tvst

Artificial Intelligence

Automated Identification of Clinically Relevant Regions in Glaucoma OCT Reports Using Expert Eye Tracking Data and Deep Learning

Ye Tian[1,3], **Anurag Sharma**[1,3], **Shubh Mehta**[1], **Shubham Kaushal**[2,3], **Jeffrey M. Liebmann**[4], **George A. Cioffi**[4], and **Kaveri A. Thakoor**[1–5]

[1] Department of Biomedical Engineering, Columbia University, New York, New York, USA
[2] Data Science Institute, Columbia University, New York, New York, USA
[3] Artificial Intelligence for Vision Science Laboratory, Edward S. Harkness Eye Institute, Department of Ophthalmology, Columbia University Irving Medical Center, New York, New York, USA
[4] Bernard and Shirlee Brown Glaucoma Research Laboratory, Edward S. Harkness Eye Institute, Department of Ophthalmology, Columbia University Irving Medical Center, New York, New York, USA
[5] Department of Computer Science, Columbia University, New York, New York, USA

Correspondence: Kaveri A. Thakoor, Artificial Intelligence for Vision Science Laboratory, Edward S. Harkness Eye Institute, Department of Ophthalmology, Columbia University Irving Medical Center, 622 West 168th St., New York 10032, USA. e-mail: k.thakoor@columbia.edu

Received: June 16, 2023
Accepted: September 4, 2024
Published: October 15, 2024

Keywords: convolutional neural networks; glaucoma; optical coherence tomography; U-Net; eye tracking

Citation: Tian Y, Sharma A, Mehta S, Kaushal S, Liebmann JM, Cioffi GA, Thakoor KA. Automated identification of clinically relevant regions in glaucoma OCT reports using expert eye tracking data and deep learning. Transl Vis Sci Technol. 2024;13(10):24, https://doi.org/10.1167/tvst.13.10.24

Purpose: To propose a deep learning–based approach for predicting the most-fixated regions on optical coherence tomography (OCT) reports using eye tracking data of ophthalmologists, assisting them in finding medically salient image regions.

Methods: We collected eye tracking data of ophthalmology residents, fellows, and faculty as they viewed OCT reports to detect glaucoma. We used a U-Net model as the deep learning backbone and quantized eye tracking coordinates by dividing the input report into an 11 × 11 grid. The model was trained to predict the grids on which fixations would land in unseen OCT reports. We investigated the contribution of different variables, including the viewer's level of expertise, model architecture, and number of eye gaze patterns included in training.

Results: Our approach predicted most-fixated regions in OCT reports with precision of 0.723, recall of 0.562, and f1-score of 0.609. We found that using a grid-based eye tracking structure enabled efficient training and using a U-Net backbone led to the best performance.

Conclusions: Our approach has the potential to assist ophthalmologists in diagnosing glaucoma by predicting the most medically salient regions on OCT reports. Our study suggests the value of eye tracking in guiding deep learning algorithms toward informative regions when experts may not be accessible.

Translational Relevance: By suggesting important OCT report regions for a glaucoma diagnosis, our model could aid in medical education and serve as a precursor for self-supervised deep learning approaches to expedite early detection of irreversible vision loss owing to glaucoma.

Introduction

Surveys have found that 1 out of 40 adults >40 years of age suffers from glaucoma, meaning that approximately 60 million people worldwide are affected by this disease. Of those, 8.4 million are completely blind in both eyes.[1] Even in developed countries, approximately 50% of glaucoma cases go unnoticed until the later stages, when the affected individual starts to experience symptoms such as loss of peripheral vision or tunnel vision. Unfortunately, vision loss resulting from glaucoma cannot be reversed, and there is a need for better ways to diagnose and detect cases of glaucoma.[2–4]

Optical coherence tomography (OCT) is a powerful imaging technique capable of producing high-resolution images that yield precise and measurable data regarding optic disc parameters and retinal nerve fiber layer thickness. As a result of its high accuracy and objectivity, OCT has become widely used as a reliable tool for the detection and monitoring of glaucoma damage and progression.[2] In addition to their usefulness in glaucoma, OCT images are also used for the diagnosis, monitoring, and management of other retinal conditions such as diabetic macular edema and age-related macular degeneration (AMD).[3]

Artificial intelligence (AI) is becoming a promising screening tool for identifying retinal diseases, including retinopathy of prematurity, diabetic retinopathy, glaucoma, and AMD, with human expert–level performance. In particular, deep learning (DL) techniques such as convolutional neural networks (CNNs) have demonstrated successfully their ability to predict the need for anti-vascular endothelial growth factor treatment in patients with AMD with an accuracy rate of 95%.[3] Although AMD-related clinical problems are well-addressed, more and more studies are focusing on the automation of glaucoma detection using DL now.[4]

A number of studies have developed approaches based on various DL architectures to detect glaucoma from OCT images.[5–9] Predicting regions of interest in an OCT report can aid in highlighting the features most relevant for detecting glaucoma and monitoring its progression. If the identification of these important regions can be automated, ophthalmologists could be guided to focus their attention on these regions during the diagnosis process (of particular value to trainees still learning systematic viewing behavior), potentially improving the accuracy and efficiency of glaucoma diagnosis. Additionally, by monitoring changes in these regions over time, ophthalmologists can track the progression of the disease and adjust treatment plans accordingly. Furthermore, by accurately predicting clinically relevant regions in OCT reports from both glaucomatous and nonglaucomatous patients, patterns of differences between these two classes could be used to train self-supervised DL systems, requiring fewer expert labels for supervised training. Overall, predicting regions of interest in OCT reports is anticipated to provide a valuable tool for improving the efficient diagnosis and management of glaucoma.

Eye tracking is a well-studied technique that measures where individuals focus their gaze within their field of view. It has been used extensively to study human visual processing and has recently found increasing application in medical imaging. By tracking eye movements, researchers can gain valuable insights into how people perform visual recognition and search tasks. These insights have the potential to enhance our understanding of how medical images are perceived, interpreted, and acted upon by clinicians, ultimately leading to improved clinical performance and thus better patient outcomes.[10]

Medical experts efficiently direct their gaze to clinically relevant information using learned features in their peripheral vision.[11] It was observed in past work that medical experts with more experience are better at searching (i.e., finding abnormalities faster than novices), because they need fewer eye movements to foveate an abnormality that they first detect peripherally.[12,13] Novice ophthalmologists can thus improve their skills in scanning OCT reports by gaining insights from AI-generated regions of interest on OCT reports (using an AI system trained on expert fixations). Previous studies involving the eye tracking of novice and expert clinicians from various fields demonstrate how novices make interpretive decision errors, indicating a need for improved training, whereas experts exhibit more efficient eye movements and focus on critical diagnostic features.[14] Li et al.[15] trained an attention-guided CNN to predict high-attention regions in fundus images based on simulated clinician eye tracking (clinicians deblurred fundus images based on mouse-clicks, thereby indicating their areas of interest [AOIs]); a second CNN predicted glaucoma based on the localized regions of importance guided by human attention. Although not using OCT or true clinician eye tracking, the attention-guided CNN achieved glaucoma detection accuracy of >95%,[15] showcasing the value of integrating human attention into DL model training. Some studies also highlighted the importance of using eye tracking to monitor skill development during medical education and training[16] and how this learning could be aided by leveraging AI models.[17,18] Another study explored new approaches for analyzing the eye movement behavior of radiologists viewing brain magnetic resonance images to investigate how radiologists and nonexperts view and interpret magnetic resonance images differently. The authors presented a new method of analyzing eye movement patterns called gaze density, which provides a more detailed visualization of where participants look while viewing the images. This study also found that experts spent more time looking at AOIs and had more efficient eye movements than nonexperts. The authors suggest that gaze density analysis can provide valuable insights into the visual information processing strategies experts use in medical image interpretation.[18]

