Detection of glaucomatous optic neuropathy with spectral-domain optical coherence tomography: a retrospective training and validation deep-learning analysis

An Ran Ran, Carol Y Cheung, Xi Wang, Hao Chen, Lu-yang Luo, Poemen P Chan, Mandy O M Wong, Robert T Chang, Suria S Mannil, Alvin L Young, Hon-wah Yung, Chi Pui Pang, Pheng-Ann Heng, Clement C Tham (The Lancet Digital Health, 2019)

Fall, 2020

Section 1

Introduction

Why detection of GON?

Glaucoma is a major cause of irreversible visual morbidity worldwide

- Projected to affect 111.8 million people by 2040

Early detection of glaucomatous optic neuropathy (GON)

- Difficult as early stages usually experience no symptoms
- Subjective (and expensive) examination of optic neuropathic features
- Essential for timely treatment and minimization of irreversible vision loss

Why SDOCT?

Spectral-domain optical coherence tomography (SDOCT)

- A non-contact and non-invasive imaging technology for cross-sectional and three-dimensional (3D) viewing of the retina and optic nerve head
- Able to diagnose mild to moderate glaucoma (not just late-stage)
- Monitor disease progression

Advantages of SDOCT

- Good sensitivity and specificity for glaucoma detection
- Useful for screening GON in high-risk communities

Limitations of SDOCT

- Professionals are required to interpret SDOCT results
- Diagnosis based on built-in normative databases could be unreliable
  - small optic discs + long optic axes = high chances of abnormal diagnostic classification, which results in false-positive errors

Why deep learning?

Deep learning with convolutional neural networks (CNNs)

- Proved to be effective in automated classification of diabetic retinopathy, age-related macular degeneration and other retinal diseases
- Some models are potentially more accurate than certified specialists

Existing CNN-based glaucoma detection

- Seems only 2D-based in the literature before Apr 2019
- Usually quantify glaucomatous damages via fundus photographs
  - thickness of the retinal nerve fibre layer
  - minimum rim width relative to Bruch's membrane opening
- No model available for other features such as
  - optic nerve head structure in the 3D cube
  - inner retinal neuronal layers
  - morphological changes in lamina cribrosa

Ideas on SDOCT

Time-domain optical coherence tomography (TDOCT)

- See this link for an introduction and comparison
- SDOCT can be proved to be intrinsically more sensitive than TDOCT

Idea 1

Can we apply techniques from spectral analysis to SDOCT data?

Unreliability related to normative database

Idea 2

Can we use external data (i.e., outside the medical image) to supplement the analysis? We will discuss this later.

Ideas on dimension

Contributions of this paper

- Develop a 3D deep-learning system based on SDOCT volumetric data
- Investigate the system's ability to detect GON

Idea 3

Can we improve the system with panel data (3D + time = 4D)? This is possible as the patients may revisit the clinic at another time.

Idea 4

Can we improve the system with video data? Note that the time dimension here is different from the above.

Idea on detection

Glaucomatous damage may not be glaucoma

- See this link for examples
- Detecting GON to predict glaucoma may not be the best choice
- Deep learning works well in prediction (but not yet causation)

Idea 5

Do we need other relationships (e.g., causation) to prevent glaucoma?

Section 2

Methods

Main datasets

Training, testing, and primary validation dataset

- Source: CUHK Eye Centre and Hong Kong Eye Hospital
- Date range: March 1, 2015 to Dec 31, 2017
- Participants: 18 years or older, with reliable visual field tests and gradable SDOCT optic nerve head scans
  - also included healthy volunteers who joined for opportunistic screening
- Exclusion: with other ocular or systemic diseases that could cause visual defects, or missing data for visual field tests or SDOCT
- Device: Cirrus HD-OCT. In each gradable SDOCT scan, extract
  - raw 3D volumetric images (main system)
  - 2D line-scanning ophthalmoscope en face image (benchmark purpose)

Remark 1

Note that missing data were discarded but not imputed.

