Detecting Glaucoma Using 3D Convolutional Neural Network of Raw SD-OCT Optic Nerve Scans

Erfan Noury*†§[1,6], Suria S. Mannil†‡[2], Robert T. Chang†♮[2], An Ran Ran[3], Carol Y. Cheung[3], Suman S. Thapa[4], Harsha L. Rao[5], Srilakshmi Dasari[5], Mohammed Riyazuddin[5], Sriharsha Nagaraj[5], and Reza Zadeh†♯[1,7]

[1]Matroid
[2]Byers Eye Institute, Stanford University, Palo Alto, CA, United States
[3]Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Hong Kong
[4]Tilganga Institute of Ophthalmology, Kathmandu, Nepal
[5]Narayana Nethralaya Foundation, Bangalore, India
[6]University of Maryland at Baltimore County, Baltimore, MD, United States
[7]Stanford University, Stanford, CA, United States
§erfan1@umbc.edu, ‡suria1@stanford.edu, ♮rchang3@stanford.edu, ♯reza@matroid.com

1 Abstract

Background. Glaucoma is a chronic progressive optic neuropathy with characteristic visual field defects and corresponding structural changes, including nerve fiber layer thinning and optic nerve neuroretinal rim loss. These changes are traditionally monitored by SD-OCT (Spectral Domain Optical Coherence Tomography), which contains a large amount of 3D voxel information in a 6mm × 6mm × 2mm cube of data. However, only a fixed 3.4 mm diameter circle (2D slice) centered over the optic nerve is currently extracted using automated segmentation of the retinal nerve fiber layer thickness (RNFL). This RNFL thickness is reported relative to a normative database to help detect thinning and neuroretinal rim loss, which does not use the additional information in the optic nerve head cube. Clinicians rarely scroll through the entire cube. Therefore we propose developing and validating a three-dimensional (3D) deep learning system using the entire unprocessed OCT optic nerve volumes to distinguish true glaucoma from normals in order to discover any additional imaging biomarkers within the cube through saliency mapping. The algorithm has been validated against 4 additional distinct datasets from different countries using multimodal test results to define glaucoma rather than just the OCT alone. We hypothesize that the output from this 3D model, alongside a map of the regions where the model attends to make a prediction, can help identify novel diagnostic information in the cube.

Methods. 2076 OCT (Cirrus SD-OCT, Carl Zeiss Meditec, Dublin, CA) 6 mm cubes centered over the optic nerve, 200 × 200 × 1024 volumes of 879 eyes (390 healthy and 489 glaucoma) from 487 patients, age 18-84 years, were exported from the Glaucoma Clinic Imaging Database at the Byers Eye Institute, Stanford University, from March 2010 to December 2017. This included bilateral eyes of 391 patients and unilateral eyes of 97 patients with a right eye to left eye ratio of 1.05:1. A 3D deep neural network was trained and tested on this unique OCT optic nerve head dataset from Stanford. 570 randomly selected optic nerve head cube scans of eyes with a diagnosis of glaucoma (*True Glaucoma*) and 342 scans of eyes with a normal diagnosis (*True Normal*) were used for training. A total of 81 scans of eyes with *True Glaucoma* and 32 scans of eyes with *True Normal* annotations were included in the primary validation set. 58 scans of eyes with *True*

*Work done during internship at Matroid.
†These authors contributed equally.

Glaucoma annotation and 50 scans of eyes with a *True Normal* annotation were included in the test set. A total of 3620 scans (all obtained using the Cirrus SD-OCT device) from 1458 eyes obtained from 4 different institutions, from United States (943 scans), Hong Kong (1625 scans), India (672 scans), and Nepal (380 scans) were used for external evaluation. *True Glaucoma* for the training data was defined as glaucomatous disc changes along with defects on SD-OCT RNFL and/or GCIPL (thickness and/or deviation) maps with corresponding visual field defects as well as intraocular pressure lowering treatment upon chart review. The range of glaucoma patients included mild to severe without excluding high myopes. True normal was defined as cases with non-glaucomatous optic disc with no structural defects on OCT RNFL/GCIPL deviation or sector maps and normal visual fields upon chart review.
Results. The 3D deep learning system achieved an area under the receiver operation characteristics curve (AUROC) of 0.8883 in the primary Stanford test set identifying true normal from true glaucoma. The system obtained AUROCs of 0.8571, 0.7695, 0.8706, and 0.7965 on OCT cubes from United States, Hong Kong, India, and Nepal, respectively. We also analyzed the performance of the model separately for each myopia severity level as defined by spherical equivalent and the model was able to achieve F1 scores of 0.9673, 0.9491, and 0.8528 on severe, moderate, and mild myopia cases, respectively. Saliency map visualizations highlighted a significant association between the optic nerve lamina cribrosa region in the glaucoma group.
Conclusions. A 3D convolutional neural network using SD-OCT optic nerve head cubes can distinguish true glaucoma from normal with good accuracy and this generalized to multiple diverse external SD-OCT datasets. Highlighted areas from saliency mapping revealed new areas within the deep lamina cribrosa. This deserves further investigation, as there is potential to monitor laminar changes even after RNFL has thinned.

2 Introduction

Glaucoma is a chronic degenerative disease that eventually leads to optic nerve damage and is one of the leading causes of blindness worldwide[1, 2]. Currently, the main modifiable risk factor is elevated intraocular pressure (IOP), which, in combination with structural and functioning longitudinal imaging, is used as one of the main diagnostic parameters. Spectral Domain Optical Coherence Tomography (SD-OCT) provides high-resolution cross-sectional imaging of the macula and optic nerve head, which is powerful for detecting the presence of glaucoma as well as glaucoma progression. OCTs operate on the principle of laser constructive and destructive interferometry, so they are similar to ultrasound reflectivity, except using light instead of sound. Glaucomatous structural changes include RNFL and ganglion cell inner plexiform layer (GCIPL) thinning, but norms need to be age adjusted and refractive error adjusted, and patients need to be compared to themselves over time. It is also known that glaucoma damage extends deep into the optic nerve head at the level of the lamina cribrosa (LC), a network of columns supporting the neuronal axon connections as they traverse from the surface of the retina to the visual cortex of the brain (see Figure 1) [3]. However, only qualitative enhanced depth imaging EDI SD-OCT protocols have been able to visualize these changes in the past. Based on current understanding of high pressure induced glaucoma, biomechanical deformation and remodeling of the ONH leads to posterior displacement of the lamina cribrosa relative to the sclera as well as progressive loss of ganglion cell axons and cell bodies, resulting in RNFL thinning [3]. Thus it seems reasonable to hypothesize that there is additional information in a standard SD-OCT optic nerve head cube which currently is not being tracked, but could be discovered through deep learning pattern recognition as a differentiator of true glaucoma from normal.

The current OCT ONH output is an RNFL map algorithm (Carl Zeiss Meditec, Inc., Dublin, CA, USA) representing a 6mm $\times$ 6mm $\times$ 2mm cube of A-scan data centered over the optic nerve in which a 3.4 mm diameter circle of RNFL data is extracted to create a TSNIT 2D map (temporal, superior, nasal, inferior, temporal). Thickness data from the 2D map is displayed as a four color scale referenced to an age-adjusted normative database. The RNFL and GCIPL parameters in the OCT summary reports are represented by a white color for those in the hyper-normal range (95^{th} to 100^{th} percentiles), green backgrounds for those in the normal range (5^{th} to 95^{th} percentiles), yellow backgrounds for those abnormal at the 1^{st} to 5^{th} percentile level, and red backgrounds for those abnormal at the 1^{st} percentile level [4]. The normative database for the Cirrus SD-OCT consists of 284 healthy individuals with an age range between 18 and 84 years (mean age of 46.5 years). Ethnically, 43% were Caucasian, 24% were Asian, 18% were African American, 12% were Hispanic, 1% were Indian, and 6% were of mixed ethnicity. The refractive error ranged from -12.00 D to $+8.00$ D [5]. Due to the relatively small normative database, there are a lot of false positive results from high myopia disc changes, or thin RNFL due to other non-glaucomatous or artifactual reasons. One of the difficulties in diagnosing glaucoma is that there is no single test with a high sensitivity and specificity to

confirm the diagnosis. So, to improve upon the color code labeling of the OCT cubes, we decided to use multimodal ground truth to provide a more accurate diagnosis, including fundus photo appearance, visual fields, and clinical history (intraocular pressure and treatment information), to more accurately confirm true glaucoma. In a typical glaucoma patient evaluation workflow, multiple tests are acquired, similar to the diagram in Figure 2. This demonstrates the complexity of true glaucoma diagnosis and a lack of single biomarker currently. In this work, we concentrate on training a 3D neural network which can predict whether an optic nerve head cube scan belongs to the normal (*True Normal*) versus glaucoma (*True Glaucoma*) class based on better ground truth and and hold out of glaucoma suspects.

3 Data

The study adhered to the tenets of the Declaration of Helsinki, and the protocols were approved by the respective institutional review boards of Stanford School of Medicine (United States), Chinese University of Hong Kong (Hong Kong), Narayana Nethralaya Foundation (India), and Tilganga Institute of Ophthalmology (Nepal). Funding to extract data and store it in a de-identified, encrypted cloud storage was supported by Santen, Inc and Stanford Global Health Seed Grant by Stanford Center for Innovation in Global Health (CIGH). Informed consent was waived based on the study's retrospective design, anonymized dataset of OCT images and test data, minimal risk, and confidentiality protections.