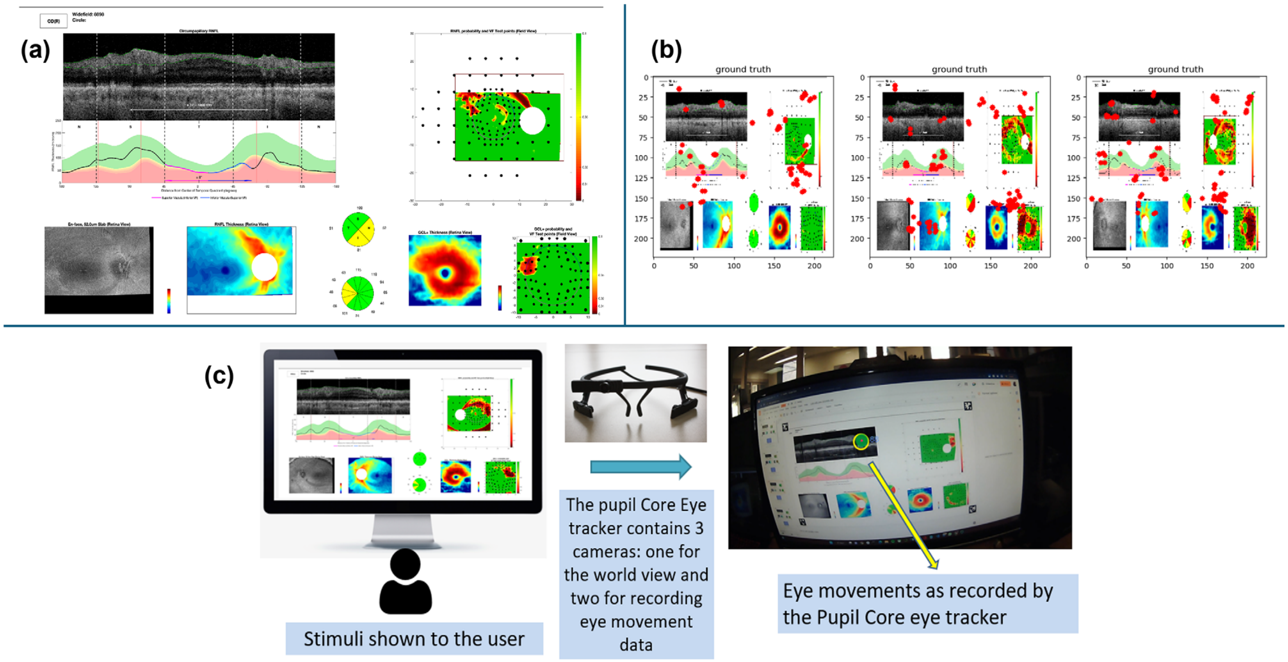
In this work, we used U-Net–based DL techniques in a novel setting to predict the most clinically relevant and important regions fixated by medical experts by tracking their eye movements on OCT reports while they diagnosed glaucoma. Unlike past approaches in nonmedical settings that use U-Net for grid-based depth estimation[19] or fine-grained weather forecasting,[20] using a grid or even for saliency prediction in natural images using a CNN and transformer-based backbone called TranSalNet,[21] our approach is unique in medical imaging in that our goal is not pixel-based image segmentation, but rather fixation map prediction quantized via a grid. By using our fixation prediction U-Net model on unseen OCT reports, trainees and clinicians may be able to identify regions of interest within OCT reports more efficiently and accurately, ultimately leading to more accurate diagnoses and improved patient outcomes.

Methods

Eye Tracking Data Collection

Fifteen ophthalmologists of varying expertise, including residents, fellows, and faculty from the Edward S. Harkness Eye Institute, Columbia University Irving Medical Center, viewed OCT reports like the one depicted in Figure 1a. The demographic makeup of these clinicians consisted of three Caucasians, six Asians, five Africans/Middle Easterners, and one Hispanic (eight female and seven male overall). For expert 1 to expert 7, a random sample of 20 OCT reports was taken from a pool of 231 reports (collected in a prior study[22]), following a noncontrolled sampling approach. Starting from expert 8, a control set, consisting of a fixed set of 20 reports, were chosen out of the 231 OCT reports and were shown to each expert. The control set was introduced to examine the impact on model performance when experts were shown the same images. From expert 8 onward, experts were presented with both the control set and the noncontrol set. Thus, every expert viewed an identical set of 20 OCT reports in the control set; note this set was composed of straightforward established glaucoma[22] or healthy cases. Whereas, in the noncontrol set, 20 different reports were selected randomly (with replacement) from the pool of 231 reports for each expert; note that this set consisted of more complex or ambiguous glaucoma suspect cases.

The current study, protocol AAAU4079, was approved by the Columbia University Irving Medical Center Institutional Review Board and was conducted in accordance with the tenets set forth by the Declaration of Helsinki. Informed consent was obtained from all study participants. Expert eye tracking data was recorded using a Pupil Labs Core eye tracker while experts examined OCT reports for glaucoma. Eye tracking coordinates were normalized to match the dimensions of the OCT reports. Figure 1a shows



Stimuli shown to the user

The pupil Core Eye tracker contains 3 cameras: one for the world view and two for recording eye movement data

Eye movements as recorded by the Pupil Core eye tracker

Figure 1. (**a**) Full Topcon OCT report. (**b**) Set of three OCT reports overlayed with expert fixations. (**c**) Experimental Setup using Pupil Labs Core eye tracker.

