vol. 35, no. 11, pp. 2371, 2015.

[21] S. Makita, *et al.*, “Optical coherence angiography,” *Opt. Express*, vol. 14, no. 17, pp. 7821-7840, Aug. 2006.

[22] L. An, and R. K. Wang, “In vivo volumetric imaging of vascular perfusion within human retina and choroids with optical micro-angiography,” *Opt. Express*, vol. 16, no. 15, pp. 11438-11452, Jul. 2008.

[23] Y. Jia, *et al.*, “Split-spectrum amplitude-decorrelation angiography with optical coherence tomography,” *Opt. Express*, vol. 20, no. 4, pp. 4710–4725, Feb. 2012.

[24] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436-444, May. 2015.

[25] G. Litjens, *et al.*, “A survey on deep learning in medical image analysis,” *Med. Image Anal.*, vol. 42, pp. 60-88, Dec. 2017.

[26] D. Shen, G. Wu, and H. I. Suk, “Deep learning in medical image analysis,” *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221-248, Jun. 2017.

[27] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy, “A review on deep learning in medical image analysis,” *Int. J. Multimed. Inf. Retr.*, vol. 11, no. 1, pp. 19-38, 2022.

[28] T. T. Hormel, *et al.*, “Artificial intelligence in OCT angiography,” *Prog. Retin. Eye Res.*, vol. 85, pp. 100965, Nov. 2021.

[29] J. He, *et al.*, “The practical implementation of artificial intelligence technologies in medicine,” *Nat. Med.*, vol. 25, no. 1, pp. 30-36, Jan. 2019.

[30] S. Gerke S, T. Minssen, and G. Cohen, “Ethical and legal challenges of artificial intelligence-driven healthcare,” *Artificial intelligence in healthcare*, Academic Press, pp. 295-336, 2020.

[31] Z. Salahuddin, *et al.*, “Transparency of deep neural networks for medical image analysis: A review of interpretability methods,” *Comput. Biol. Med.*, vol. 140, pp. 105111, 2022.

[32] P. Zang, *et al.*, Automated segmentation of peripapillary retinal boundaries in OCT combining a convolutional neural network and a multi-weights graph search. *Biomed Opt Express*, vol. 10, no. 8, pp. 4340-4352, Aug. 2019.

[33] P. Zang, *et al.*, DcardNet: Diabetic Retinopathy Classification at Multiple Levels Based on Structural and Angiographic Optical Coherence Tomography. *IEEE Trans Biomed Eng.*, vol. 68, no. 6, pp. 1859-1870, Jun 2021.

[34] P. Zang, *et al.*, A Diabetic Retinopathy Classification Framework based on Deep-learning Analysis of OCT Angiography. *Transl Vis Sci Technol.*, vol. 11, no. 7, pp. 10-10, Jul 2022.

[35] P. Zang, *et al.*, Deep-learning-aided Diagnosis of DR, AMD, and Glaucoma based on Structural and Angiographic Optical Coherence Tomography. *Ophthalmol Sci.*, vol. 3, no. 1, pp. 100245, Nov. 2022.

[36] P. Zang, *et al.*, Interpretable Diabetic Retinopathy Diagnosis based on Biomarker Activation Map. *IEEE Trans Biomed Eng.*, Jun 2023.

[37] Y. C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C. Y. Cheng, “Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis,” *Ophthalmology*, vol. 121, no. 11, pp. 2081–2090, 2014.

[38] R. N. Weinreb, T. Aung, and F. A. Medeiros, “The pathophysiology and treatment of glaucoma: a review,” *JAMA*, vol. 311, no. 18, pp. 1901–1911, 2014.

[39] Y. Jia, *et al.*, “Optical Coherence Tomography Angiography of Optic Disc Perfusion in Glaucoma,” *Ophthalmology*, vol. 121, no. 7, pp. 1322–1332, 2014.

[40] L. Liu, *et al.*, “Optical coherence tomography angiography of the peripapillary retina in glaucoma,” *JAMA Ophthalmol.*, vol. 133, no. 9, pp. 1045–1052, 2015.

[41] J. P. Campbell, *et al.*, “Detailed vascular anatomy of the human retina by projection-resolved optical coherence tomography angiography,” *Sci. Rep.*, vol. 7, no. 1, pp. 42201, 2017.