External datasets

External validation datasets

- Source: independently from
  - Prince of Wales Hospital
  - Tuen Mun Eye Centre
  - Byers Eye Institute, Stanford University
- Criteria: same inclusion, exclusion, visual field, and SDOCT device
  - only date ranges differed

Ground truth labeling

Assessment of SDOCT images by human graders

- Gradable
  - presence of GON
    - ★ defined by Collaborative Normal-Tension Glaucoma Study Group
    - ★ two glaucoma specialists for HK and Stanford datasets respectively
    - ★ discrepancies were reviewed by a senior glaucoma specialist (HK), resolved by consensus or excluded if no consensus (Stanford)
  - absence of GON (i.e., normal/healthy)
- Not gradable
  - always excluded

Deep-learning system

Brief specification

- Backbone: ResNet-34 without pretraining
- Input: randomly divided (3:1:1) into training, testing and validation
  - ratios of presence to absence of GON were similar
  - images from the same patient were confined to the same set
  - training-validation curve was assessed to avoid over-fitting

Idea 6

Can we use other backbone (e.g., VGG) or transfer learning to improve the performance? There are many other ideas in the computer vision literature.

3D deep-learning system

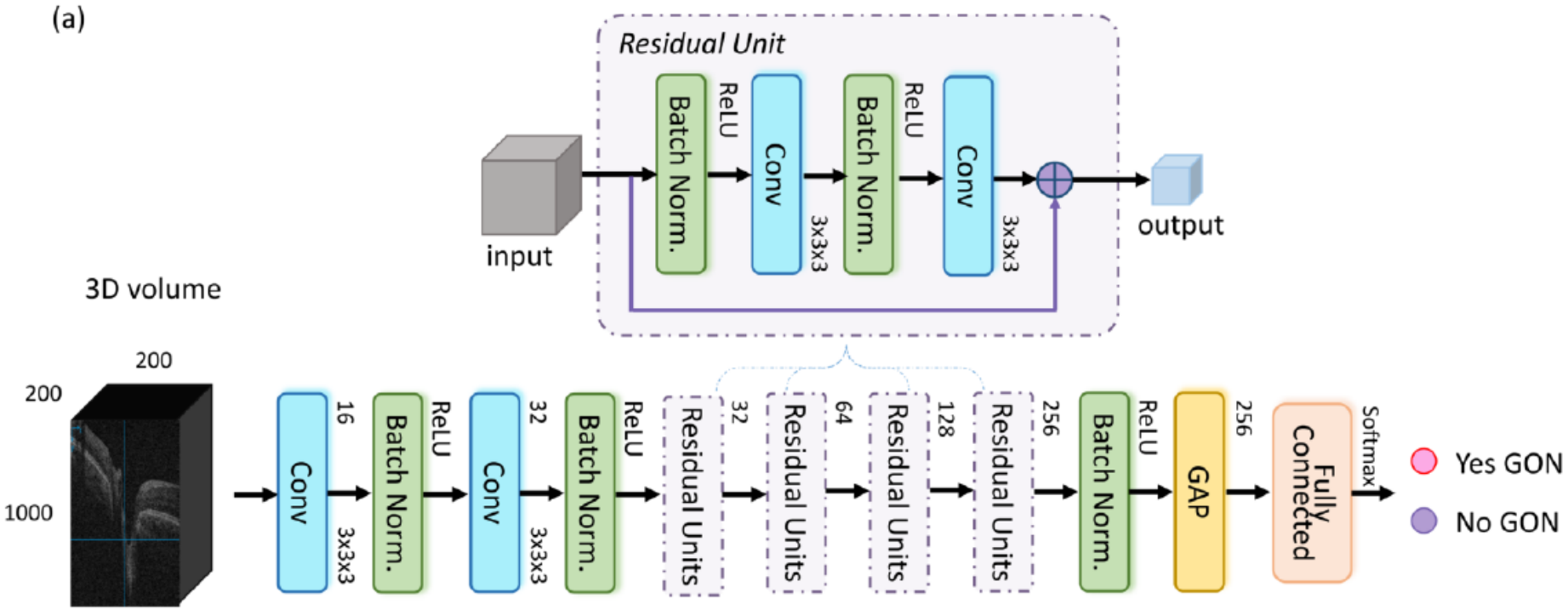

Figure 1: *A diagram showing the architecture of the 3D deep learning system (DLS). The 3D DLS was built based on the classical ResNet34 network with 3D convolutional layers and global average pooling layer. The input of the DLS was an OCT volumetric scan of size 200x1000x200 pixels after image pre-processing and the output was yes/no GON. (Source: suuplementary materials of this paper)*