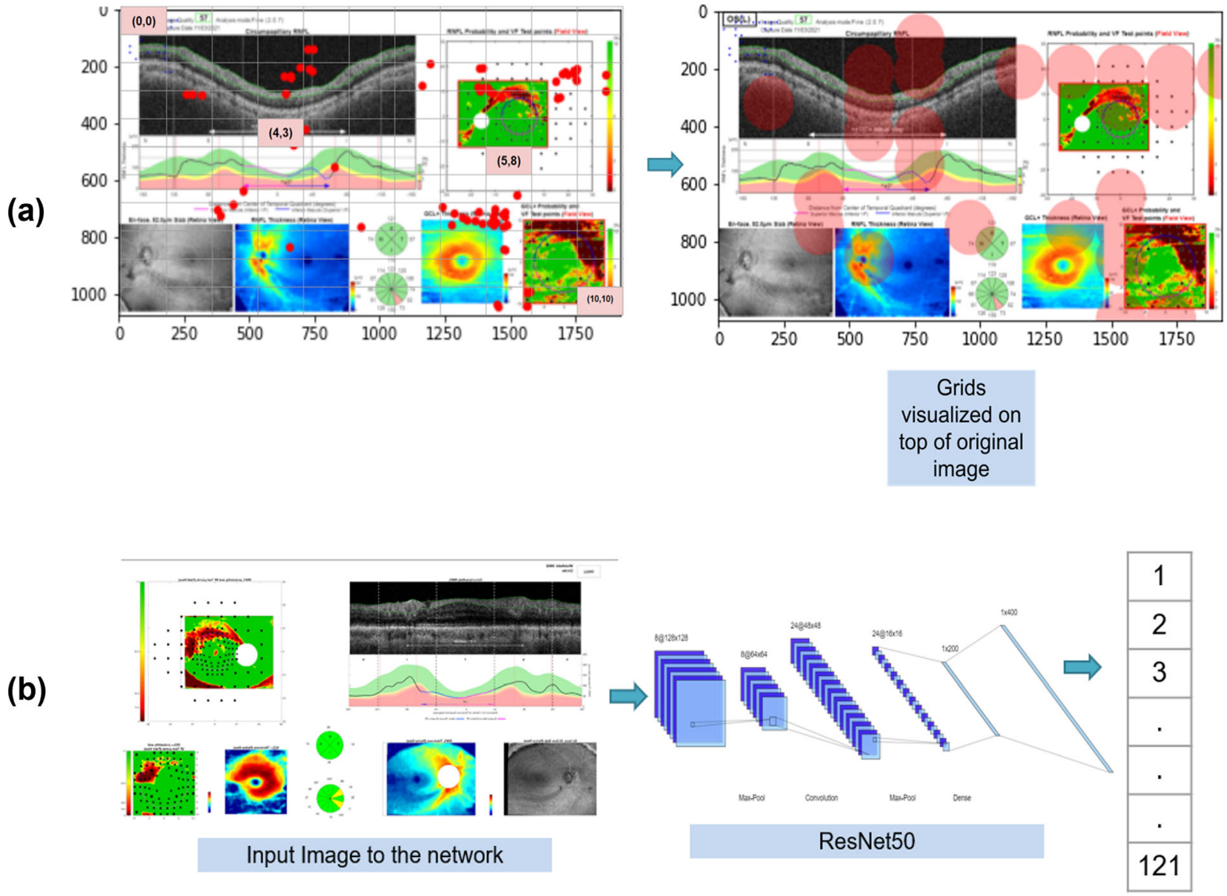


Figure 2. (**a**) Pixel-to-grid transition and overlay on OCT reports. (**b**) ResNet model pipeline.

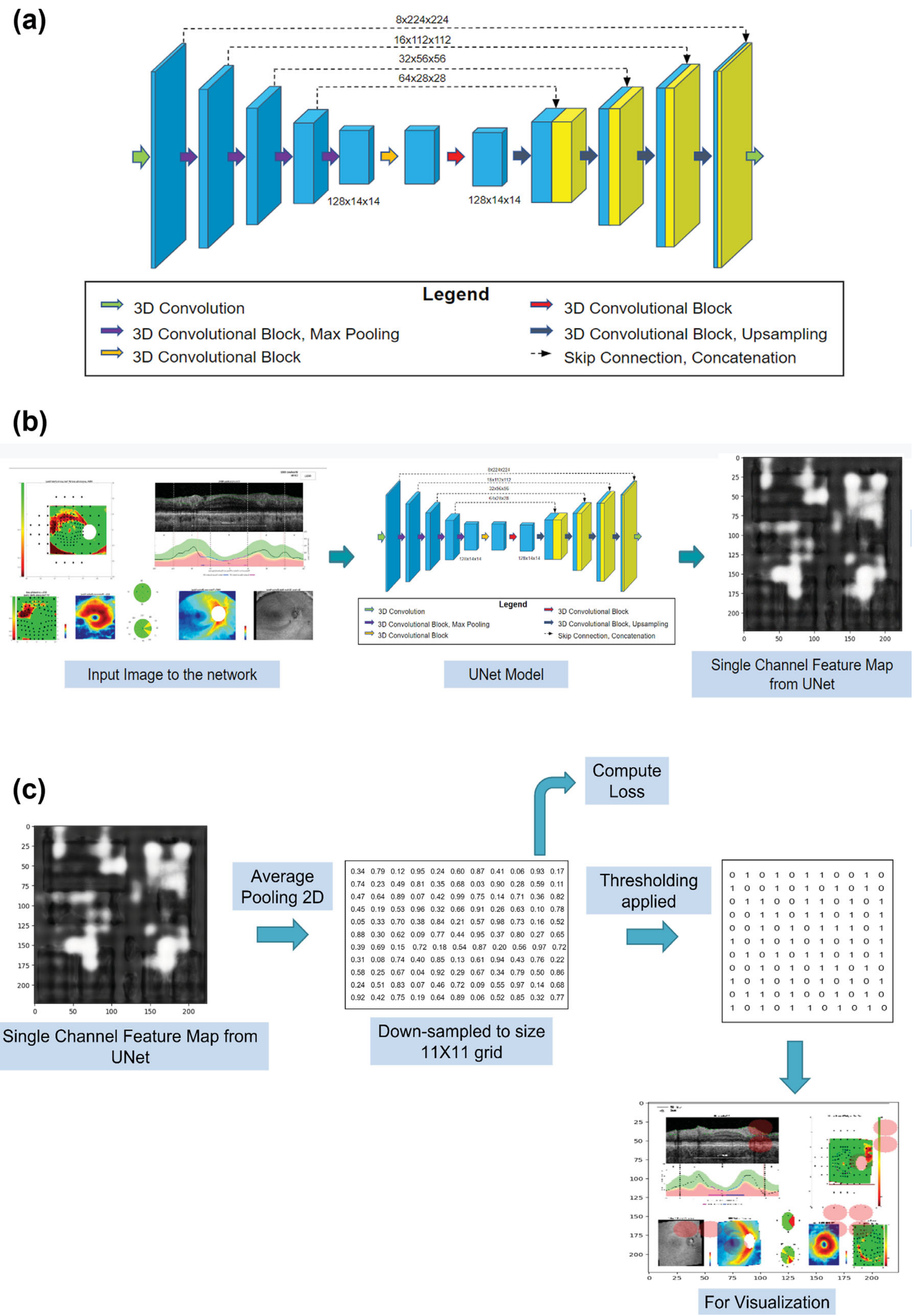
an example of a widefield OCT report used in our study. Figure 1b illustrates three OCT reports on which eye tracking coordinates are overlaid, representing the gaze points of a clinician. Figure 1c shows the eye tracker and data collection pipeline.

Data Preprocessing

The OCT reports were resized to a resolution of 224 × 224 pixels, and the corresponding eye coordinates were also mapped to the same resolution. Because each individual had a different viewing pattern, using raw eye fixation coordinates could result in too much variance in the data; therefore, we downsampled the fixation coordinates to an 11 × 11 grid map on the original image, where each grid is in the shape of a square with equal side lengths. In this way, the OCT reports were divided into 121 grids as shown in Figure 2a. Whenever a fixation fell in a grid, that grid was assigned a value of 1 (otherwise 0). The same binary transformation was applied to every OCT report observed by an expert. After the image-level preprocessing, the full dataset was further split randomly into 80% for training, 10% for validation, and 10% for testing. The performance on the test set is reported in the Results using the best-performing model on the validation set. In addition, a given patient or eye was only present in one of these partitions.

The ResNet Model Architecture

We evaluated two DL architectures to learn clinician fixation data on OCT reports. First, we used ResNet, a popular DL architecture used for various computer vision tasks, including image classification, object detection, and segmentation. ResNet was proposed in 2015[23] and has shown superior performance compared with earlier architectures, such as VGG[24] and Inception.[25] We used a ResNet model pretrained on the ImageNet dataset[26] and modified the number of connections in the dense layers. The output vector length was increased to 121, corresponding to the 121-element fixation grid map described in the previous section. Our ResNet model architecture is shown in Figure 2b.