[42] M. K. Garvin, M. D. Abramoff, R. Kardon, S. R. Russell, X. Wu, and M. Sonka, “Intraretinal layer segmentation of macular optical coherence tomography images using optimal 3-D graph search,” *IEEE Trans. Med. Imaging*, vol. 27, no. 10, pp. 1495–1505, 2008.

[43] Z. Hu, M. Niemeijer, K. Lee, M. D. Abramoff, M. Sonka, and M. K. Garvin, “Automated segmentation of the optic disc margin in 3-D optical coherence tomography images using a graph-theoretic approach,” *Proc. SPIE*, 7262, 72620U, 2009.

[44] B. J. Antony, *et al.*, “Automated 3D segmentation of intraretinal layers from optic nerve head optical coherence tomography images,” *Proc. SPIE*, 7626, 76260U, 2010.

[45] K. Lee, M. Niemeijer, M. K. Garvin, Y. H. Kwon, M. Sonka, and M. D. Abramoff, “Segmentation of the optic disc in 3-D OCT scans of the optic nerve head,” *IEEE Trans. Med. Imaging*, vol. 29, no. 1, pp. 159–168, 2010.

[46] M. S. Miri, *et al.*, “Multimodal Segmentation of Optic Disc and Cup From SD-OCT and Color Fundus Photographs Using a Machine-Learning Graph-Based Approach,” *IEEE Trans. Med. Imaging*, vol. 34, no. 9, pp. 1854–1866, 2015.

[47] Z. Hu, C. A. Girkin, A. Hariri, and S. R. Sadda, “Three-dimensional choroidal segmentation in spectral OCT volumes using optic disc prior information,” *Proc. SPIE*, 9697, 96971S, 2016.

[48] P. Zang, S. *et al.*, “Automated boundary detection of the optic disc and layer segmentation of the peripapillary retina in volumetric structural and angiographic optical coherence tomography,” *Biomed. Opt. Express*, vol. 8, no. 3, pp. 1306–1318, 2017.

[49] E. Gao, *et al.*, “Graph Search–Active Appearance Model based Automated Segmentation of Retinal Layers for Optic Nerve Head Centered OCT Images,” in *SPIE Medical Imaging*, pp. 101331Q–101331Q, 2017.

[50] K. Yu, F. Shi, E. Gao, W. Zhu, H. Chen, and X. Chen, “Shared-hole graph search with adaptive constraints for 3D optic nerve head optical coherence tomography image segmentation,” *Biomed. Opt. Express*, vol. 9, no. 3, pp. 962–983, 2018.

[51] S. Apostolopoulos, S. De Zanet, C. Ciller, S. Wolf, and R. Sznitman, “Pathological OCT Retinal Layer Segmentation using Branch Residual U-shape Networks,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 294–301, Sep. 2017.

[52] L. Fang, D. Cunefare, C. Wang, R. H. Guymer, S. Li, and S. Farsiu, “Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search,” *Biomed. Opt. Express*, vol. 8, no. 5, pp. 2732–2744, 2017.

[53] S. K. Devalla, *et al.*, “DRUNET: a dilated-residual U-Net deep learning network to segment optic nerve head tissues in optical coherence tomography images,” *Biomed. Opt. Express*, vol. 9, no. 7, pp. 3244–3265, 2018.

[54] J. Kugelman, D. Alonso-Caneiro, S. A. Read, S. J. Vincent, and M. J. Collins, “Automatic segmentation of OCT retinal boundaries using recurrent neural networks and graph search,” *Biomed. Opt. Express*, vol. 9, no. 11, pp. 5759–5777, 2018.

[55] A. Camino, *et al.*, “Deep learning for the segmentation of preserved photoreceptors on en face optical coherence tomography in two inherited retinal diseases,” *Biomed. Opt. Express*, vol. 9, no. 7, pp. 3092–3105, 2018.

[56] Y. Guo, A. Camino, J. Wang, D. Huang, T. S. Hwang, and Y. Jia, “MEDnet, a neural network for automated detection of avascular area in OCT angiography,” *Biomed. Opt. Express*, vol. 9, no. 11, pp. 5147–5158, 2018.

[57] S. S. Gao, G. Liu, D. Huang, and Y. Jia, “Optimization of the split-spectrum amplitude-decorrelation angiography algorithm on a spectral optical coherence tomography system,” *Opt. Lett.*, vol. 40, no. 10, pp. 2305–2308, 2015.