2D deep-learning system

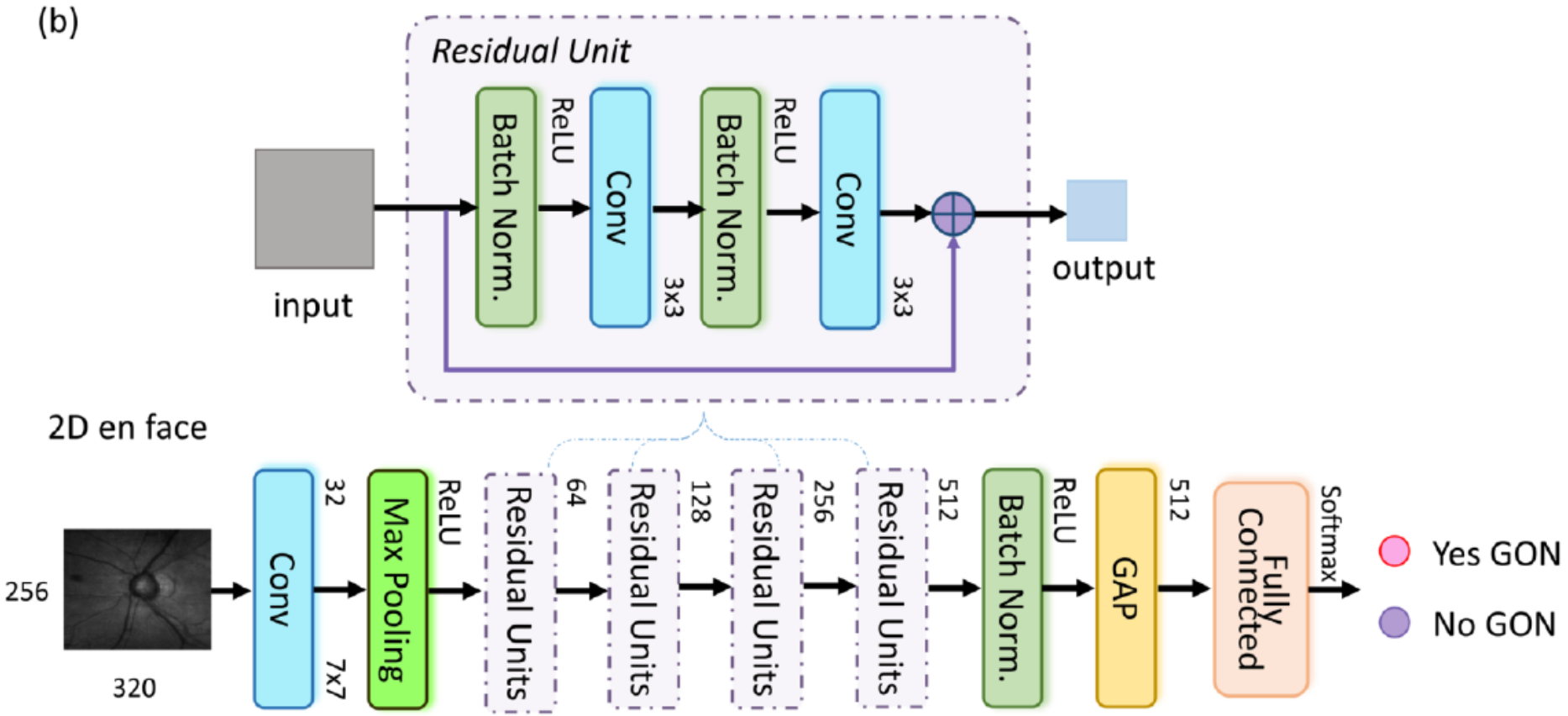

Figure 2: *A diagram showing the 2D deep learning system (DLS). We simply adopted ResNet34 with a minor modification as 2D DLS. The average pooling layer was replaced with global average pooling. The input of the DLS was an OCT en face image of size 320x256 pixels after image pre-processing and the output was yes/no GON. (Source: suuplementary materials of this paper)*

More on the deep-learning system

- Standardization and normalization (zero mean and unit variance)
- Scaling to range from 0 to 1
- Random cropping, jittering/rotating and flipping to alleviate over-fitting
- Weighted binary cross-entropy loss function
  - imbalance training dataset (e.g., 2 'yes GON' to every 1 'no GON')
- Implemented in Keras
  - learning rate: 0.0001 and 0.00001 for 3D and 2D CNN
  - optimization algorithm: Adam
- Other modifications
  - e.g., number of filters in ResNet-34 was halved due to limited memory

Remark 2

The batch size was not reported. From the limited memory consideration, perhaps they did not adjust these hyperparameters.