3D OCT cube (volume) ONH images (Cirrus HD-OCT, Carl Zeiss Meditec, Dublin,CA) of 1741 eyes of 978 patients evaluated at the Byers Eye Institute, Stanford School of Medicine, from March 2010 to December 2017 were exported for the study. Prior to labeling as *True Glaucoma* versus *True Normal*, based on chart review, 749 eyes were excluded due to the presence of other ocular pathologies and 93 eyes were excluded due to the presence of OCT artifacts or due to signal strength being less than 3, as per exclusion criteria mentioned below. 20 eyes were excluded after arbitration among 2 glaucoma specialists in which a consensus of definite glaucoma could not be made from the chart. Finally 879 eyes of 487 patients (2076 scans) were labelled and used for training, primary validation, primary testing, and external testing from Stanford.

The inclusion criteria were (1) age equal to or older than 18 years old; (2) reliable visual field (VF) tests (acceptable results defined below); and (3) availability of SD-OCT Optic Disc scans (acceptable scans defined below). A reliable visual field report is defined as (a) fixation losses less than 33%; (b) false positive rate less than 25%; (c) false negative rate less than 25%; and (d) no appearance of lid or lens rim artifacts, and no appearance cloverleaf patterns. SD-OCT scans with signal strength less than 3 or any artifact obscuring imaging of the ONH, or any artifacts or missing data areas that prevented measuring the thickness of the RNFL at 3.4 mm diameter, were excluded from the study. Artifacts included blink, motion, registration, and mirror artifacts. The reason a signal strength of < 6 was included is because the entire cube of data was being used and not the results from the machine's segmentation algorithm (which often fails at low signal strength).

True Glaucoma was defined as those eyes with glaucomatous disc changes [6] on fundus examination, with localized defects on OCT RNFL/GCIPL deviation or sector maps, that correlated with the VF defect which fulfilled the minimum definition of Hodapp-Anderson-Parrish (HAP) glaucomatous VF defect and are on or had intraocular pressure lowering treatment as per chart review [7]. Thus no pre-perimetric glaucoma were included. *True Normal* was defined as non-glaucomatous optic disc on fundus exam with no structural defects on OCT RNFL/GCIPL deviation or sector map and normal visual fields, and normal intraocular pressures.

Eyes with optic nerve head pathologies, such as non-glaucomatous optic neuropathy, optic nerve head hypoplasia, or optic nerve pit, and other retinal pathologies such as retinal detachment, age-related macular degeneration, myopic macular degeneration, macular hole, diabetic retinopathy, and arterial and venous obstruction were carefully excluded.

3.1 Training, Primary Validation, and Test Sets

The initial dataset was randomly split into three sets for patients. In total, 570 optic nerve scans of 229 eyes from 121 patients with a diagnosis of glaucoma (*True Glaucoma*) (randomly chosen), and 342 scans of 200 eyes from 112 patients of definitive normal (*True Normal*) were included in the training set. A total of 81 scans of 33 eyes from 17 patients with a *True Glaucoma* annotation, and 32 scans of 22 eyes from 14 patients with a *True Normal* annotation were included in the primary validation set. Similarly, 157 scans of 54 eyes from 29 patients with a *True Glaucoma* annotation, and 50 scans of 25 eyes from 14 patients with

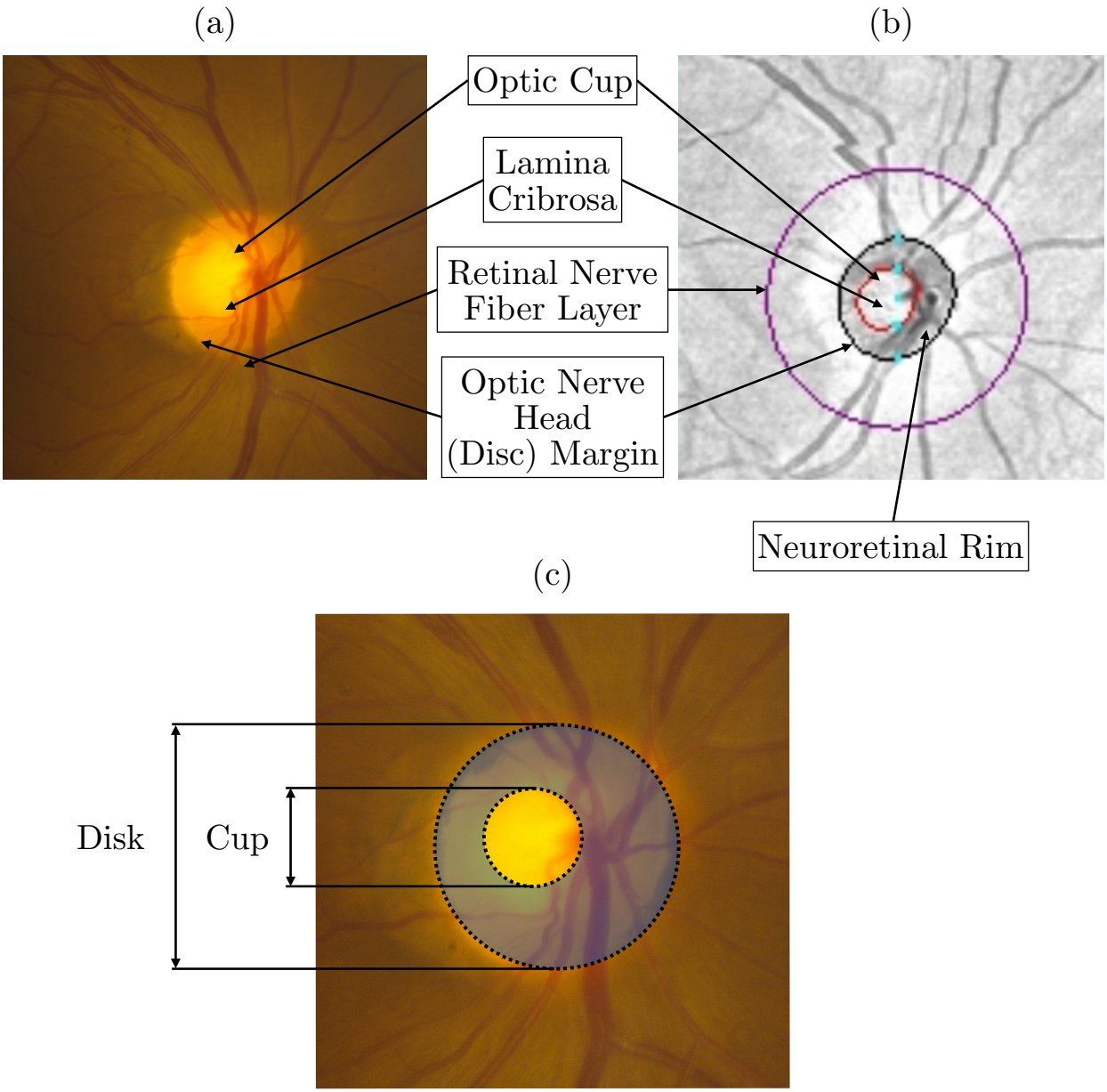


Figure 1: Primer on Optic Nerve Head (ONH) Morphology. **(a)** Color Fundus Image of the Optic Disc. **(b)** Enface OCT image reconstruction of the Optic Nerve Head. Retinal Nerve Fibers converge at the ONH (known as the optic disc boundary, marked in black in (b)) and then exit the eye as the optic nerve. The ONH consists of retinal nerve fibers from the Retinal Ganglion Cell axons leading into a central depression known as the optic cup (boundary marked in red in **(b)**) and a collagenous structure, known as the Lamina Cribrosa, which provides physical support to the exiting axon fibers. Neuroretinal Rim is the retinal nerve fiber tissue between the border of the cup and the disc. Optic disc cupping, characterized by progressive neuroretinal rim thinning, is a result of an increased ratio between the optic cup and disc, called the vertical cup-to-disc ratio **(c)**, a classic feature in glaucoma. Lamina Cribrosa forms the bottom of the optic cup on the inner surface of the ONH.

a *True Normal* annotation were included in the test set. The split was based on patients, so as to make sure that scans belonging to each patient are included in only one of the splits. Each OCT scan over the optic nerve head is a three-dimensional array of size 6mm $\times$ 6mm $\times$ 2mm divided into a cube of resolution of 200 $\times$ 200 $\times$ 1024, with numbers representing the height, width, and depth of the array, respectively. A