[58] M. F. Kraus, *et al.*, “Quantitative 3D-OCT motion correction with tilt and illumination correction, robust similarity measure and regularization,” *Biomed. Opt. Express*, vol. 5, no. 8, pp. 2591–2613, 2014.

[59] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[60] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Proceedings of Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Oct. 2015.

[61] V. K. Fisher Yu, “Multi-scale context aggregation by dilated convolutions,” arXiv:1511.07122 [cs.CV] (2016).

[62] Q. Zhang, Z. Cui, X. Niu, S. Geng, and Y. Qiao, “Image segmentation with pyramid dilated convolution based on ResNet and U-Net,” in *Neural Information Processing*, 364–372, 2017.

[63] S. Ioffe and C. Szegedy, “Batch normalization: accelerating deep network training by reducing internal covariate shift,” in *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, pp. 448–456, 2015.

[64] T. U. Djork-Arné Clevert and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (elus),” arXiv:1511.07289 [cs.LG] (2015).

[65] S. Ruder, “An overview of gradient descent optimization algorithms,” arXiv:1609.04747 [cs.LG] (2016).

[66] A. S. Reis, *et al.*, “Influence of clinically invisible, but optical coherence tomography detected, optic disc margin anatomy on neuroretinal rim evaluation,” *Invest. Ophthalmol. Vis. Sci.*, vol. 53, no. 4, pp. 1852–1860, 2012.

[67] T. S. Hwang, *et al.*, “Visualization of 3 Distinct Retinal Plexuses by Projection-Resolved Optical Coherence Tomography Angiography in Diabetic Retinopathy,” *JAMA Ophthalmol.*, vol. 134, no. 12, pp. 1411–1419, 2016.

[68] M. Zhang, *et al.*, “Projection-resolved optical coherence tomographic angiography,” *Biomed. Opt. Express*, vol. 7, no. 3, pp. 816–828, 2016.

[69] J. Wang, M. Zhang, T. S. Hwang, S. T. Bailey, D. Huang, D. J. Wilson, and Y. Jia, "Reflectance-based projection-resolved optical coherence tomography," *Biomed. Opt. Express*, vol. 8, no. 3, pp. 1536–1548, 2017.

[70] T. T. Hormel, J. Wang, S. T. Bailey, T. S. Hwang, D. Huang, and Y. Jia, "Maximum value projection produces better en face OCT angiograms than mean value projection," *Biomed. Opt. Express*, vol. 9, no. 12, pp. 6412–6424, 2018.

[71] B. Zhou *et al.*, "Learning deep features for discriminative localization," in *Proc. CVPR*, 2016, pp. 2921-2929.

[72] H. S. Sandhu, *et al.*, "Automated diagnosis and grading of diabetic retinopathy using optical coherence tomography," *Investig. Ophthalmol. Vis. Sci.*, vol. 59, no. 7, pp. 3155-3160, 2018.

[73] H. S. Sandhu, *et al.*, "Automated diabetic retinopathy detection using optical coherence tomography angiography: a pilot study," *Brit. J. Ophthalmol.*, vol. 102, no. 11, pp. 1564-1569, 2018.

[74] M. Alam, *et al.*, "Quantitative optical coherence tomography angiography features for objective classification and staging of diabetic retinopathy," *Retina*, vol. 40, no. 2, pp. 322-332, 2020.

[75] Heisler M, *et al.*, "Ensemble Deep Learning for Diabetic Retinopathy Detection Using Optical Coherence Tomography Angiography," *Transl. Vis. Sci. Technol.*, vol. 9, no. 2, pp. 20-20, 2020.

[76] Le D, *et al.*, "Transfer learning for automated OCTA detection of diabetic retinopathy," *Transl. Vis. Sci. Technol.*, vol. 9, no. 2, pp. 35-35, 2020.

[77] K. He, *et al.*, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770-778.

[78] M. Zhang, *et al.*, "Advanced image processing for optical coherence tomographic angiography of macular diseases," *Biomed. Opt. Express*, vol. 6, no. 12, pp. 4661–4675, 2015.

[79] N. Srivastava, *et al.*, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929-1958, 2014.

[80] G. Huang, *et al.*, “Densely connected convolutional networks,” in *Proc. CVPR*, 2017, pp. 4700-4708.

[81] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proc. 27th ICML*, 2010, pp. 807-814.