(a)
8x224x224
16x112x112
32x56x56
64x28x28
128x14x14
128x14x14
Legend
3D Convolution
3D Convolutional Block, Max Pooling
3D Convolutional Block
3D Convolutional Block
3D Convolutional Block, Upsampling
Skip Connection, Concatenation
(b)
Input Image to the network
UNet Model
Single Channel Feature Map from UNet
(c)
Single Channel Feature Map from UNet
Average Pooling 2D
Down-sampled to size 11X11 grid
Compute Loss
Thresholding applied
For Visualization

0.34	0.79	0.12	0.95	0.24	0.60	0.87	0.41	0.06	0.93	0.17
0.74	0.23	0.49	0.81	0.35	0.68	0.03	0.90	0.28	0.59	0.11
0.47	0.64	0.89	0.07	0.42	0.99	0.75	0.14	0.71	0.36	0.82
0.45	0.19	0.53	0.96	0.32	0.66	0.91	0.26	0.63	0.10	0.78
0.05	0.33	0.70	0.38	0.84	0.21	0.57	0.98	0.73	0.16	0.52
0.88	0.30	0.62	0.09	0.77	0.44	0.95	0.37	0.80	0.27	0.65
0.39	0.69	0.15	0.72	0.18	0.54	0.87	0.20	0.56	0.97	0.72
0.31	0.08	0.74	0.40	0.85	0.13	0.61	0.94	0.43	0.76	0.22
0.58	0.25	0.67	0.04	0.92	0.29	0.67	0.34	0.79	0.50	0.86
0.24	0.51	0.83	0.07	0.46	0.72	0.09	0.55	0.97	0.14	0.68
0.92	0.42	0.75	0.19	0.64	0.89	0.06	0.52	0.85	0.32	0.77

0	1	0	1	0	1	1	0	0	1	0
1	0	0	1	0	1	0	1	0	1	0
0	1	1	0	0	1	1	0	1	0	1
0	0	1	1	0	1	1	0	1	0	1
0	0	1	0	1	0	1	1	1	0	1
1	0	1	0	1	0	1	0	1	0	1
0	1	0	1	0	1	1	0	1	0	1
0	0	1	0	1	0	1	1	0	1	0
1	0	1	0	1	0	1	0	1	0	1
0	1	1	0	1	0	0	1	0	1	0
1	0	1	0	1	1	0	1	0	1	0

Figure 3. (**a**) U-Net architecture: it has four downsampling and four upsampling layers that are further connected by skip connections. (**b**) U-Net training. The model takes three-channel images as input and outputs a single-channel feature map. (**c**) Use of feature map to create grids: the output feature map from U-Net is fed into an AveragePool2D layer, which downsamples it to an 11 × 11 grid map giving probabilities for presence of fixations in grids. A 0.5 threshold is applied to these probabilities to obtain fixation predictions in 0 (fixation absent) or 1 (fixation present).

The U-Net Model Architecture

The second architecture we investigated was U-Net, another popular DL architecture that is applied widely for biomedical image segmentation and reconstruction tasks, and many downstream models have been derived from it since its emergence in 2015.[27] We modified the U-Net (Fig. 3a) to take as input the OCT report at 224 × 224 resolution and output a single-channel grayscale feature map of the same resolution, as shown in Figure 3b. The skip connections concatenate feature maps in the encoder and decoder to preserve spatial details and improve accuracy. The model consists of two convolution layers followed by four downsampling layers and four upsampling layers, as shown in Figure 3a. The network architecture applied batch normalization and ReLU activation function in all layers, except for the last convolutional layer, which used Sigmoid as the activation function. The final layer of the U-Net produced a single-channel feature map with a shape of 224 × 224 pixels.

Subsequently, this feature map underwent downsampling using an average-pooling layer, resulting in an 11 × 11 grid map, as depicted in Figure 3c. This 11 × 11 grid represents the probabilities of the presence or absence of fixation within each grid cell. To convert these probabilities into binary values, a threshold of 0.5 was compared with the probabilities generated for each grid, as illustrated in Figure 3c. Following this thresholding step, the resulting 11 × 11 grid map, consisting of 0s and 1s, was used for calculating various metrics such as dice score, weighted precision, recall, f1-score, and accuracy.

Additionally, for visualization purposes, we overlaid this 11 × 11 grid map on top of the original OCT report, as demonstrated in Figure 3c.

Loss was calculated at the grid level after pooling and before thresholding as explained in the next section. We envision our task as image segmentation rather than classification on each grid.

Training and Model Optimization

We implemented experiments with ResNet and U-Net models which are summarized in Figures 2 and 3, respectively. The output of the ResNet model was a 121-element vector; in contrast, for U-Net, after applying average pooling to U-Net model output from the last layer, the final prediction was an 11 × 11 grid map. The loss for each element in the 11 × 11 grid map was calculated as a separate binary classification. The choice of loss function for both models was binary cross entropy (BCE), formulated as:

$$\textit{Binary Cross Entropy} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij}\log\left(p_{ij}\right) + \left(1-y_{ij}\right)\log\left(1-p_{ij}\right)\right)$$

where N is the total number of samples (images) in the batch, M is the total number of fixation labels per sample (121 in our case), y_{ij} is the true label (either 0 or 1) for the i-th sample and j-th label, and p_{ij} is the predicted probability for the i-th sample and j-th label.

During training, the BCE loss was calculated and averaged over each of the 121 fixation predictions. We also implemented cosine similarity loss and dice loss, anticipating they might be better at evaluating feature similarity and giving grid predictions. However, training with BCE loss performed the best consistently; thus, we selected BCE to continue with all subsequent experiments. The optimizer chosen for minimizing the loss was Adam, with a learning rate of 1×10^{-4} and a step decay scheduler.

Our model generated pixel-level predictions, but after pooling the output, model weights were updated via the grid-level loss. We monitored the change in Sørensen–Dice coefficient (Dice score) during model training to measure the overlap between the predicted grid mask and the ground truth and used the Dice score as the metric for finding the model with the best validation performance.

Considering our grid map dataset is imbalanced toward the nonfixation class, we measured the weighted average precision, which calculates the precision for each class individually and then computes a weighted average of those precisions based on the class frequencies in the dataset. Compared with macro average precision, the weighted average precision is formulated as:

$$\textbf{Macro-Avg Precision} = 0.5 \times \text{Precision}_{\text{glaucoma}} + 0.5 \times \text{Precision}_{\text{healthy}}$$

$$\textbf{Weighted-Avg Precision} = \text{W}_{\text{glaucoma}} \times \text{Precision}_{\text{glaucoma}} + \text{W}_{\text{healthy}} \times \text{Precision}_{\text{healthy}}$$

Where the weights and precision are defined as:

$$\mathbf{W_{glaucoma}} = \frac{\#\,\text{glaucoma}}{\#\,\text{glaucoma} + \#\,\text{healthy}}$$

$$\mathbf{Precision_{glaucoma}} = \frac{\text{TP glaucoma}}{(\text{TP} + \text{FP})\text{ glaucoma}}$$

Table 1. Model Performance Using ResNet

Data	Precision	Recall	F1-Score	Accuracy
Expert 1–6 (Topcon)	0.590 (0.493–0.689)	**0.533** (0.435–0.631)	0.546 (0.448–0.644)	**0.533** (0.435–0.631)
Expert 1–6 (Zeiss)	0.638 (0.540–0.736)	0.506 (0.408–0.604)	0.543 (0.445–0.641)	0.506 (0.408–0.604)
Expert 1–6 (Topcon + Zeiss)	**0.682** (0.584–0.780)	0.502 (0.404–0.600)	**0.554** (0.456–0.652)	0.502 (0.404–0.600)

The input for each row is the noncontrol set of full OCT report images without augmentation. In each row, we compare the same cohort of expert 1–6 data from either TopCon, Zeiss, or both combined.

Metric values and 95% confidence intervals are presented. The highest metric in each column is bolded.