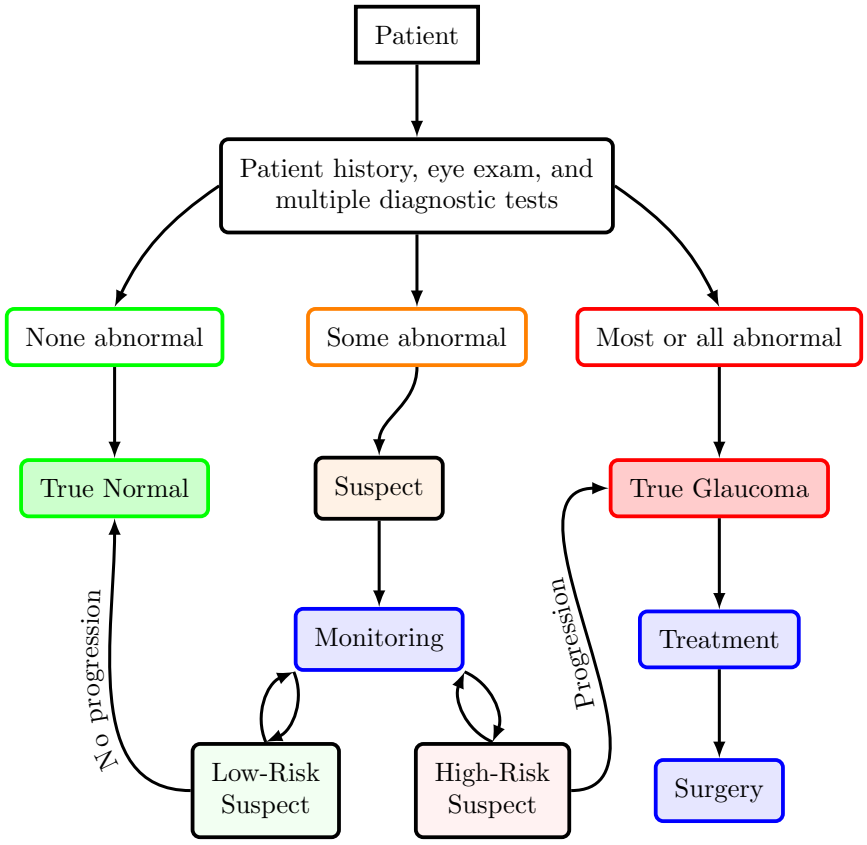


Figure 2: Glaucoma screening procedure diagram.

downsampled version of this three-dimensional array (of size $100 \times 100 \times 128$) is given to the deep neural network as input and the probability of each of the *True Normal* or *True Glaucoma* classes is predicted. For the dataset from Stanford, two Glaucoma fellowship trained Ophthalmologists (RC and SM) worked separately to label all the eyes with SD-OCT scans. Images were labeled as per criteria (Table 1) into *True Glaucoma* and *True Normal*. 36 disagreeable cases were reviewed again by both glaucoma specialists to make the final decision and 20 cases out of the 36 were eliminated based on difficulty determining *True Glaucoma*.

3.2 External Test Sets

Four datasets were used for external evaluation of the model. The SD-OCTs of all the 4 external datasets were acquired using Cirrus HD-OCT (Carl Zeiss Meditec, Dublin, CA, USA) according to the optic disc cube scanning protocol. **Dataset A** is composed of 943 additional OCT 3D cube images from the Glaucoma Clinic at Stanford University that were annotated after the initial set was annotated and used for training, validation, and primary testing. Of those, 297 OCT 3D cube volumes were from 143 eyes (of 85 patients) that were labeled as *True Normal*, and 646 scans of 207 eyes (of 124 patients) were labeled as *True Glaucoma*. **Dataset B** consists of 1625 OCT 3D cube images from Chinese University of Hong Kong, with 666 OCT 3D cubes of 196 eyes (of 99 patients) labeled as *True Normal*, and 959 OCT 3D cubes of 277 eyes (of 155 patients), labeled as *True Glaucoma*. **Dataset C** is composed of 672 OCT 3D cube images of ONH from Narayana Nethralaya Foundation, India. 211 scans from 147 eyes of 98 patients were labeled as *True Normal* and 461 OCT 3D cubes from 171 eyes of 101 patients had a *True Glaucoma* annotation. Finally, **Dataset D** contained 380 OCT 3D cube images of ONH from the Tilganga Institute of Ophthalmology, Nepal. In this dataset, 158 scans from 143 eyes of 89 patients were labeled as *True Normal*, and 222 scans from 174 eyes of 109 patients were labeled as *True Glaucoma*.

For SD-OCT data from Hong Kong (Dataset B), two trained medical students and a postgraduate ophthalmology trainee (with more than 3 years' of experience in glaucoma) did the initial quality control and then graded the SD-OCT scans into gradable or non-gradable SD-OCT scans, according to the aforementioned criteria. Two glaucoma specialists then worked separately to label all the eyes with gradable SD-OCT

Table 1: Criteria used for labelling cases as *True Glaucoma* and *True Normal*.

Labels	**True Glaucoma**	**True Normal**
Criteria	• Clinical Glaucomatous Disc changes (as per ISGEO classification [6]), *and* • OCT Glaucomatous defects on deviation maps and not all green on OCT RNFL and/or OCT GCIPL maps, *and* • 2 repeatable VF defects as per HAP criteria [7]. Reliably measured data were used, *i.e.* with a fixation loss $< 20\%$, false positive errors $< 15\%$, and false negative errors $< 33\%$, *or* total cupping of the optic nerve and unable to perform VF evaluation, *and* • On Treatment for Glaucoma or has undergone surgery/SLT-ALT.	• No disc changes for glaucoma (few cases have high cup disc ratio > 0.6 but no other glaucomatous disc changes), *and* • No OCT glaucomatous defects on deviation maps and all green OCT RNFL and OCT GCIPL maps, *and* • No visual field defects, *and* • No treatment/review after a duration no lesser than a year as per chart review.

Table 2: Demographic background of the training, primary validation, and primary test sets (from Stanford).

	True Glaucoma	**True Normal**
Age	69.41 (±14.70)	61.84 (±15.20)
Gender (F:M)	47%:53%	60%:40%
Asian	136 (n)	119 (n)
Caucasian	102 (n)	74 (n)
African American	14 (n)	16 (n)
Hispanic	21 (n)	24 (n)
Data of ethnicity unavailable	10 (n)	14 (n)
Average MD	-9.75 (±7.50)	-0.79 (±1.20)
Mean Refractive Error	-3.57 (±3.37)	-2.20 (±2.34)

Table 3: Demographic background of the external test set from Stanford (Dataset A).

	True Glaucoma	**True Normal**
Age	69.82 (±16.15)	63.00 (±16.93)
Gender (F:M)	41%:59%	60%:40%
Asian	85 (n)	78 (n)
Caucasian	87 (n)	50 (n)
African American	14 (n)	4 (n)
Hispanic	21 (n)	4 (n)
Data of ethnicity unavailable	0 (n)	0 (n)
Average MD	-9.01 (±7.52)	-0.79 (±0.98)
Mean Refractive Error	-2.64 (±2.86)	-1.92 (±2.03)

Table 4: Demographic background of the external test set from Hong Kong (Dataset B), such as gender and ethnicity distribution, and mean values (standard deviations) for visual field parameter mean deviation (MD) and Mean Refractive error.

	True Glaucoma	**True Normal**
Age	65.90 (±9.30)	61.05 (±8.50)
Asian	277 (n)	196 (n)
Female:Male	70%:30%	No Data
Average MD	-8.005 (±6.81)	-0.900 (±1.30)
Mean Refractive Error	-0.85 (±2.57)	-0.51 (±2.15)

Table 5: Demographic background of the external test set from India (Dataset C), such as gender and ethnicity distribution, and mean values (standard deviations) for visual field parameter mean deviation (MD) and Mean Refractive error.

	True Glaucoma	**True Normal**
Age	63.84 (±11.72)	54.76 (±14.95)
Asian	173 (n)	130 (n)
Female:Male	38%:62%	40%:60%
Average MD	-12.74 (±9.22)	-2.10 (±1.30)
Mean Refractive Error	-0.483 (±2.25)	-0.440 (±2.19)

Table 6: Demographic background of the external test set from Nepal (Dataset D), such as gender and ethnicity distribution, and mean values (standard deviations) for visual field parameter mean deviation (MD) and Mean Refractive error.

	True Glaucoma	**True Normal**
Age	45.34 (±17.08)	39.17 (±12.28)
Asian	184 (n)	173 (n)
Female:Male	44%:56%	31%:69%
Average MD	-8.30 (±7.04)	-2.32 (±1.47)
Mean Refractive Error	-1.38 (±2.38)	-1.17 (±1.36)

Table 7: Distribution of cases in terms of Glaucoma severity. Classification based on Mean Deviation (Severe: MD ≤ -12, Moderate: $-12 <$ MD ≤ -6, Mild: $-6 <$ MD).

	Primary (Stanford)	**Dataset A**	**Dataset B**	**Dataset C**	**Dataset D**
Severe Glaucoma	27.20%	28.40%	24.00%	44.80%	21.10%
Moderate Glaucoma	26.80%	18.93%	26.10%	17.20%	22.76%
Mild Glaucoma	45.50%	52.66%	49.70%	37.90%	56.10%

Table 8: Comparison of myopia severity between the primary test set, and Datasets A, B, C, and D. **TG** stands for *True Glaucoma* and **TN** stands for *True Normal*. Chi-squared test was used for severe myopia distribution analysis (Myopia severity distribution: Severe: $D \leq -6$, Moderate: $-6 < D \leq -3$, Mild: $-3 < D$, where D is diopter).

Subset	Severe Myopia	Moderate Myopia	Mild Myopia	Emmetropia	Hypermetropia
Primary (TG)	11.00%	14.21%	34.80%	8.50%	31.70%
Primary (TN)	4.09%	6.14%	31.57%	8.40%	49.70%
Dataset A (TG)	8.88% ($p = 0.70$)	8.10%	42.20%	11.11%	20.00%
Dataset A (TN)	4.20% ($p = 0.98$)	10.08%	31.09%	5.88%	47.89%
Dataset B (TG)	4.70% ($p = 0.12$)	12.50%	37.50%	5.90%	39.20%
Dataset B (TN)	0.0% ($p < 0.001$)	21.01%	15.70%	10.50%	47.30%
Dataset C (TG)	0.0% ($p < 0.001$)	3.94%	43.20%	22.30%	38.10%
Dataset C (TN)	0.0% ($p < 0.001$)	16.60%	30.30%	15.15%	37.87%
Dataset D (TG)	2.50% ($p < 0.001$)	14.28%	43.80%	10.70%	25.00%
Dataset D (TN)	0.0% ($p < 0.001$)	6.38%	53.00%	0.0%	40.40%

Table 9: Comparison of additional clinical data between the primary set and four external evaluation datasets. The statistical analysis was performed with the Statistical Package for Social Sciences (SPSS) 10.1 (SPSS Inc., Chicago, IL, USA). Results are expressed as mean (± standard deviation) and paired Student's t-test was used to evaluate the level of significance. A p-value of 0.001 or less was considered significant. Chi square test was used for comparisons of categorical demographic data for proportions. **TG** stands for *True Glaucoma* and **TN** stands for *True Normal*.