[82] X. Glorot, *et al.*, “Deep sparse rectifier neural networks,” in *Proc. 14th AISTATS*, 2011, pp. 315-323.

[83] C. Szegedy, *et al.*, “Rethinking the inception architecture for computer vision,” in *Proc. CVPR*, 2016, pp. 2818-2826.

[84] G. Pereyra, *et al.*, “Regularizing neural networks by penalizing confident output distributions,” arXiv preprint arXiv:1701.06548, 2017.

[85] I. Loshchilov and F. Hutter, “SGDR: stochastic gradient descent with warm restarts,” arXiv preprint arXiv:1608.03983, 2016.

[86] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” arXiv preprint arXiv:1905.11946, 2019.

[87] K. Simonyan, and A. Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556, 2014.

[88] K. He, *et al.*, “Identity mappings in deep residual networks,” arXiv:1603.05027, 2016.

[89] C. Szegedy, *et al.*, “Inception-v4, inception-resnet and the impact of residual connections on learning.” in *AAAI*, 2017, pp. 4278–4284.

[90] E. Swanson and D. Huang. “Ophthalmic OCT reaches $1 billion per year.” *Retin. Physician*, vol. 8, no. 4, pp. 58-59, 2011.

[91] I. Ghorbel, *et al.*, “Automated segmentation of macular layers in OCT images and quantitative evaluation of performances.” *Pattern Recognit.*, vol. 44, no. 8, pp. 1590-1603, 2011.

[92] P. P. Srinivasan, *et al.*, “Automatic segmentation of up to ten layer boundaries in SD-OCT images of the mouse retina with and without missing layers due to pathology.” *Biomed. Opt. Express*, vol. 5, no. 2, pp. 348-365, 2014.

[93] Z. Gao, *et al.*, “Automated layer segmentation of macular OCT images via graph-based SLIC superpixels and manifold ranking approach.” *Comput. Med. Imag. Grap.*, vol. 55, pp. 42-53, 2017.

[94] S. J. Chiu, *et al.*, “Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema.” *Biomed. Opt. Express*, vol. 6, no. 4, pp. 1172-1194, 2015.

[95] Y. Guo, *et al.*, “Automated segmentation of retinal layer boundaries and capillary plexuses in wide-field optical coherence tomographic angiography.” *Biomed. Opt. Express*, vol. 9, no. 9, pp. 4429-4442, 2018.

[96] C. S. Lee, *et al.*, “Deep-learning based, automated segmentation of macular edema in optical coherence tomography.” *Biomed. Opt. Express*, vol. 8, no. 7, pp. 3440-3448, 2017.

[97] A. C. Onishi, *et al.*, “Importance of considering the middle capillary plexus on OCT angiography in diabetic retinopathy.” *Investig. Ophthalmol. Vis. Sci.*, vol. 59, no. 5, pp. 2167-2176, 2018.

[98] E. Borrelli, *et al.*, “In vivo rotational three-dimensional OCTA analysis of microaneurysms in the human diabetic retina.” *Sci. Rep.*, vol. 9, no. 1, pp. 1-8, 2019.

[99] G. Virgili, *et al.*, “Optical coherence tomography (OCT) for detection of macular oedema in patients with diabetic retinopathy,” *Cochrane Database Syst Rev.*, vol. 1, pp. CD008081, 2015.

[100] J. Kinyoun, *et al.*, “The ETDRS Research Group. Detection of diabetic macular edema: ophthalmoscopy versus photography—Early Treatment Diabetic Retinopathy Study Report Number 5,” *Ophthalmology*, vol. 96, no. 6, pp. 746–750, 1989.

[101] K. V. Bhavsar, and M. L. Subramanian, “Risk factors for progression of subclinical diabetic macular oedema,” *Br J Ophthalmol.*, vol. 95, no. 5, pp. 671–674, 2011.

[102] N. M. Bressler, *et al.*, “Diabetic Retinopathy Clinical Research Network. Observational study of subclinical diabetic macular edema,” *Eye (Lond)*, vol. 26, no. 6, pp. 833–840, 2012.

[103] D. J. Browning, and C. M. Fraser, “The predictive value of patient and eye characteristics on the course of subclinical diabetic macular edema,” *Am J Ophthalmol.*, vol. 145, no. 1, pp. 149–154, 2008.