Table 2. Model Performance Using U-Net With Data From Different Cohort of Experts and Faculty

Data	Set	Precision	Recall	F1-Score	Accuracy
Expert 1–6 (Topcon)	Noncontrol	0.707 (0.609–0.805)	0.536 (0.438–0.634)	0.583 (0.485–0.681)	0.536 (0.438–0.634)
All experts (Topcon)	Noncontrol	0.747 (0.649–0.845)	0.479 (0.381–0.577)	0.554 (0.456–0.652)	0.479 (0.381–0.577)
All faculty (Topcon)	Noncontrol	0.643 (0.545–0.741)	0.530 (0.432–0.628)	0.564 (0.466–0.662)	0.530 (0.432–0.628)
Faculty 10	Noncontrol	**0.801** (0.703–0.899)	0.477 (0.379–0.575)	0.564 (0.466–0.662)	0.477 (0.379–0.575)
Faculty 10	Control	0.723 (0.625–0.821)	**0.562** (0.464–0.660)	**0.609** (0.511–0.707)	**0.562** (0.464–0.660)
Faculty 13	Control	0.784 (0.686–0.882)	0.479 (0.381–0.577)	0.575 (0.477–0.673)	0.479 (0.381–0.577)
Faculty 10,13	Control	0.735 (0.637–0.833)	0.446 (0.348–0.544)	0.513 (0.415–0.611)	0.446 (0.348–0.544)
Faculty 10,13, and expert 12	Control	0.760 (0.662–0.858)	0.525 (0.427–0.623)	0.596 (0.498–0.694)	0.525 (0.427–0.623)

The highest metric in each column is bolded.

Metric values and 95% confidence intervals are presented.

$$\mathbf{W}_{\mathbf{healthy}} = \frac{\#\,\text{healthy}}{\#\,\text{glaucoma} + \#\,\text{healthy}}$$

$$\mathbf{Precision}_{\mathbf{healthy}} = \frac{\text{TP healthy}}{(\text{TP} + \text{FP})\ \text{healthy}}$$

Similarly, we computed weighted average recall and weighted average f1-score. Recall focuses on capturing as many positive instances as possible and is defined as the proportion of true positive predictions to the sum of true-positive predictions and false-negative predictions. We can derive and find the weighted average recall is equal to accuracy, which is consistent with the equal recall and accuracy numbers in result Tables 1–5:

Weighted-Avg Recall

$$\begin{aligned}
&= \text{Recall}_{\text{glaucoma}} \times \frac{(\text{TP} + \text{FN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \\
&+ \text{Recall}_{\text{healthy}} \times \frac{(\text{TN} + \text{FP})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \\
&= \frac{\text{TP}}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} + \frac{\text{TN}}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})} \\
&= \text{Accuracy}
\end{aligned}$$

Using the universal training parameters described above, we explored potential factors that could influence the model performance:

Table 3. Model Performance Comparison Between Our Customized U-Net Model and Machine Learning (Random Forest) and Benchmark Multilayer Perceptron models

	Fails to Learn and Predicts Chance			
Model	Precision	Recall	F1-Score	Accuracy
Train and test on all experts				
Random Forest				
Multilayer perceptron	0.638	0.506	0.543	0.506
	(0.540–0.736)	(0.408–0.604)	(0.445–0.641)	(0.408–0.604)
U-Net	0.747	0.479	**0.554**	0.479
	(0.649–0.845)	(0.381–0.577)	(0.456–0.652)	(0.381–0.577)
Train and test on faculty 10 (control) data				
Multilayer perceptron	0.752	0.488	0.560	0.488
	(0.654–0.850)	(0.390–0.586)	(0.463–0.659)	(0.390–0.586)
U-Net	0.723	0.562	**0.609**	0.562
	(0.625–0.821)	(0.464–0.660)	(0.511–0.707)	(0.464–0.660)

U-Net outperformed the benchmark on F1-score when experimenting on either all experts or faculty 10.
Metric values and 95% confidence intervals are presented. The highest F-1 scores across all experiments are bolded.

Table 4. Model Performance Comparison Among Different Grid Sizes of 10 × 10, 11 × 11, and 12 × 12 Using Our Customized U-Net Model

Grid Size	Precision	Recall	F1-Score	Accuracy
Train and test on all experts				
10 × 10	0.660	0.467	0.521	0.467
	(0.562–0.758)	(0.369–0.565)	(0.423–0.619)	(0.369–0.565)
11 × 11	0.747	0.479	**0.554**	0.479
	(0.649–0.845)	(0.381–0.577)	(0.456–0.652)	(0.381–0.577)
12 × 12	0.664	0.503	0.547	0.503
	(0.566–0.762)	(0.405–0.601)	(0.449–0.645)	(0.405–0.601)
Train and test on faculty 10 (control) data				
10 × 10	0.705	0.530	0.581	0.530
	(0.607–0.803)	(0.432–0.628)	(0.482–0.679)	(0.432–0.628)
11 × 11	0.723	0.562	**0.609**	0.562
	(0.625–0.821)	(0.464–0.660)	(0.511–0.707)	(0.464–0.660)
12 × 12	0.783	0.483	0.556	0.483
	(0.685–0.881)	(0.385–0.581)	(0.458–0.654)	(0.385–0.581)

The choice of grid size = 11 × 11 achieved the best F1-score when experimenting on either all experts or faculty 10.
Metric values and 95% confidence intervals are presented. The highest F-1 scores across all experiments are bolded.

1. Model architecture: the trained-from-scratch ResNet50 which learned to predict a 121-element binary classification vs. the U-Net, which simulated an 11 × 11-grid image segmentation task, where each grid represented a pixel labeled as 0 or 1.
2. The expertise level of participants at glaucoma diagnosis: resident group 1 ($<$12 months experience), resident group 2 (24–36 months experience), and fellow or faculty group (undergoing glaucoma fellowship training of $\geq$36 months through $\geq$30 years of experience).
3. Types of OCT devices that generated the OCT reports: Topcon vs. Zeiss.
4. Number of participants included in the analysis: using a superset of data from all participants or a subset of data from the first six participants. We also incrementally added fellow and

Table 5. Model Performance With or Without White Space Removal on the OCT Report Using Our Customized U-Net Model

White Space Removal	Precision	Recall	F1-Score	Accuracy
Train and test on all experts data				
Yes	0.674 (0.576–0.772)	0.527 (0.429–0.625)	**0.572** (0.475–0.671)	0.527 (0.429–0.625)
No	0.747 (0.649–0.845)	0.479 (0.381–0.571)	0.554 (0.456–0.652)	0.479 (0.381–0.577)
Train and test on faculty 10 (control) data				
Yes	0.690 (0.592–0.788)	0.475 (0.377–0.573)	0.527 (0.429–0.625)	0.475 (0.377–0.573)
No	0.723 (0.625–0.821)	0.562 (0.464–0.660)	**0.609** (0.511–0.707)	0.562 (0.464–0.660)

Removing white space could improve the F1-score when training and testing the model on all experts' data, but led to a significant decrease on only faculty 10's data.

Metric values and 95% confidence intervals are presented. The highest F-1 scores across all experiments are bolded.

faculty experts to evaluate the impact of number of participants on performance.

Results

Seven sets of experiments were conducted with different input data modalities, participant cohorts, and model architectures, all of which are summarized in Table 1 and Table 2. Table 1 specifically contains results pertaining to the ResNet model performance, while Table 2 describes the U-Net model performance. The 95% confidence intervals are presented along with each metric value. We showcase results with the following performance metrics: precision, recall, f1-score, and accuracy, respectively, on the test set.

The precision, recall and f1-score are weighted, calculating an overall score that considers not only the performance for individual classes but also the distribution of those classes in the dataset. Given the high imbalance between nonfixation and fixation points in our data, f1-score was considered as the most representative metric of the model's predictive power. The f1-score is the harmonic mean of precision and recall; thus, a f1-score of 0.5 is not the same as an accuracy of 0.5; f-1 scores of <0.5 also have meaning. The model's accuracy was defined as the number of correctly predicted grids (with or without fixations) divided by the total number of 121 grids.