Statistical analysis

Hypothesis: performance of the 3D deep-learning system is similar to that of experienced human assessors

- in terms of detection of GON from SDOCT volumetric data
- in both primary and external validations

Tests (all two-sided with $\alpha = 0.05$)

- Numerical demographic data: Wilcoxon rank sum test
- Categorical demographic data: χ^2 test
  - also used to analyze variances of data between the different datasets

Additional analyses

- Subgroups stratified by age, sex, eye, signal strength, severity of GON, size of disc area and ethnicity
- Number and proportion of eyes with pre-perimetric glaucoma that were predicted to have GON

Section 3

Results

Summary of study participants

I omit the table in the paper here

- it kind of reinforces which factors are more important in detecting GON
- so I consult my friend who is a medical student

TABLE 1
Risk factors for POAG[21]

Risk factor	Relative risk
Age (per 10 years after age 40)	x2
African ethnicity	x4
Family history (first degree relative)	x2-4
Low diastolic perfusion pressure	x3
IOP > 30mmHg (vs. <15mmHg)	x40
IOP 22-29	x13
IOP 19-21	x3
Myopia	x1.5 - 3

Figure 3: *POAG: primary open-angle glaucoma; IOP: intraocular pressure (Source: Optician)*

Performance of the systems

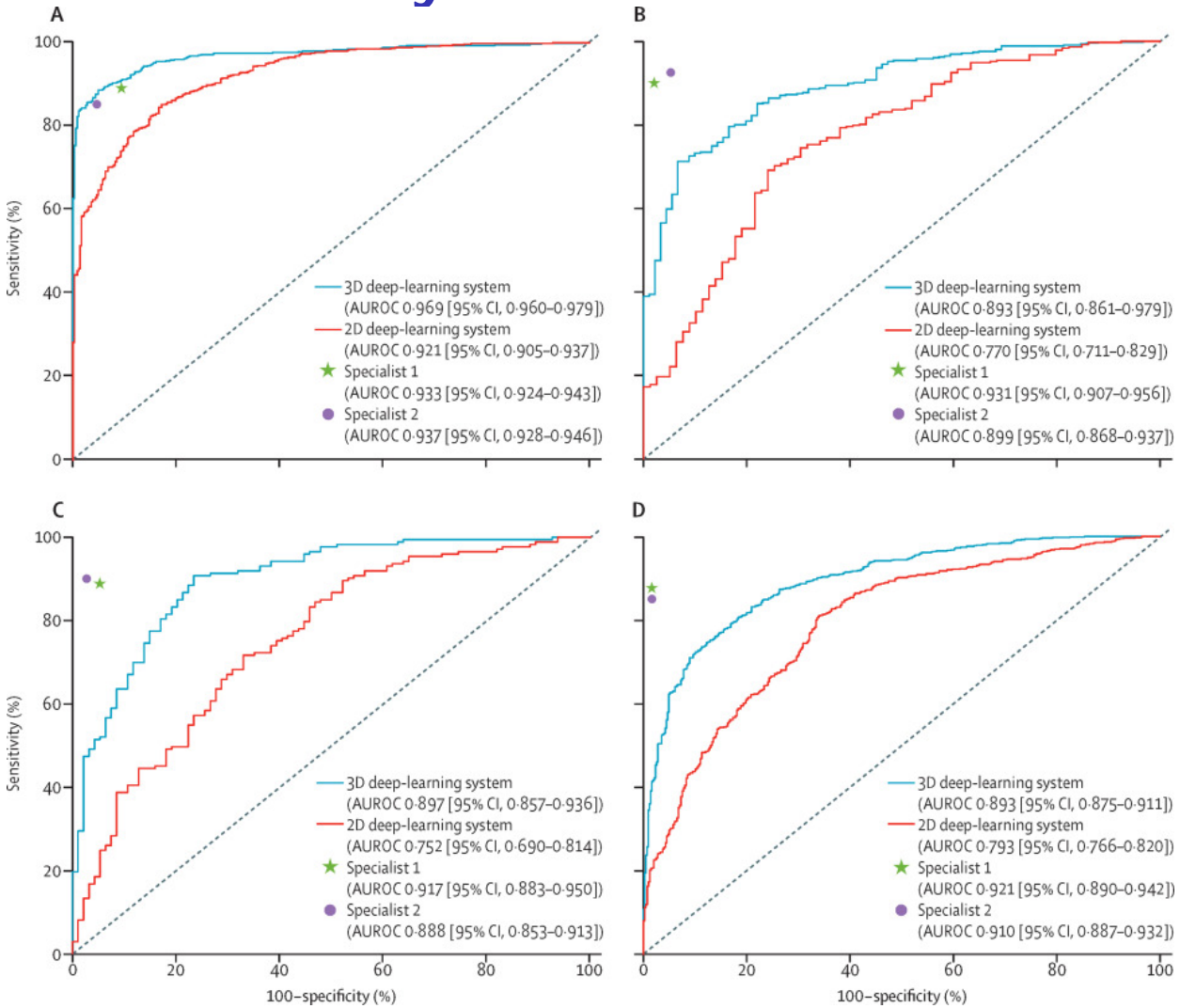

Figure 4: *Receiver operating characteristic curves for the 3D deep-learning system and 2D deep-learning system compared with two glaucoma specialists in the primary validation dataset (A), and in exernal validation datasets 1 (B), 2 (C), and 3 (D)*