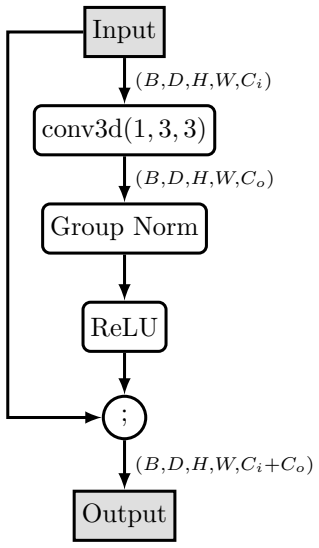
Subset	Cup-Disc Ratio	IOP	Gender Distribution	PSD	VFI
Primary (TG)	0.80 (±0.12) $p < 0.001$	20.07 (±4.75)	55:45	7.71 (±6.66)	74.4% ($p < 0.001$)
Primary (TN)	0.46 (±0.16) ($p < 0.001$)	15.67 (±2.72)		1.83 (±0.53)	98.46% ($p < 0.001$)
Dataset A (TG)	0.79 (±0.19) $p = 0.5262$	19.56 (±5.47) $p = 0.2739$	49:51 ($p = 0.2194$)	6.37 (±4.46) $p = 0.0413$	77.01% ($p = 0.6191$)
Dataset A (TN)	0.45 (±0.16) $p = 0.4764$	16.00 (±2.72) $p = 0.2475$		2.12 (±1.07) $p = 0.0413$	98.06% ($p = 0.7678$)
Dataset B (TG)	No Data	16.19 (±4.17) $p < 0.001$	67:33 ($p = 0.0048$)	6.44 (±4.21) $p = 0.0076$	79.83% ($p = 0.5239$)
Dataset B (TN)	No Data	13.44 (±2.72) $p < 0.0001$		1.46 (±0.30) $p = 0.0076$	99.61% ($p = 0.2346$)
Dataset C (TG)	No Data	No Data	40:60 ($p = 0.0031$)	7.68 (±3.81) $p = 0.9640$	65.38% ($p = 0.0331$)
Dataset C (TN)	No Data	No Data		2.54 (±1.39) $p = 0.9640$	93.17% ($p = 0.0068$)
Dataset D (TG)	No Data	16.56 (±4.74) $p < 0.001$	40:60 ($p = 0.0031$)	5.37 (±3.30) $p < 0.001$	77.00% ($p = 0.5791$)
Dataset D (TN)	No Data	15.68 (±2.90) $p = 0.9701$		1.99 (±1.08) $p < 0.001$	97.58% ($p = 0.5362$)

scans into yes/no glaucoma combined with VF results. In this dataset glaucoma was defined as RNFL defects on thickness or deviation maps that correlated in position with the VF defect which fulfilled the definition of glaucomatous VF defects [7]. Most of the images were labelled as yes/no glaucoma when the two graders arrived at the same categorization separately, but a few disagreeable cases were reviewed by a senior Glaucoma specialist to make the final decision.

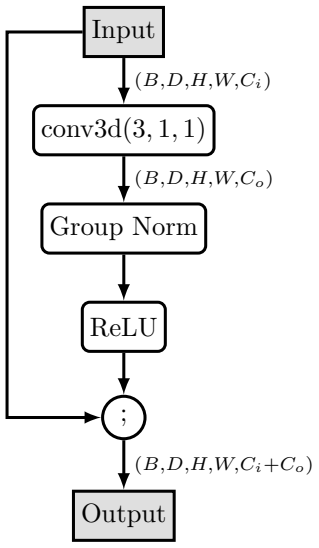
For external evaluation set from India (Dataset C) and Nepal (Dataset D), glaucoma specialists each with experience of more than 10 years in Glaucoma labeled the cases into *True Glaucoma* and *True Normal*. Definitions of *True Glaucoma* and *True Normal* in this dataset were similar to those used at Stanford (Table 1).

4 Method

The deep neural network used in this work is based on the "End-to-End Classification Network" of De Fauw *et al.* [8]. This network uses multiple layers of dense convolutional blocks [9] that is applied to 3D volumes of OCT scans. Each dense convolutional block consists of one 3D spatial convolutional block (Figure 3a) followed by a 3D depth-wise convolutional block (Figure 3b). Each convolutional block applies a convolutional operation, followed by normalization and non-linearity to the input, and the output is concatenated to the input of the convolutional block along the channel dimension.



(a) Spatial 3D Convolutional block.



(b) Depth-wise 3D convolutional block.

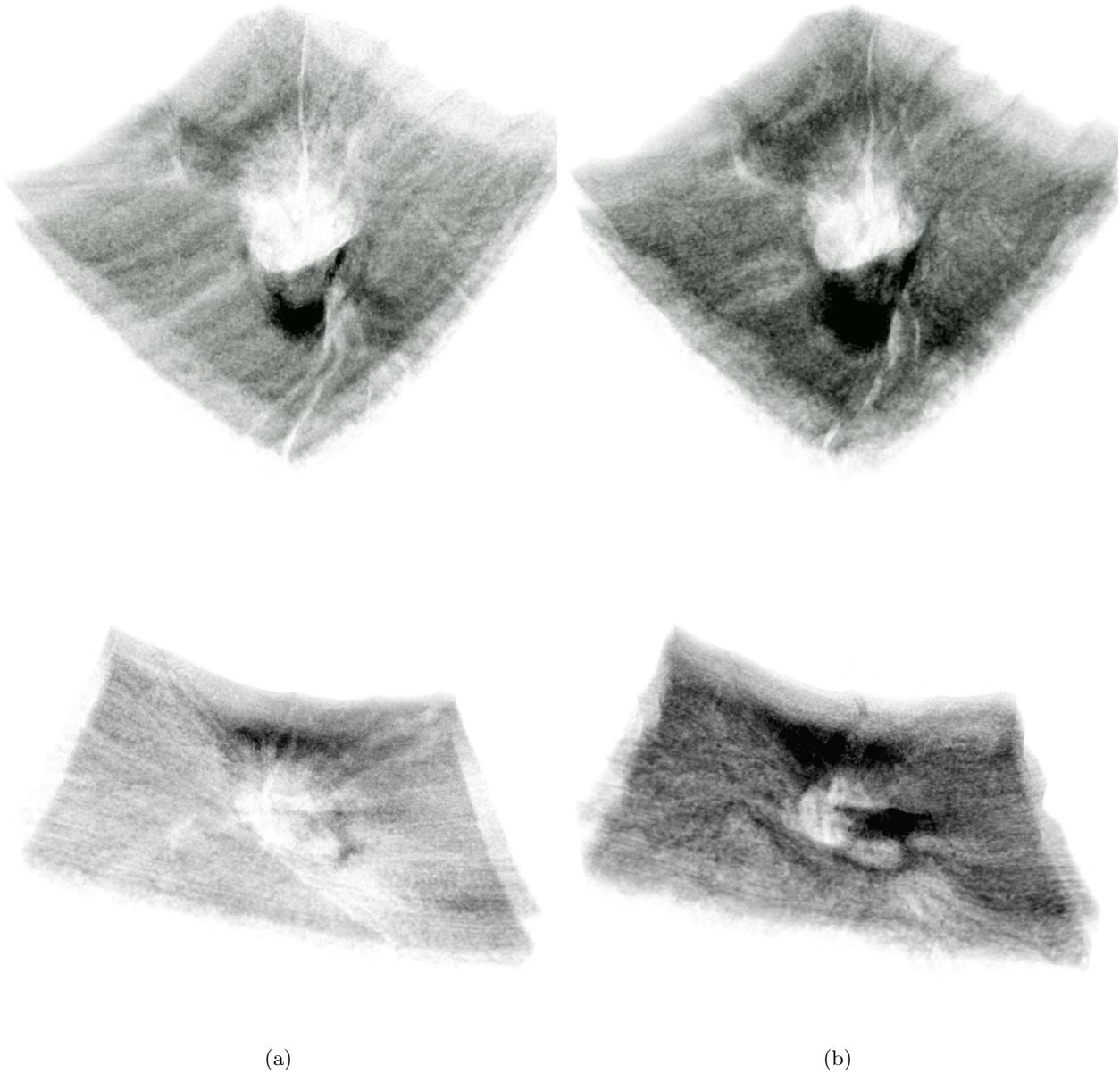
Figure 3: Building blocks of the dense convolutional blocks used in the convolutional neural network.

Different from De Fauw *et al.* [8], Group Normalization [10] is used instead of Batch Normalization [11]. This modification was necessary, due to the fact that the network could not be trained using Batch Normalization. To increase the amount of effective training data, random flipping and dense elastic deformation was used as data augmentation during training (see Figure 4). Adam optimizer with weight decay [12] was used for training. After training, model checkpoint with the best results on the validation set was selected as the final model.