[104] D. J. Browning, C. M. Fraser, and S. Clark, “The relationship of macular thickness to clinically graded diabetic retinopathy severity in eyes without clinically detected diabetic macular edema,” *Ophthalmology*, vol. 115, no. 3, pp. 533–539, 2008.

[105] S, Ruia, S. Saxena, C. M. Gemmy Cheung, J. S. Gilhotra, and T. Y. Lai, “Spectral domain optical coherence tomography features and classification systems for diabetic macular edema: a review,” *Asia Pac J Ophthalmol (Phila)*, vol. 5, no. 5, pp. 360–367, 2016.

[106] Olson J, *et al.*, “Improving the economic value of photographic screening for optical coherence tomography–detectable macular oedema: a prospective, multicentre, UK study,” *Health Technol Assess*, vol. 17, no. 51, pp. 1–142, 2013.

[107] U. Schmidt-Erfurth, *et al.*, “Guidelines for the management of diabetic macular edema by the European Society of Retina Specialists (EURETINA),” *Ophthalmologica.*, vol. 237, no. 4, pp. 185–222, 2017.

[108] S. H. Paz, *et al.*, “Noncompliance with vision care guidelines in Latinos with type 2 diabetes mellitus: the Los Angeles Latino Eye Study,” *Ophthalmology*, vol. 113, no. 8, pp. 1372-1377, 2006.

[109] Q. You, *et al.*, “Comparison of central macular fluid volume with central subfield thickness in patients with diabetic macular edema using optical coherence tomography angiography,” *JAMA Ophthalmol.*, vol. 139, no. 7, pp. 734–741, 2021.

[110] J. Cohen, “A coefficient of agreement for nominal scales,” *Educ Psychol Meas*, vol. 20, no. 1, pp. 37-46, 1960.

[111] T. T. Hormel, D. Huang, and Y. Jia, “Artifacts and artifact removal in optical coherence tomographic angiography,” *Quant. Imaging Med. Surg.*, vol. 11, no. 3, pp. 1120, 2021.

[112] J. Krause, *et al.*, “Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy,” *Ophthalmology*, vol. 125, no. 8, pp. 1264–1272, 2018.

[113] T. S. Hwang, *et al.*, “Automated quantification of nonperfusion areas in 3 vascular plexuses with optical coherence tomography angiography in eyes of patients with diabetes,” *JAMA Ophthalmol.*, vol. 136, no. 8, pp. 929–936, 2018.

[114] Z. L. Teo, *et al.*, “Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis,” *Ophthalmology*, vol. 128, pp. 1580e1591, 2021.

[115] J. Beagley, L. Guariguata, C. Weil, and A. A. Motala, “Global estimates of undiagnosed diabetes in adults,” *Diabetes Res Clin Pract.*, vol. 103, pp. 150e160, 2015.

[116] C. C. Klaver, *et al.*, “Genetic risk of agerelated maculopathy: population-based familial aggregation study,” *Arch. Ophthalmol.*, vol. 116, pp. 1646e1651, 1998.

[117] P. J. Rosenfeld, *et al.*, “Ranibizumab for neovascular age-related macular degeneration,” *N Engl J Med.*, vol. 355, pp. 1419e1431, 2006.

[118] Y. C. Tham, *et al.*, “Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis,” *Ophthalmology*, vol. 121, pp. 2081e2090, 2014.

[119] P. Mitchell, G. Liew, B. Gopinath, T. Y. Wong, “Age-related macular degeneration,” *Lancet*, vol. 392, pp. 1147e1159, 2018.

[120] M. A. Kass, *et al.*, “The Ocular Hypertension Treatment Study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma,” *Arch Ophthalmol.*, vol. 120, pp. 701e713, 2002.

[121] Y. Jia, *et al.*, “Quantitative optical coherence tomography angiography of vascular abnormalities in the living human eye,” *Proc. Natl. Acad. Sci.*, vol. 112, no. 18, pp. E2395-402, 2015.

[122] M. Adhi, and J. S. Duker, “Optical coherence tomography–current and future applications,” *Curr Opin Ophthalmol.*, vol. 24, pp. 213–221, 2013.

[123] Z. Yehoshua, P. J. Rosenfeld, G. Gregori, and F. Penha, “Spectral domain optical coherence tomography imaging of dry agerelated macular degeneration,” *Ophthalmic Surg. Lasers Imaging Retina*, vol. 41, pp. S6–S14, 2010.