It is important to note that there are multiple reasons why chance in our setting is not equal to 0.5. First, the average number of positive fixations on all images viewed was 22.64 (18.71%), implying a class imbalance. Conventional AI performance metrics are not the best fit to evaluate predictions of eye tracking fixations, because the fixated grids are spatially and temporally connected. For example, Liebmann et al.[28] established a method to guide clinicians on the order of examining an OCT report. Despite not knowing about this educated human behavior, our AI model usually predicts a part of the recommended fixation series, leading to its high precision, but a seemingly worse recall performance. Rather than a probability of 0.5, a more valid probability of chance in our setting could be ($AOI_1 \times 0.5 + AOI_2 \times 0.5 + AOI_3 \times 0.5 + \ldots\ldots + AOI_n \times 0.5$), where the number n of AOIs varies based on the number of correlated regions in an OCT report. These grid inter-relationships explain why the translational power of our model could be underestimated by simply looking at conventional values of accuracy and similar metrics.

We found that the U-Net model combined with a grid-based approach performed best when trained on expert 10 control data; it achieved a precision of 0.72, recall of 0.56, f1-score of 0.61, and accuracy of 61% at predicting medically relevant regions of interest. These performances were the best among all experiments, except that using data from faculty/fellow 10 noncontrol, which resulted in a best accuracy of 80%. Our model's high precision indicates that it is capable of making more conservative and precise predictions.

Choice of DL Model: ResNet Vs. U-Net

We investigated the performance of two models: ResNet50 for multiclass classification and U-Net for grid-wise image segmentation. Comparing the first row of Table 1 and the first row of Table 2, when experimenting on experts 1 through 6 Topcon data, the U-Net model significantly surpasses ResNet50 in precision and f1-score ($P < 0.05$, Wilcoxon signed-rank test). The reason behind this disparity is that ResNet50 predicted a large number of the same fixation grids for every report input, resulting in a higher recall. In contrast, the U-Net model demonstrated superior performance by generating fewer but more diverse and precise predicted output fixation maps for different OCT report inputs.

Level of Glaucoma Expertise

We divided all 15 participants into different groups based on their expertise at diagnosing glaucoma. Among them were 4 glaucoma faculty, 1 glaucoma fellow, and 10 residents with training experience of <36 months. From the first and second rows of Table 2, we observed that the model's fixation prediction on the Topcon reports by six residents performed with f1-score of 0.58; inputting all available expert data confused the model and made the overall prediction worse, with f1-score of 0.55. This result could be because each expert had distinct habits of reading OCT reports, such as speed, area of interest, and gazing trajectory. In contrast, residents still undergoing training seemed to exhibit a more routine style of observing OCT reports. They preferred to view most of the regions on the reports before giving their final diagnosis, leading to a relatively greater number of fixations than the combined expert data. This difference could explain why the precision is higher and recall is lower for results from all experts combined compared with those from the six residents only.

Type of OCT Reports

The third variable of interest was the type of OCT reports: Topcon and Zeiss OCTs both capture information of the retina and optic nerve to diagnose and monitor glaucoma. They have various sizes, laterality, and orientations of images on each report. The first two rows of Table 1 indicate the ResNet50 model performed better on Topcon reports alone than on Zeiss reports alone, in terms of recall, and accuracy. In the third row, when we combined both datasets of OCT reports by resizing and registering them, only the precision showed a significant increase, whereas all other metrics obtained similar or worse results, compared with using just Topcon or just Zeiss data alone. With data augmentations such as random flipping applied during training, the combined dataset did not generate more promising results. We, therefore, used only Topcon reports for further experiments and analysis. We attribute this result primarily to our Zeiss report layout being different for left vs. right eyes, making it difficult for the model to learn a standard format for Zeiss reports; in future work, we plan to use Zeiss bilateral OCT reports (both left and right eye on the same report), so model performance may differ.

Optimizing Dataset Size Vs. Expertise

DL model training also is highly related to the amount of input information and size of the dataset. When input data were used from only six residents (first row of Table 2), the model tended to predict a very limited number of fixations. But when input data from all experts (resident, fellow, and faculty) were used (second row of Table 2), all metrics except precision showed a performance drop. This decrease with an enlarged input dataset size seems to be counterintuitive to the common DL principle that usually more training data improves the model's ability to learn features. However, in eye tracking, increasing the number of participants does not necessarily increase common fixation patterns for a convolutional AI model to extract; instead, adding more participants to a cohort of only faculty participants during the training process introduces greater variability, making the ground truth data more random.

To optimize and find a balance between increasing dataset size and finding consistent eye gaze patterns, we implemented experiments with incremental input information from the faculty/fellow–only group (those experts with most experience), with results shown in the third to sixth rows of Table 2. To better compare how the fixations from faculty-level vs. resident-level experience affected model performance, we experimented on the U-Net model using data from single faculty/fellow #10, single faculty/fellow #13, and single faculty/fellow #12 alone. The visualization of three example pairs of ground truth vs. fixation predictions are shown in Figure 4. On faculty/fellow #13 data alone, the baseline f-1 score was only 0.58. When we trained the model with data from one more faculty, (i.e., faculty #10 and #13 together), the f1-score significantly decreased from 0.58 to 0.51, indicating an essential disagreement in the fixation information from these two faculty members. It should be noted that faculty #10 data alone trained the model to have an f1-score of 0.61. Finally, when data from participant #12 (with less experience) was combined, the f1-score increased back to 0.60, showing the influence of agreement and consensus on the model's capability of predicting fixations using data from participants at different glaucoma expertise levels.