For saliency visualization, the Grad-CAM method [13] was combined with the Guided Backpropagation [14] to generate better and more clear visualizations. Gradients of the predicted class were obtained with respect to the input and the middle layers.

Area under the curve (AUC) and F1 scores have been used to quantify the performance of machine learning models. The F1 score is a measure of a model's accuracy and is defined as the weighted harmonic mean of the model's precision and recall. In a binary classification model (like the proposed model in this paper), different discrimination threshold values will result in different values of precision and recall, due to changing values of true positive, true negative, false positive, and false negative. The Area under the (Receiver Operating) Curve summarizes the performance of the binary classifier for different values of discrimination

threshold. AUC is also a measure of the probability of the binary classifier giving a random positive sample a higher probability of belonging to the positive class compared to a random negative data point [15].



(a) (b)

Figure 4: **(a)** Original OCT scans. **(b)** Elastic Deformation applied to the OCT scans. Darker regions are tissues in the eye that are less transparent against the light beamed to the eye.

5 Results

Demographic background of the combined training, primary validation, primary test sets are presented in Table 2. The demographic data includes age, gender, and ethnicity distribution, visual field mean deviation (MD), and mean refractive error as these are parameters known to affect the OCT cube tissue thicknesses independent of glaucoma. Note that for some patients, demographic data was incomplete and therefore, aggregate numbers do not necessarily add up to the dataset size. Demographic information for the external test sets from Stanford (Dataset A), Hong Kong (Dataset B), India (Dataset C), and Nepal (Dataset D) are presented in Table 3, Table 4, Table 5, Table 6, respectively.

Among the training, primary validation, and primary test set, and datasets A, B, and C, there was no significant difference in the average age ($p > 0.001$), but the average age of patients in Dataset D was significantly lower than the other datasets ($p < 0.001$). There was significant difference in the mean refractive error between the *True Glaucoma* and *True Normal* subsets in the training/primary validation/primary test set compared to Dataset B (Hong Kong), Dataset C (India), and Dataset D ($p < 0.001$), while there was no significant difference with Dataset A ($p > 0.001$). The distribution of cases according to severity of refractive error is shown in Table 8. There is significantly higher percentage of severe myopia cases in the *True Glaucoma* subset in the training/primary validation/primary test set, Dataset A, and Dataset B, compared to Dataset C and Dataset D. Also there is significantly higher number of severe myopia in the *True Normal* subset of training/primary validation/primary test sets and Dataset A, compared to Datasets B, C, and D ($p < 0.001$). There was no significant difference in severity of glaucoma between the training/primary validation/primary test set, Dataset A ($p = 0.2821$), Dataset B ($p = 0.004$) and Dataset D ($p = 0.0385$) while it was significant compared to the Dataset C ($p < 0.001$). The percentage of severe glaucoma cases in Dataset C was significantly higher ($p < 0.001$) compared to training set/primary validation/primary test set, Dataset A, Dataset B, and Dataset D (Table 7). Severity distribution of datasets from United States, Hong Kong, India, and Nepal are shown in Table 7. Details of additional clinical information such as cup-to-disc ratio, IOP, gender distribution, pattern standard deviation (PSD), and visual field index (VFI) are shown in Table 9.

Our model achieved a peak test AUC of 0.8883 and F1 score of 0.8834 to differentiate between healthy and normal eyes. The results of the model on the primary test set are shown in the "Primary Test" row of Table 10. Additionally, the same model was also evaluated on SD-OCT data obtained from external test sets and the results are shown in Table 10. The model was able to achieve an AUC value of 0.8571 and F1 score of 0.8705 on Dataset A (Stanford), AUC value of 0.7695 and F1 score of 0.7449 on Dataset B (Hong Kong), AUC value of 0.8706 and F1 score of 0.8860 on Dataset C (India), and AUC of 0.7965 and F1 score of 0.7736 on Dataset D (Nepal). Note that the deep neural network was not fine-tuned on the external data, hence the difference in performance. Fine-tuning the model on the external data sources will result in increased accuracy on the external test set.

False predictions were analysed on the external Dataset A (from Stanford), as can be seen in Table 12. Among the 13 false positive cases identified, large cup-disc ratio with larger disc areas were identified as the cause in 7 cases (53.8%) and in 6 cases (46.2%), older age (above 80) was an identifiable correlation. Among the 41 cases identified as false negative by the model, 21 (51.2%) cases had total average RNFL falling in the age matched normative range and had normal OCT RNFL thickness sector maps but had defects on GCIPL deviation and/or sector maps with corresponding visual field defects, and in 16 cases (39.0%), small disc area was the identifiable cause. Myopia was not associated with either false positive or false negative predictions.

We also analyzed the performance of the model separately for each myopia severity level. We defined severity of myopia by slightly modifying the Blue Mountain Eye Study (BMES) [16]. We modified the BMES category of moderate to severe myopia ($> 3D$) by further subdividing it into mild myopia (up to -3) moderate myopia (3 up to $-6D$) and severe myopia ($> 6D$), using cutoffs established in the Beijing Eye Study [17].

As can be seen in Table 11, the model was able to achieve a maximum F1 score of 0.9673 on severe myopia cases, and maximum accuracy of 0.9370 on severe myopia cases. Model was also able to achieve a F1 score of 0.9491 and accuracy of 0.9035 on moderate myopia cases. Performance on the mild myopia cases were lower than the severe and moderate myopia cases. The model achieved an F1 score of 0.8528, and accuracy of 0.7437 on mild myopia cases.

Saliency visualizations show that in most of the cases in which the model makes a *True Glaucoma* prediction, the Lamina Cribrosa is highlighted (see Figure 5a and Figure 5b). Out of the 182 cases predicted as *True Glaucoma* by the model on the Dataset A, all the cases had Lamina Cribrosa highlighted on the saliency visualizations, with or without retina highlighting. However, when the prediction is *True Normal*, superficial retina is highlighted in a high number of cases (see Figure 5c and Figure 5d). Out of the 134 cases predicted as *True Normal*, 70 cases (52.24%) had highlighting of the superficial retina and 38 out of 134 (28.36%) cases had superficial retina highlighting along with LC highlighting. Of the latter 38 cases, all had cup-to-disc ratio less than 0.4, which can indicate that the model utilizes low cup-to-disc ratio as a feature to predict normals.

Saliency maps of 158 *True Glaucoma* cases predicted as *True Glaucoma* by the model from the external Dataset A test set were evaluated. Out of the 158 aforementioned eyes, 76 eyes were moderate or severe glaucoma (Mean Deviation worse than -6). Among 49 out of 76 moderate to severe glaucoma cases (64.5%) had diffuse highlighting of the lamina cribrosa with minimal or no highlighting of the rest of the retina.

Out of the 81 cases identified as mild glaucoma, 49 cases (60.5%) had highlighting of the retina with focal highlighting of the LC.

In Figure 6a and Figure 6b, where the prediction of the model was *True Glaucoma* while the ground truth label was *True Normal*, the LC region is highlighted. In Figure 6c and Figure 6d, where the prediction of the model was *True Normal* while the ground truth label was *True Glaucoma*, the retinal layer is highlighted.

Table 10: Results of the proposed model on the internal and external test sets. Bootstrapping with 1000 trials was used for computing the standard deviation of the metrics.

Dataset	F1 Score	AUC	Accuracy
Primary Test	0.8834 (±0.0200)	0.8883 (±0.0188)	0.8410 (±0.0253)
Dataset A	0.8705 (±0.0104)	0.8571 (±0.0115)	0.8367 (±0.0121)
Dataset B	0.7449 (±0.0121)	0.7695 (±0.0096)	0.7447 (±0.0108)
Dataset C	0.8860 (±0.0115)	0.8706 (±0.0132)	0.8542 (±0.0137)
Dataset D	0.7736 (±0.0239)	0.7965 (±0.0188)	0.7740 (±0.0214)

Table 11: Results of the proposed model on the Dataset A (United States) external test set for each myopia severity level. Bootstrapping with 1000 trials was used for computing the standard deviation of the metrics.

Myopia Severity	Number of eyes	F1 Score	Accuracy
Mild	71	0.8528 (±0.0159)	0.7437 (±0.0240)
Moderate	29	0.9491 (±0.0161)	0.9035 (±0.0289)
Severe	38	0.9673 (±0.0135)	0.9370 (±0.0253)

Table 12: Observed causes of false predictions of *True Glaucoma* versus *True Normal* on the Dataset A (United States) external test set.

False Predictions	Number of eyes
False Positives	13
Large CD (> 0.5) with large disc area	7 (53.8%)
Age > 80	6 (46.2%)
False Negatives	41
Normal RNFL thickness maps with GCIPL deviation and/or sector map defects	21 (51.2%)
Small disc area	16 (39.0%)
Cause unidentifiable	4 (9.8%)

6 Discussion

In this study we developed and validated a 3D deep learning system using real world raw OCT optic nerve head volumes to detect glaucomatous optic neuropathy from normals. The labeled ground truth of glaucoma was assessed by reviewing fundus photos, OCT RNFL and macula results, visual field results, and IOP and treatment data over several visits to make sure there was no question glaucoma was present. Since the definition of glaucoma is very important when training an algorithm, we realized there is a limitation in diagnosing glaucoma with just the OCT red/yellow/green printout, and thus the qualitative RNFL and GCIPL thickness and deviation maps were reviewed (Table 1). While many of the recent studies (*e.g.* [18]) define structural changes in glaucoma based on OCT RNFL thickness and/or deviation maps alone, we think the additional multimodal test results allow for training an algorithm on a wider variation in the population instead of narrowing the inclusion criteria.