[124] Y. Jia, *et al.*, “Quantitative optical coherence tomography angiography of choroidal neovascularization in age-related macular degeneration,” *Ophthalmology*, vol. 121, no. 7, pp. 1435-44, 2014.

[125] Q. You, *et al.*, “Detection of reduced retinal vessel density in eyes with geographic atrophy secondary to age-related macular degeneration using projection-resolved optical coherence tomography angiography,” *Am. J. Ophthalmol.*, vol. 209, pp. 206-212, 2020.

[126] J. C. Mwanza, *et al.*, “Cirrus Optical Coherence Tomography Normative Database Study Group. Ability of cirrus HD-OCT optic nerve head parameters to discriminate normal from glaucomatous eyes,” *Ophthalmology*, vol. 118, pp. 241–248, 2011.

[127] L. Liu, *et al.*, “Sectorwise visual field simulation using optical coherence tomographic angiography nerve fiber layer plexus measurements in glaucoma,” *Am. J. Ophthalmol.*, vol. 212, pp. 57-68, 2020.

[128] Centers for Medicare & Medicaid Services. Medicare Provider Utilization and Payment Data: Physician and Other Supplier.

[129] T. T. Hormel, *et al.*, “Plexus-specific retinal vascular anatomy and pathologies as seen by projection-resolved optical coherence tomographic angiography,” *Prog. Retin. Eye Res.*, vol. 80, pp. 100878, 2021.

[130] R. Hazin, M. K. Barazi, and M. Summerfield, “Challenges to establishing nationwide diabetic retinopathy screening programs,” *Curr Opin Ophthalmol.*, vol. 22, no. 3 pp. 174-179, 2011.

[131] F. G. Venhuizen, *et al.*, "Automated staging of age-related macular degeneration using optical coherence tomography," *Investig. Ophthalmol. Vis. Sci.*, vol. 58, no. 4, pp. 2318-2328, 2017.

[132] Y. Peng, *et al.*, "DeepSeeNet: a deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs," *Ophthalmology*, vol. 126, no. 4, pp. 565-575, 2019.

[133] A. C. Thompson, A. A. Jammal, and F. A. Medeiros, "A review of deep learning for screening, diagnosis, and detection of glaucoma progression," *Transl. Vis. Sci. Technol.*, vol. 9, no. 2, pp. 42-42, 2020.

[134] A. C. Thompson, *et al.*, "Assessment of a segmentation-free deep learning algorithm for diagnosing glaucoma from optical coherence tomography scans," *JAMA Ophthalmol.*, vol. 138, no. 4, pp. 333-339, 2020.

[135] R. Asaoka, *et al.*, "Using deep learning and transfer learning to accurately diagnose early-onset glaucoma from macular optical coherence tomography images," *Am. J. Ophthalmol.*, vol. 198, pp. 136-145, 2019.

[136] E. B. Mariottoni, et al., "An objective structural and functional reference standard in glaucoma," *Sci. Rep.*, vol. 11, no. 1, pp. 1-10, 2021.

[137] M. F. Marmor, and J. G. Ravin, "Fluorescein angiography: insight and serendipity a half century ago," *Arch. Ophthalmol.*, vol. 129, pp. 943e948, 2011.

[138] H. L. Takusagawa, *et al.*, "Projection-resolved optical coherence tomography angiography of macular retinal circulation in glaucoma," *Ophthalmology*, vol. 124, no. 11, pp. 1589-1599, 2017.

[139] C. L. Chen, *et al.*, "Peripapillary retinal nerve fiber layer vascular microcirculation in eyes with glaucoma and single-hemifield visual field loss," *JAMA Ophthalmol.*, vol. 135, no. 5, pp. 461-468, 2017.

[140] R. N. Weinreb, and P. T. Khaw, "Primary open-angle glaucoma," *Lancet*, vol. 363, no. 9422, pp. 1711-1720, 2004.

[141] O. Daanouni, B. Cherradi, and A. Tmiri, “Automatic detection of diabetic retinopathy using custom CNN and grad-cam,” in *Proc. ICACIn*, 2021. pp. 15-26.

[142] J. D. Fauw *et al.*, “Clinically applicable deep learning for diagnosis and referral in retinal disease,” *Nat. Med.*, vol. 24, no. 9, pp. 1342-1350, Aug. 2018.