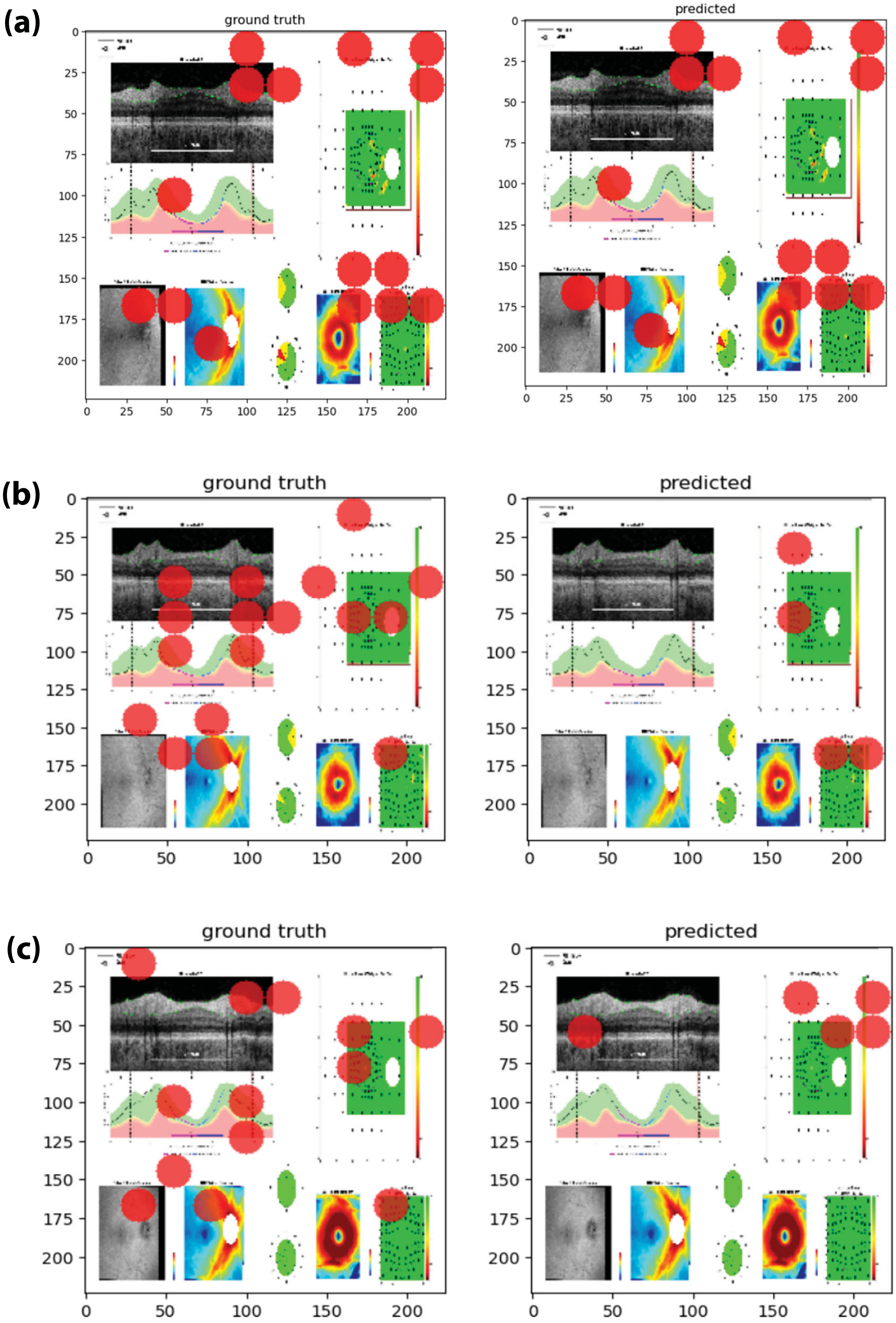


Figure 4. Fixation prediction results when incrementally adding faculty/fellow participants. (**a**) Results with faculty/fellow 10 data alone. (**b**) Results with faculty/fellow 10 and 13. (**c**) Results with faculty/fellow 10, 13, and 12.

Baseline Models

To prove the robustness of our customized U-Net model, we obtained baseline results from benchmark models such as machine learning models and the multilayer perceptron model. The results are shown in Table 3. We picked a random forest model because its learning mechanism aligns the best with the nature of our data format, but it predicted all zeros for all input, leading to a blank predicted (fixation-less) grid map. The multilayer perceptron model's performance was surpassed by our U-Net whether experimenting on all experts or faculty 10 alone, indicating the superiority of our approach.

Grid Size

We also tuned the grid size (i.e., number of grids in one OCT report image), as shown in Table 4. Experimenting on a smaller grid size of 10 × 10 or a larger grid size of 12 × 12 worked comparably well when keeping other variables the same. The grid size of 11 × 11 performed best among all, confirming its fit for our task of predicting eye tracking.

White Space Removal

Removing fixations on white space indeed increased f1-score from 0.55 to 0.57 on all experts' data, as shown in Table 5. It enabled the removal of unreasonable labels resulting from participants fixating outside the AOIs and retained only information within useful regions such as the b-scans and probability maps. However, this enhancement was not always exhibited; in fact, the removed labels could still be useful when training and testing on smaller cohorts of data. For example, when only using faculty 10 data, the f1-score significantly decreased from 0.61 to 0.53 after white space removal. This finding suggests that data augmentation techniques such as image background thresholding can be useful when the dataset is large, but may not always be suitable for particular participants.

Discussion

In this study, our objective was to develop a DL model capable of predicting the regions of interest that medical experts focused on while examining OCT reports. To achieve this goal, we collected eye tracking data from experts when they classified OCT reports as either glaucomatous or nonglaucomatous. We used several strategies during the training of our DL models, including experimenting with different model architectures, considering the expertise level of the participants, analyzing different types of OCT reports and evaluating the tradeoff of the number of participants included in the analysis.

Our model excels in high precision. Greater precision underscores the effectiveness of our approach in assisting ophthalmologists with the diagnosis of glaucoma. Although our model potentially predicts fewer grids than the ground truth, those conservatively predicted grids are informative and precise. Considering our study's purpose is to investigate AOIs on OCT reports to educate novices, fewer and more accurate grid predictions can better hint and remind trainees of regions not to be missed on OCT reports to expedite diagnosis.

Several factors influenced precision and recall in our study. Training the model on individual experts allowed it to capture the unique viewing behavior of each expert, leading to fewer predicted grids with greater precision. Each expert had their own distinctive patterns of fixations, and training on individual experts helped the model to learn and replicate these patterns effectively. Also, when we added more experts to the training process, the ground truth data became more random, introducing greater variability. This strategy increased variability in the data helped the model to learn generalized patterns, resulting in a higher recall and precision. The f1-scores, as the harmonic mean of precision and recall, were also influenced in a similar way: training on individual experts increased f1-scores.

To optimize and find a balance between dataset size and common eye gaze patterns, we examined the diagnosis accuracy of the most experienced participants, faculty 10 and 13. Faculty 10, who had the highest level of experience among the faculty, demonstrated 100% glaucoma classification accuracy (compared with ground truth labels obtained via consensus), as shown in Figure 4a, when classifying the noncontrol set of OCT reports. In contrast, expert 13, who had less experience, showed an 85% accuracy rate on the same dataset. These findings align with the expectation that increased experience correlates with higher classification accuracy when using expert eye tracking data.

In this study, our approach predicts clinically relevant regions in OCT reports using DL models trained on expert eye movements, thus aiding novice clinicians in learning important OCT report regions to which experts attend. For example, the predictions from our models could be overlayed in a virtual reality/augmented reality environment, to guide ophthalmology trainees to important regions in OCT reports. Alternatively, medically salient regions predicted by our model (after being trained by experts) could be used to assess skill progression of trainees by measuring their eye fixation similarity to model-predicted regions.

Limitations and Future Directions

Our study has scope for improvement in the following ways: first, we will collect our data using a higher frame-rate screen-based eye tracker such as the Tobii Pro Fusion. Owing to the pioneering nature of our approach, we only experimented with data from a single kind of Pupil Labs eye tracking device that had a 200-Hz frequency (sampling rate). By adding and comparing data from the Tobii device that has 250-Hz frequency and thus delivers more accurate data on what captures a user's attention, we may be able to gain deeper insights and potentially better AI model results. Second, we will explore more possibilities for implementing AI models on the technical side. Diffusion-based and attention-based models that have emerged in recent years have proved their outstanding performance in image comprehension and segmentation tasks, as well as their strong compatibility with medical data such as OCT images. The success of the current U-Net approach gives us the confidence to extend AI's predictive power via larger nonconvolutional neural network models. Last but not least, we will overcome the limitation of our small dataset size, model scale, and the number of tunable parameters by leveraging pretrained weights and finetuning strategies to enhance downstream accuracy.

To validate the feasibility of our method and establish a precedent for DL-based eye fixation prediction, we used a relatively straightforward U-Net architecture with approximately 30 million trainable parameters. Moving forward, our objective is to enhance the robustness of our approach by exploring more sophisticated image-to-image translation algorithms. A promising avenue for improvement involves the use of generative adversarial networks, a type of neural network architecture capable of learning the underlying distribution of a dataset to generate new images.

In the future, we will also focus on expanding this approach to other medical reports where eye tracking data from experienced medical professionals could aid in AI models in predicting important regions of interest, such as in radiology, cardiology, and neurology. Furthermore, along with predicting regions of interest, the AI model could be trained to predict the sequence (order) in which regions of interest are viewed. This would completely capture the gaze behavior (timing and location) of experts.

Moreover, future work could investigate alternative AI training techniques, including the use of the predicted 11 × 11 grid to train a self-supervised model. Using this approach, our model would predict whether an OCT report is indicative of glaucoma or not based on eye fixations overlaid on OCT reports. Consequently, given the accurate prediction of eye fixations from an OCT report, we could further train a model to predict the diagnosis of glaucoma based on eye fixations alone. Instead of relying on explicit and costly hand-provided labels by experts, the differences in fixation patterns between different disease classes would serve as labels. This approach could significantly reduce the need for manual labeling and make the training process more efficient.