In our study, the machine learning system performed with an AUC of 0.8883 to differentiate between healthy and definite glaucomatous eyes of all ranges from early perimetric to late perimetric glaucoma. Our

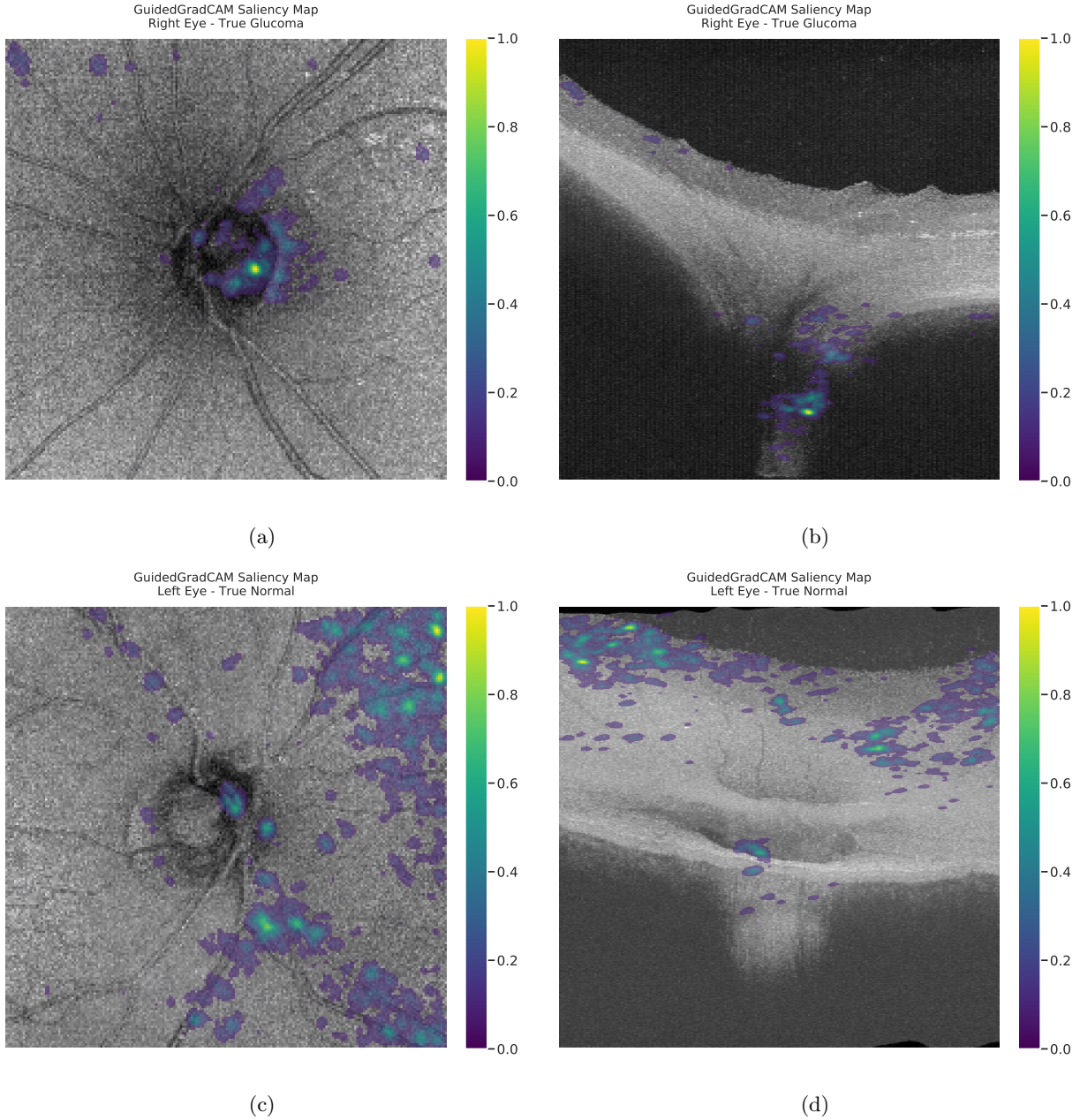


Figure 5: Saliency visualizations for two cases from the Stanford Test set. **(a)** Top, and **(b)** Side side view of saliency visualizations of a correctly classified glaucomatous eye. **(c)** Top, and **(d)** Side view of saliency visualizations of a correctly classified normal eye. As can be seen, in most of the cases, a highlight in the lamina cribrosa region is mostly correlated with *True Glaucoma* prediction, while for cases with *True Normal* prediction, the retinal layer is mostly highlighted. Saliency visualization have been obtained with respect to the predicted class. Regions with higher value are more salient for the model in making the final prediction.

performance with external testing generalized across multinational datasets where there are differing patient populations with varying disease severities. The performance was also very good on the external test set from India (Dataset C), which had an AUC value of 0.8706. We hypothesize that this is because there was a significantly higher percentage of eyes with severe disease in this dataset compared to other external datasets (Table 8). These cases would likely be easier to differentiate from normal cases. The performance was reduced on the dataset from Hong Kong with an AUC of 0.7695 (Dataset B). This can be explained due to the differences in their labeling criteria which defined structural changes in glaucoma based on the RNFL

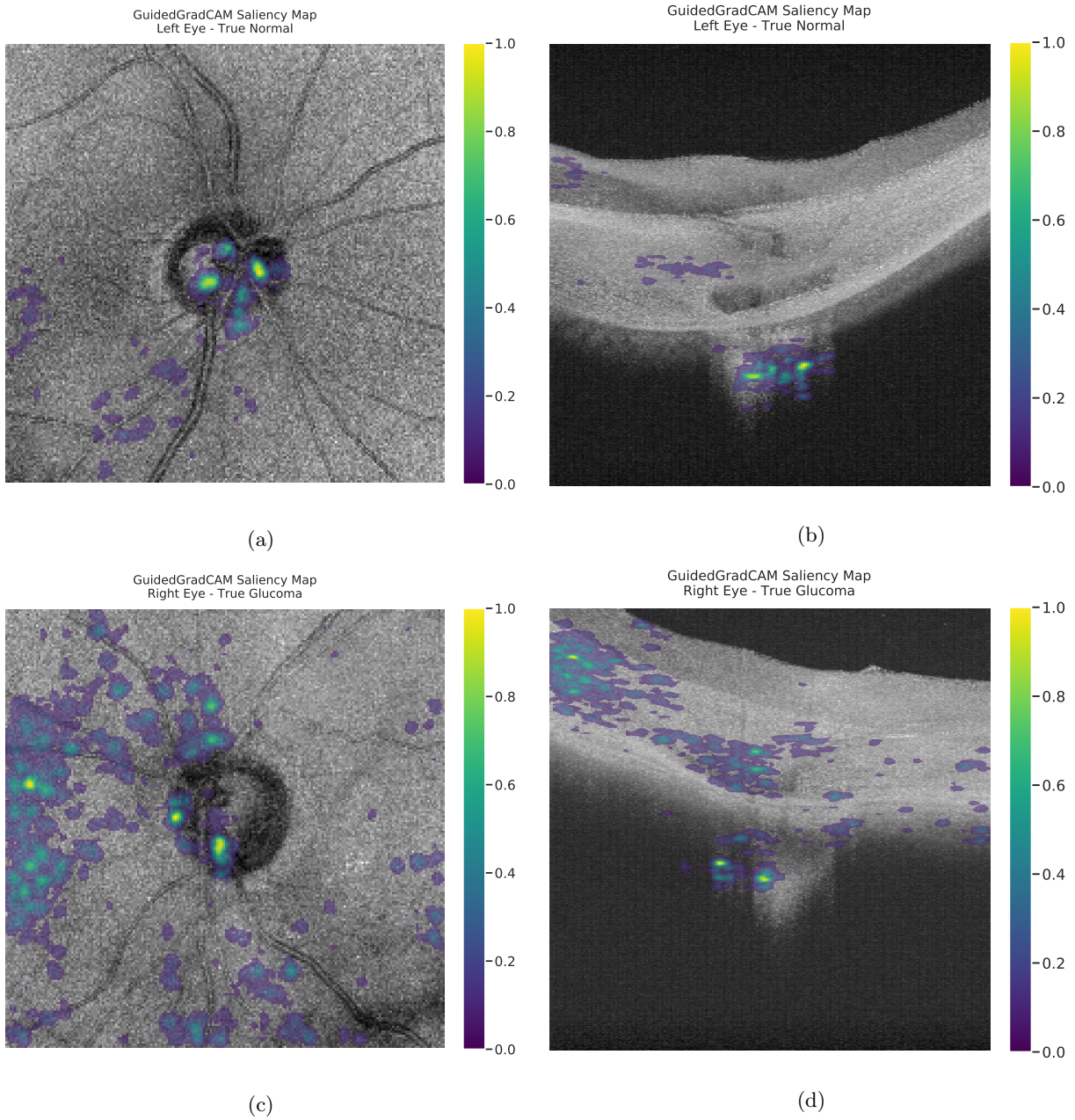


Figure 6: Saliency visualizations for two cases with wrong predictions. **(a)** Top, and **(b)** Side side view of saliency visualizations of a false positive case from the Hong Kong dataset. **(c)** Top, and **(d)** Side view of saliency visualizations of a false negative case from the India dataset. Saliency visualization have been obtained with respect to the predicted class. Regions with higher value are more salient for the model in making the final prediction.