[143] Q. Zhang, and S. C. Zhu, “Visual interpretability for deep learning: a survey,” *Front. Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 27-39, Jan. 2018.

[144] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, “Machine learning interpretability: A survey on methods and metrics,” *Electronics*, vol. 8, no. 8, pp. 832, Jul. 2019.

[145] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable AI: a review of machine learning interpretability methods,” *Entropy*, vol. 23, no. 1, pp. 18, Dec. 2020.

[146] I. Goodfellow *et al.*, “Generative adversarial nets,” *Commun. ACM*, vol. 63, no. 11, pp. 139-144, Nov. 2020.

[147] M. Mirza, and S. Osindero, “Conditional generative adversarial nets,” arXiv preprint arXiv:1411.1784, 2014.

[148] P. Isola *et al.*, “Image-to-image translation with conditional adversarial networks,” in *Proc. IEEE CVPR*, 2017, pp. 1125-1134.

[149] J. Y. Zhu *et al.*, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proc. of the IEEE ICCV*, 2017, pp. 2223-2232.

[150] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *Proc. ICML PMLR*, 2017, pp. 3319–3328.

[151] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. “Not just a black box: Learning important features through propagating activation differences,” arXiv preprint arXiv:1605.01713, 2016.

[152] D. Smilkov, N. Thorat, B. Kim, F. Vi´egas, and M. Wattenberg, “Smoothgrad: removing noise by adding noise,” arXiv preprint arXiv:1706.03825, 2017.

[153] S. Srinivas, and F. Fleuret, “Full-gradient representation for neural network visualization,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 4126–4135.

[154] H. Chefer, S. Gur, and L. Wolf, “Transformer interpretability beyond attention visualization,” in *Proc. IEEE CVPR*. 2021, pp. 782-791.

[155] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE ICCV*, 2017, pp. 618-626.

[156] A. Chattopadhay, A. Sarkar, P. Howlader, V. N. Balasubramanian, “Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks,” in *Proc IEEE Winter Conf. on Appl. Comput. Vis.*, 2018, pp. 839-847.

[157] K. Li, Z. Wu, K. Peng, J. Ernst, and Y. Fu, “Tell me where to look: guided attention inference network,” in *Proc. IEEE CVPR*, 2018, pp. 9215-9223.

[158] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K. M¨uller, “Explaining nonlinear classification decisions with deep taylor decomposition,” *Pattern Recognit.*, vol. 65, pp. 211–222, May. 2017.

[159] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. M¨uller, and W. Samek, “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation,” *PloS one*, vol. 10, no. 7, pp. e0130140, Jul. 2015.

[160] W. Nam, S. Gur, J. Choi, L. Wolf, and S. Lee. “Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks,” arXiv preprint arXiv:1904.00605, 2019.

[161] S. Gur, A. Ali, and L. Wolf. “Visualization of supervised and self-supervised neural networks via attribution guided factorization,” in *Proc. AAAI*, 2021, pp. 11545-11554.

[162] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning important features through propagating activation differences,” in *Proc. ICML PMLR*, 2017, pp. 3145–3153.

[163] S. M. Lundberg, and S. Lee, “A unified approach to interpreting model predictions,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.

[164] J. Gu, Y. Yang, and V. Tresp, “Understanding individual decisions of cnns via contrastive backpropagation,” in *Proc. ACCV*, 2018, pp. 119–134.

[165] B. K. Iwana, R. Kuroki, and S. Uchida, “Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation,” arXiv preprint arXiv:1908.04351, 2019.

[166] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, “Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning,” *Nat. Biotechnol.*, vol. 33, no. 8, pp. 831–838, Jul. 2015.

[167] M. D. Zeiler, and R. Fergus, “Visualizing and understanding convolutional networks,” in *Proc. ECCV*, 2014, pp. 818–833.

[168] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Object detectors emerge in deep scene cnns,” arXiv preprint arXiv:1412.6856, 2014.

[169] J. Zhou, and O. G. Troyanskaya, “Predicting effects of noncoding variants with deep learning–based sequence model,” *Nat. Methods*, vol. 12, no. 10, pp. 931–934, Aug. 2015.

[170] A. Mahendran, and A. Vedaldi, “Visualizing deep convolutional neural networks using natural pre-images,” *Int. J. Comput. Vis.*, vol. 120, no. 3, pp. 233–255, May. 2016.