The models developed in this study could also be integrated with other existing models,[25] which specifically focus on using eye tracking data from ophthalmology experts as a substitute for positional embeddings in a Vision Transformer model.[29] This integration would aim to use predicted eye movements from our model instead of an average gaze pattern for downstream Vision Transformer–based classification of OCT reports as glaucomatous or healthy. This dual capability of predicting both eye fixations and diagnosis from an OCT report holds significant potential for advancing the effectiveness and capabilities of our approach.

Furthermore, the higher f1-score observed in the diagnoses made by faculty 10 prompts an inquiry into the specific features this expert prioritized during their assessments. Consequently, training the AI model to predict these specific features (shown in Figure 4a) can provide valuable insights into the underlying mechanisms used by both the experts and the AI model. This direction holds significant implications for medical education, AI training, and the interpretability of the diagnostic process.

Overall, these potential applications highlight the broad applicability and future directions of incorporating eye tracking data into AI research and glaucoma detection, offering innovative approaches for training models and improving the efficiency of the labeling processes. In the clinic, such fixation prediction can also aid in the training of novice clinicians and provide insights into significant regions on OCT reports that may not yet be recognized by clinical research, enabling broader application of such AI approaches in healthcare.

Conclusions

Our study aimed to predict the AOIs in OCT images by analyzing the eye tracking behavior of ophthalmologists with varying levels of experience. The results showed that the performance of the model was influenced by factors including the expertise of participants at glaucoma diagnosis, the types of OCT devices that generated the reports, the number of participants included in the analysis, and the model architecture. The best-performing model had a U-Net backbone, was trained on expert 10 control data, and achieved a precision of 0.723, recall of 0.562, and an f1-score of 0.609. This model was trained on clinicians with the most experience (glaucoma faculty). The study found that the AI model was able to predict the fixations made by expert participants on OCT reports, which could potentially aid in the training of self-supervised eye movement-informed AI systems, could shed light on new ocular biomarkers by objectively showing AI-predicted AOIs learned from expert eye movements, and could offer new eye movement–based medical education paradigms.

Acknowledgments

Supported in part by an unrestricted grant to the Columbia University Department of Ophthalmology from Research to Prevent Blindness, Inc., New York, New York, USA.

Disclosure: **Y. Tian**, None; **A. Sharma**, None; **S. Mehta**, None; **S. Kaushal**, None; **J.M. Liebmann**, None; **G.A. Cioffi**, None; **K.A. Thakoor**, None

References

1. Quigley HA. Glaucoma. *Lancet*. 2011;377(9774): 1367–1377.
2. Berisha F, Hoffmann EM, Pfeiffer N. Optical Coherence Tomography in Glaucoma. In: Bernardes R, Cunha-Vaz J, eds. *Optical Coherence Tomograph: Biological and Medical Physics, Biomedical Engineering*. Berlin, Heidelberg: Springer; 2012.
3. Moraru AD, Costin D, Moraru RL, Branisteanu DC. Artificial intelligence and deep learning in ophthalmology - present and future (Review). *Exp Ther Med*. 2020;20(4):3469–3473.
4. Fang H, Shang F, Fu H, Li F, Zhang X, Xu Y. Multi-modality images analysis: a baseline for glaucoma grading via deep learning. In: Fu H, Garvin MK, MacGillivray T, Xu Y, Zheng Y, eds. *Ophthalmic Medical Image Analysis. OMIA. Lecture Notes in Computer Science*, vol. 12970. Cham, Berlin, Heidelberg: Springer; 2021.
5. Thakoor KA, Li X, Tsamis E, Sajda P, Hood DC. Enhancing the accuracy of glaucoma detection from OCT probability maps using convolutional neural networks. *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2019:2036–2040.
6. Muhammad H, Fuchs TJ, De Cuir N, et al. Hybrid deep learning on single wide-field optical coherence tomography scans accurately classifies glaucoma suspects. *J Glaucoma*. 2017;26(12):1086–1094.
7. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24(9):1342–1350.
8. Asaoka R, Murata H, Hirasawa K, et al. Using deep learning and transfer learning to accurately diagnose early-onset glaucoma from macular optical coherence tomography images. *Am J Ophthalmol*. 2019;198:136–145.
9. García G, del Amor R, Colomer A, Naranjo V. Glaucoma detection from raw circumpapillary OCT images using fully convolutional neural networks. *arXiv*. 2020. preprint arXiv:2006.00027.
10. Lévêque L, Bosmans H, Cockmartin L, Liu H. State of the art: eye-tracking studies in medical imaging. *IEEE Access*. 2018;6:37023–3703.
11. Kundel HL. Visual search and lung nodule detection on CT scans. *Radiology*. 2015;274(1): 14–16.
12. Drew T, Vo ML, Olwal A, Jacobson F, Seltzer SE, Wolfe JM. Scanners and drillers: characterizing expert visual search through volumetric images. *J Vis*. 2013;13(10):3:1–13.
13. Manning D, Ethell S, Donovan T, Crawford T. How do radiologists do it? The influence of experience and training on searching for chest nodules. *Radiography*. 2006;12(2):134–142.
14. Brunyé TT, Nallamothu BK, Elmore JG. Eye-tracking for assessing medical image interpretation: a pilot feasibility study comparing novice vs expert cardiologists. *Perspect Med Educ.* 2019;8:65–73.
15. Li L, Xu M, Wang X, Jiang L, Liu H. Attention based glaucoma detection: a large-scale database and CNN model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019;10571–10580.
16. Krupinsky EA. On the development of expertise in interpreting medical images. *Proceedings of the SPIE 8291, Human Vision and Electronic Imaging XVII*, 82910R, 2012.
17. Brunyé TT, Trafton D, Kerr KF, Shucard H, Weaver DL, Elmore JG. Eye tracking reveals expertise-related differences in the time-course of medical image inspection and diagnosis. *J Med Imaging.* 2020:7(5):051203.
18. Crowe EM, Gilchrist ID, Kent C. New approaches to the analysis of eye movement behaviour across

expertise while viewing brain MRIs. *Cogn Res*. 2018;3:12.

19. Trebing K, Stañczyk T, Mehrkanoon S. SmaAt-U-Net: precipitation nowcasting using a small attention-U-Net architecture. *Pattern Recognit Lett*. 2021;145:178–186.
20. Sharma M, Sharma A, Tushar KR, Panneer A. A novel 3D-U-Net deep learning framework based on high-dimensional bilateral grid for edge consistent single image depth estimation. *2020 International Conference on 3D Immersion (IC3D)*. IEEE. 2020; 01–08.
21. Lou J, Lin H, Marshall D, Saupe D, Liu H. TranSalNet: towards perceptually relevant visual saliency prediction. *Neurocomputing*. 2022;494: 455–467.
22. Leshno A, Tsamis E, Hirji S, et al. Detecting established glaucoma using OCT alone: utilizing an OCT reading center in a real-world clinical setting. *Transl Vis Sci Technol*. 2024;13(1):4. Accepted November 27, 2023.
23. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
24. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv*. 2014. preprint arXiv:1409.1556.
25. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. *arXiv*. 2015.
26. Deng J, Dong W, Socher R, et al. Imagenet: a large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*. *arXiv*. 2009.
27. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. *arXiv*. 2015.
28. Liebmann JM, Hood DC, de Moraes CG, e al. Rationale and development of an OCT-based method for detection of glaucomatous optic neuropathy. *J Glaucoma*. 2022;31(6):375–381.
29. Kaushal S, Sun Y, Zukerman R, Chen RWS, Thakoor KA. Detecting eye disease using vision transformers informed by ophthalmology resident gaze data. *IEEE Engineering in Medicine and Biology Conference (EMBC)*. 2023.