thickness and/or deviation maps alone. Also there was significant differences in the refractive error between the *True Glaucoma* and *True Normal* cases in the training/primary validation/test dataset and the Hong Kong external test set. Mean refractive error in *True Glaucoma* cases in Dataset B (Hong Kong) was -0.85 (± 2.57) versus -3.57 (± 3.37) in the training, primary validation, and primary test sets. Mean refractive error in the *True Normal* cases was -0.51 (± 2.15) in Dataset B (Hong Kong) versus -2.2 (± 2.34) in the training, primary validation and test sets. Another reason for the difference in performance on the test set from Hong Kong could be the inclusion of solely gradable images with signal strength ≥ 5. We included cases with signal strength ≥ 3 and excluded images with artifacts which obscured imaging of ONH and the area

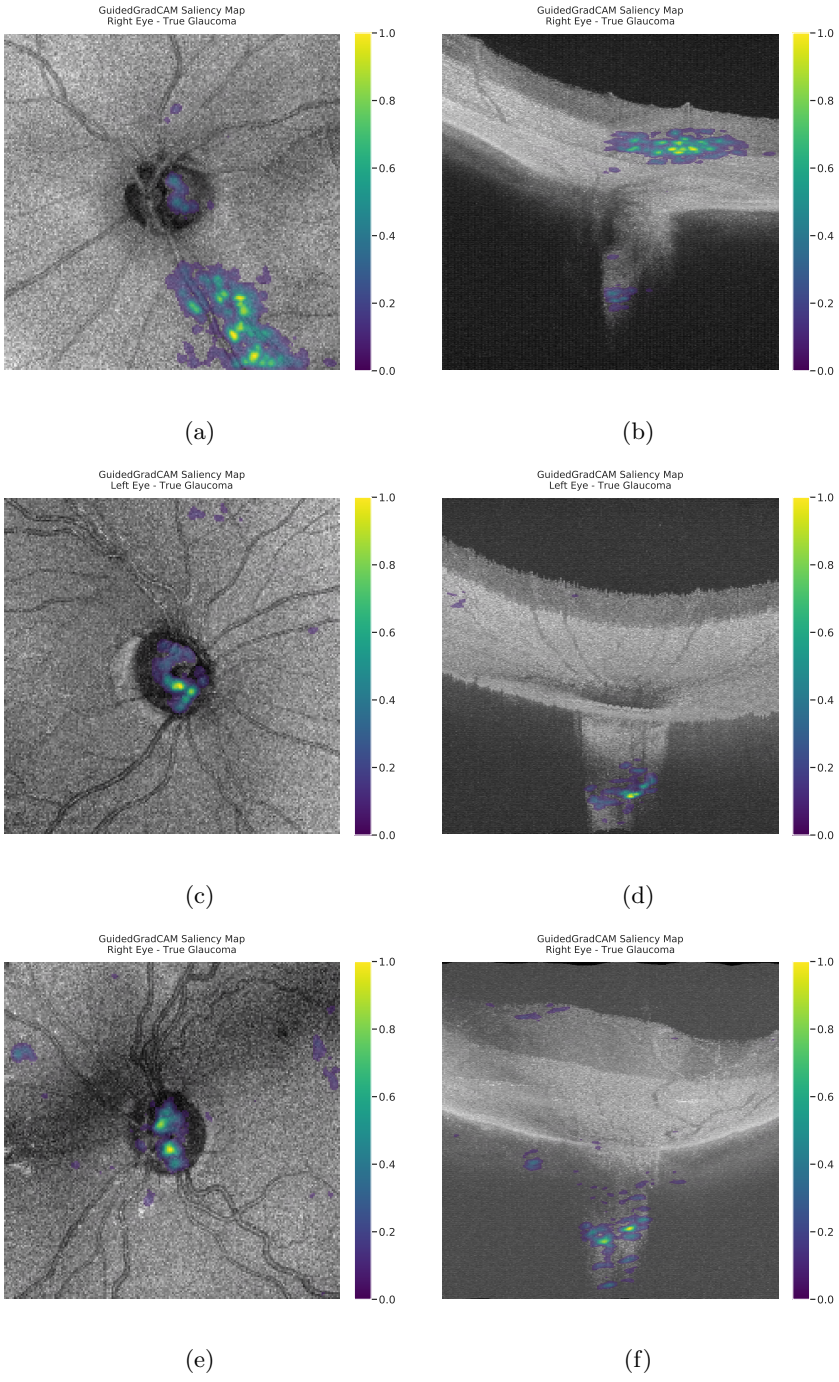


Figure 7: Saliency visualizations for cases with different glaucoma severity from the Dataset A external set. **(a)**, **(b)** Saliency visualization of a correctly classified case with mild glaucoma. **(c)**, **(d)** Saliency visualization of a correctly classified case with moderate glaucoma. **(e)**, **(f)** Saliency visualization of a correctly classified case with severe glaucoma.

inside and including the RNFL measurement circle at 3.4 mm from the center of the ONH. This is because many at times, clinicians are deprived of high quality OCT images for diagnosis and evaluation of glaucoma, due to medial opacity, tear film issues, small pupils, or other limitations. Our aim was to train the algorithm to be able to identify representations to detect glaucoma even on low quality images, hence replicating real world presentations. Even though it is recommended to obtain scans of signal strength higher or equal to 6 to facilitate the longitudinal quantitative progression calculated parameters, qualitative patterns in thickness

and deviation maps can still be seen at lower signal strength, and it is suggested that signal strength of > 3 is acceptable to obtain reproducible scanning images among patients with ocular media opacities [19].

With our fourth external dataset from Nepal (Dataset D), the model performed with an AUC of 0.7965. Possible explanations for the difference in performance could be due to the differences in the dataset. The mean age of the subjects were significantly lower in this dataset. The percentage of eyes with severe myopia in the *True Glaucoma* and *True Normal* subsets were lower compared to the training/primary validation/primary test sets from Stanford. The mean refractive error was significantly lower in this dataset. Another possible reason for the differences in performance across the external datasets could be possible inter- and intra-grader variability in labeling of cases based on the criteria.

A novel output of our model is its ability to detect glaucoma across different ranges of myopia (Table 11). The model was able to achieve an accuracy of 0.9370 on severe myopia cases, accuracy of 0.9491 on moderate myopia, and accuracy of 0.7437 on mild myopia cases. It is known that diagnosing glaucoma in the setting of myopia is a common challenge due to alteration of the appearance of the optic nerve and OCT. Myopic refractive error impacts RNFL and macular thickness measurements due to stretching and thinning of these layers due increased axial length and optical projection artifact of the scanning area [20]. This often results in many false positive diagnoses, also known as "Red Disease". Using the entire cube and highlighting the lamina cribrosa may help researchers study this LC region more closely in myopes when trying to differentiate glaucoma from normal. The difference in the performance in the myopia subsets compared to the total dataset could be due to the fewer number of cases in each subgroup (Table 11).

What was most interesting from our model were the saliency maps of the regions in the scan where the model attends to make a prediction. Normally, we expected the RNFL to be a majority of the differentiation of true glaucoma from normal, but in many cases, the lamina cribrosa was just as important, or sometimes more important since the RNFL can be thinned for other reasons such as myopia. Given that clinicians do not routinely review every single slice of the cube, and there is no OCT printout highlighting the lamina, we were excited to discover that, by training a model on every single slice, saliency visualization highlighted the lamina cribrosa region along with exiting nerve fibers posterior to LC, and in most cases are correlated with *True Glaucoma* prediction. For cases with *True Normal* prediction, the areas on superficial retina were mostly highlighted in saliency visualizations. This corresponds with clinical practice, whereby when an OCT RNFL is all normal (all RNFL quadrants colored green or white), then likely it has a very high negative predictive value for glaucoma. In a smaller subset of cases predicted as *True Normal* when lamina cribrosa was highlighted along with retina, smaller cup-to-disc ratio (≤ 0.4) was noted as an association suggesting that the model might have identified smaller cup-to-disc ratio as a feature to identify normals. Peripheral LC or regions posterior to large blood vessels typically remain difficult for OCT image interpretation without enhanced depth imaging. It was also observed that in moderate to severe glaucoma, there was diffuse highlighting of the lamina cribrosa with minimal or absent highlighting of the retina, whereas in mild glaucoma, retina was being highlighted with focal or diffuse highlighting of the lamina cribrosa (see Figure 7). This correlates with the fact that in advanced disease, RNFL thickness levels off, falling below 50 μm and almost never below 40 μm for the Cirrus machine, due to the assumed presence of residual glial or non-neural tissue including blood vessels and hence making RNFL measurement less clinically useful at this stage [21]. However, lamina cribrosa may not be limited by this floor effect and if 3D information was used, a new method to monitor structure progression in end stage glaucoma could be created. This needs more analysis with a larger distribution of glaucoma severity based datasets.

Our assessment of false predictions by the 3D deep neural model showed no correlation with myopia, despite the fact that myopia is one of the most common reason for misdiagnosis of glaucoma in clinical presentations [20]. This suggests that by training the model on all scans including high myopes and low signal strength ones as long as there were no data loss artifacts, could provide enough training examples within the volumes of slices to avoid myopia affecting the result. An interesting observation was false negative prediction of cases diagnosed as *True Glaucoma* based on structural defects on GCIPL maps alone. This emphasizes the need for evaluation of optic nerve head and macula parameters in detecting glaucoma [22].