[171] C. Olah, A. Mordvintsev, and L. Schubert, “Feature visualization,” *Distill*, vol. 2, no. 11, pp. e7, Nov. 2017.

[172] P. Dabkowski, and Y. Gal, “Real time image saliency for black box classifiers,” in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 6970–6979, 2017.

[173] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” arXiv preprint arXiv:1312.6034, 2013.

[174] B. Mittelstadt, C. Russell, and S. Wachter, “Explaining explanations in ai,” in *Proc. ACM FAccT*, pp. 279–288, 2019.

[175] B. Zhou, D. Bau, A. Oliva, and A. Torralba, “Interpreting deep visual representations via network dissection,” *IEEE PAMI*, vol. 41, no. 9, pp. 2131-2145, Jul. 2018.

[176] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, “Visualizing higher-layer features of a deep network,” *University of Montreal*, vol. 1341, no. 3, pp. 1, Jan. 2009.

[177] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, “Top-down neural attention by excitation backprop,” *Int. J. Comput. Vis.*, vol. 126, no. 10, pp. 1084–1102, 2018.

[178] R. Fong, M. Patrick, and A. Vedaldi, “Understanding deep networks via extremal perturbations and smooth masks,” in *Proc. IEEE ICCV*, pp. 2950–2958, 2019.

[179] R. C. Fong, and A. Vedaldi, “Interpretable explanations of black boxes by meaningful perturbation,” in *Proc. IEEE ICCV*, pp. 3429–3437, 2017.

[180] J. Wang *et al.*, “Robust non-perfusion area detection in three retinal plexuses using convolutional neural network in OCT angiography,” *Biomed. Opt. Express*, vol. 11, no. 1, pp. 330-345, Dec. 2020.

[181] Y. Guo *et al.*, “Automated segmentation of retinal fluid volumes from structural and angiographic optical coherence tomography using deep learning,” *Transl. Vis. Sci. Technol.*, vol. 9, no. 2, pp. 54-54, Oct. 2020.

[182] P. Kovesi, “Good colour maps: How to design them,” arXiv preprint arXiv:1509.03700, 2015.

[183] F. Crameri, G. E. Shephard, and P. J. Heron, “The misuse of colour in science communication,” *Nat. Commun.*, vol. 11, no. 1, pp. 1-10, Oct. 2020.

[184] J. Adebayo, J. Glimer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. "Sanity checks for saliency maps," in *Proc. Adv. Neural. Inf. Process. Syst*, 2018.

[185] M. Temme, "Algorithms and transparency in view of the new General Data Protection Regulation," *Eur. Data Prot. L. Rev.*, vol. 3, pp. 473, 2017.

[186] C. F. Baumgartner, L. M. Koch, K. C. Tezcan, J. X. Ang, and E. Konukoglu, "Visual feature attribution using wasserstein gans," in *Proc. CVPR*, 2018, pp. 8309-8319.

[187] T. Xia, C. Agisilaos, and S. A. Tsaftaris, "Pseudo-healthy synthesis with pathology disentanglement and adversarial learning," *Med. Image Anal.*, vol. 64, pp. 101719, Aug. 2020.

[188] Z. Yang, L. Zhao, S. Wu, and C. Y. Chen, "Lung lesion localization of COVID-19 from chest CT image: A novel weakly supervised learning method," *IEEE J. of Biomed. Health Inform.*, vol. 25, no. 6, pp. 1864-1872, Mar. 2021.

[189] Q. Yang, X. Guo, Z. Chen, P. Y. Woo, and Y. Yuan, "D2-Net: Dual Disentanglement Network for Brain Tumor Segmentation with Missing Modalities," *IEEE Trans. Med. Imaging*, vol. 41, no. 10, pp. 2953-2964, May. 2022.

[190] A. Vaswani, *et al.*, "Attention is all you need. In Advances in Neural Information Processing Systems," pp. 6000–6010, 2017.

[191] H. Wang, and D. -Y. Yeung, "A survey on Bayesian deep learning," *ACM Comput. Surv.*, vol. 53, no. 5, Sep. 2020.

[192] D. P. Kingma, *et al.*, "An introduction to variational autoencoders," *Foundations and Trends in Machine Learning*, vol. 12, no. 4, pp. 307-392, 2019.

[193] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning. ICML*, 2015.

[194] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” in *Proceedings of CVPR*, pp. 10684–10695, 2022.