Recently Maetschke *et al.* [23] employed 3D convolutional neural networks to classify eyes as healthy or glaucomatous directly from raw, unsegmented OCT volumes (1110 scans) of the optic nerve head obtained using Cirrus SD-OCT scanner (Carl Zeiss Meditec Inc., Dublin, CA, USA) and achieved a substantially high AUC of 0.94 against logistic regression, which was found to be the best performing classical machine learning technique with an AUC of 0.89. In their study, glaucomatous eyes were defined as those with glaucomatous visual field defects alone and was not based on any structural parameters. This work used a convolutional

neural network for the task of glaucoma classification, however, the architecture used for the neural network was different from the architecture of the proposed model. Another difference from our study was that they included scans with signal strength ≥ 7. Despite the differences in definition and inclusion criteria, it is interesting to note that our saliency maps had similar findings. Similar to our study, for healthy eyes, the network in [23] tends to focus on a section across all layers and ignores the optic cup/rim and the lamina cribrosa. In contrast, for glaucomatous eyes, the optic disc cupping, neuroretinal rims, as well as the lamina cribrosa and its surrounding regions were highlighted. The strength of our study compared to [23] is that we included more information about our training population and had multiple external datatests for validation.

In the recent study by Ran *et al.* [18], the 3D deep learning system had an AUC of 0.969. The study showed good performance with external test set from United States with an AUC of 0.893. Similar to our study, the heatmaps generated in their study showed neuro-retinal rim and areas covering the lamina cribrosa to be highlighted in detection of glaucomatous optic neuropathy. Apart from this, the retinal nerve layer and choroid were also potentially found be related to detection of glaucomatous optic neuropathy in their study. The difference in their study from ours was in the definitions used for glaucoma and inclusion of images with signal strength ≥ 5. They defined glaucomatous structural defect based on OCT RNFL thickness and deviation maps.

While it is unclear about the distribution or inclusion of different degrees of myopia in their study, our cohort had 11 percentage of total eyes with severe myopia (≥ -6) in our *True Glaucoma* subset and 4.09 percentage of total eyes with severe myopia in the *True Normal* subset in training, primary validation, and primary test sets. Another difference was the distribution of ethnicity in their training set which consisted exclusively of Chinese Asian eyes, while our training, primary validation and, primary test sets included subjects of Caucasian, Asian (which included Chinese Asians, Non-Chinese Asians, and Indians), African American, and Hispanic origin.

The major differences between the recent studies [23, 18] and ours was the diversity in the ethnicity of the datasets used for training of the model, inclusion of high refractive errors in both glaucoma and normal cases for training, and inclusion of eyes with lower signal strength, hence representing the real world clinical presentations. Our work used external datasets from United States, India, Hong Kong, and Nepal, while similar works (*e.g.* [18]) did not have similar variety in the external tests sets used.

Our study has several strengths. Multiple international datasets provide diversity in our database for evaluation purposes, which is rare to have for glaucoma datasets. We had images from patients of different ethnicities, including Caucasian, Asian (including both Chinese Asian and Non-Chinese Asian), African American, Hispanic, and of Indian origin. The performance of our model was promising across multiple geographies and ethnicities to distinguish glaucoma from normal.

Another significant strength of our method was that our main training dataset was not cleaned for this experiment to more closely follow the challenges that are faced in real world clinical settings. While strict exclusion criteria such as axial length, small and large disc sizes, and high myopia are common, our cohort included all ranges of myopia, disc sizes, and axial lengths, reflecting real world presentations. One other major highlight of our study was the criteria used to classify cases as *True Glaucoma* versus *True Normal* in the training and validation dataset, which included both multimodal longitudinal structural and functional evaluations. This closely replicates real world clinical settings where multimodal longitudinal evaluation is used to arrive at the diagnosis.

Our study has few drawbacks. We did not include "Suspect" cases in our datasets. This was mainly because of the difficulty in obtaining consensus for glaucoma suspect definition among experts. We are now working on a separate dataset and are trying to achieve consensus among multiple glaucoma experts to classify high- and low-risk suspect cases or referral cases. Additionally, we have not included "Preperimetric" glaucoma in the training due to the unavailability of adequate number of cases in the subset.

Even though we have not excluded any cases based on disc sizes or presence of myopic tilted discs in our datasets, and have included cases with low signal strength, we have not looked into the performance of our model across subsets.

Going forward, we plan to develop a 3D deep learning algorithm using a wider range of data including high- and low-risk suspect cases that would help in identifying cases which require referral for management by glaucoma specialists. Secondly, we also plan to evaluate the performance across severity of glaucoma cases and look closely at the patterns in each severity subset by including larger number of cases in each subset. Further, we plan to include raw OCT macula cube scans along with optic nerve head scans for better algorithm correspondence. Finally, we intend to study cropped images including LC, regions adjacent, and posterior to LC to further characterize the saliency mapping highlights of true glaucoma from normal.

7 Conclusion

Our 3D deep learning model was trained and tested using the largest OCT glaucoma dataset so far from multinational data sources, and has been able to detect glaucoma from raw SD-OCT volumes across severity of myopia and severity of glaucoma. By using a multimodal definition of glaucoma, we could include more scans from the real world. The saliency visualizations highlighted the lamina cribrosa as an important component in the 3D optic nerve head cube in differentiating glaucoma.

References

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

[2] A. Coleman and L. Brigatti, "The glaucomas.," *Minerva medica*, vol. 92, no. 5, pp. 365–379, 2001.

[3] I. A. Sigal, B. Wang, N. G. Strouthidis, T. Akagi, and M. J. Girard, "Recent advances in oct imaging of the lamina cribrosa," *British Journal of Ophthalmology*, vol. 98, no. Suppl 2, pp. ii34–ii39, 2014.

[4] C. Y. Kim, J. W. Jung, S. Y. Lee, and N. R. Kim, "Agreement of retinal nerve fiber layer color codes between stratus and cirrus oct according to glaucoma severity," *Investigative ophthalmology & visual science*, vol. 53, no. 6, pp. 3193–3200, 2012.

[5] Zeiss, "Cirrus HD-OCT with RNFL, Macular, Optic Nerve Head, and Ganglion Cell Normative Databases.." `https://www.zeiss.com/meditec/int/products/ophthalmology-optometry/glaucoma/diagnostics/oct/oct-optical-coherence-tomography/cirrus-hd-oct.html`, 2019.

[6] P. J. Foster, R. Buhrmann, H. A. Quigley, and G. J. Johnson, "The definition and classification of glaucoma in prevalence surveys," *British journal of ophthalmology*, vol. 86, no. 2, pp. 238–242, 2002.

[7] D. R. Anderson and V. M. Patella, "Automated static perimetry," 1992.

[8] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. ODonoghue, D. Visentin, *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature medicine*, vol. 24, no. 9, p. 1342, 2018.

[9] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "Densenet: Implementing efficient convnet descriptor pyramids," *arXiv preprint arXiv:1404.1869*, 2014.

[10] Y. Wu and K. He, "Group normalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

[11] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, pp. 448–456, 2015.

[12] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," *arXiv preprint arXiv:1711.05101*, 2017.

[13] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

[14] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[15] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[16] P. Mitchell, F. Hourihan, J. Sandbach, and J. J. Wang, "The relationship between glaucoma and myopia: the blue mountains eye study," *Ophthalmology*, vol. 106, no. 10, pp. 2010–2015, 1999.

[17] L. Xu, Y. Wang, S. Wang, Y. Wang, and J. B. Jonas, "High myopia and glaucoma susceptibility: the beijing eye study," *Ophthalmology*, vol. 114, no. 2, pp. 216–220, 2007.

[18] A. R. Ran, C. Y. Cheung, X. Wang, H. Chen, L.-y. Luo, P. P. Chan, M. O. Wong, R. T. Chang, S. S. Mannil, A. L. Young, *et al.*, "Detection of glaucomatous optic neuropathy with spectral-domain optical coherence tomography: a retrospective training and validation deep-learning analysis," *The Lancet Digital Health*, vol. 1, no. 4, pp. e172–e182, 2019.

[19] M. M. Ha, J. M. Kim, H. J. Kim, K. H. Park, M. Kim, and C. Y. Choi, "Low limit for effective signal strength in the stratus oct in imperative low signal strength cases," *Korean Journal of Ophthalmology*, vol. 26, no. 3, pp. 182–188, 2012.

[20] J. C. Mwanza, F. E. Sayyad, A. A. Aref, and D. L. Budenz, "Rates of abnormal retinal nerve fiber layer and ganglion cell layer oct scans in healthy myopic eyes: Cirrus versus rtvue," *Ophthalmic Surgery, Lasers and Imaging Retina*, vol. 43, no. 6, pp. S67–S74, 2012.

[21] D. C. Hood and R. H. Kardon, "A framework for comparing structural and functional measures of glaucomatous damage," *Progress in retinal and eye research*, vol. 26, no. 6, pp. 688–710, 2007.

[22] J. W. Jeoung and Y. J. Choi, "Macular ganglion cell imaging study: Glaucoma diagnostic accuracy of spectral-domain optical coherence tomography," *Invest. Ophthalmol. Vis. Sci.*, vol. 54, no. 7, pp. 4422–4429, 2013.

[23] S. Maetschke, B. Antony, H. Ishikawa, G. Wollstein, J. Schuman, and R. Garnavi, "A feature agnostic approach for glaucoma detection in oct volumes," *PloS one*, vol. 14, no. 7, p. e0219126, 2019.