Performance of the systems

	AUROC (95% CI)	p value	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy (95% CI)
Primary validation dataset					
3D deep-learning system	0·969 (0·960–0·979)	..	89%(83–93)	96% (92–99)	91% (89–93)
2D deep-learning system	0·921 (0·905–0·937)	<0·0001	85% (77–89)	85% (79–91)	84% (82–87)
Glaucoma specialist 1	0·933 (0·924–0·943)	<0·0001	89% (86–92)	94% (91–97)	91% (89–93)
Glaucoma specialist 2	0·937 (0·928–0·946)	<0·0001	87% (84–90)	96% (93–98)	90% (88–93)
External validation dataset 1					
3D deep-learning system	0·893 (0·861–0·979)	..	79% (76–83)	84% (76–90)	80% (77–83)
2D deep-learning system	0·770 (0·711–0·829)	<0·0001	72% (64–83)	75% (59–85)	73% (67–80)
Glaucoma specialist 1	0·931 (0·907–0·956)	0·022	89% (85–92)	97% (94–100)	90% (88–93)
Glaucoma specialist 2	0·899 (0·868–0.931)	0·42	85% (81–89)	95% (89–99)	87% (84–89)
External validation dataset 2					
3D deep-learning system	0·897 (0·857–0·936)	..	90% (79–95)	79% (69–88)	86% (80–90)
2D deep-learning system	0·752 (0·690–0·814)	<0·0001	78% (49–94)	64% (43–88)	73% (63–80)
Glaucoma specialist 1	0·917 (0·883–0·950)	0·33	89% (84–93)	95% (89–99)	91% (87–94)
Glaucoma specialist 2	0·888 (0·853–0·913)	0·80	81% (75–87)	97% (93–100)	87% (82–90)
External validation dataset 3					
3D deep-learning system	0·893 (0·875–0·911)	..	78% (71–88)	86% (76–93)	81% (77–84)
2D deep-learning system	0·793 (0·766–0·820)	<0·0001	82% (78–88)	66% (58–71)	76% (73–79)
Glaucoma specialist 1	0·921 (0·890–0·942)	0·48	86% (82–90)	98% (96–100)	91% (89–94)
Glaucoma specialist 2	0·910 (0·887–0·932)	0·88	84% (80–88)	98% (96–100)	90% (88–93)

Figure 5: *Comparison of the 3D deep-learning system, 2D deep-learning system, and assessments by two glaucoma specialists in the datasets.The Z test was used to calculate p values for comparison of AUROCs between groups.*

Output of the systems

Videos comparisons available in the online supplement

- Red–orange-coloured area has the most discriminatory power
- Heatmaps shows potential with other areas covering the lamina cribrosa and choroid in detecting GON
  - in contrast to the traditional retinal nerve fibre layer and neuroretinal rim
- False-negative results: mainly due to small disc area
- False-positive results: mainly due to large disc area

Subgroups analyses

	Primary validation dataset		External validation dataset 1		External validation dataset 2		External validation dataset 3	
	AUROC (95% CI)	p value	AUROC (95% CI)	p value	AUROC (95% CI)	p value	AUROC (95% CI)	p value
Age								
<60 years	0·932 (0·895–0·969)	0·0060	0·871 (0·801–0·942)	0·65	0·930 (0·884–0·975)	0·079	0·889 (0·854–0·924)	0·97
≥60 years	0·980 (0·973–0·988)	..	0·890 (0·850–0·929)	..	0·853 (0·781–0·925)	..	0·888 (0·866–0·911)	..
Eyes								
Right	0·970 (0·956–0·984)	0·97	0·914 (0·877–0·950)	0·15	0·916 (0·867–0·965)	0·47	0·896 (0·871–0·921)	0·95
Left	0·971 (0·958–0·983)	..	0·863 (0·806–0·921)	..	0·888 (0·828–0·947)	..	0·895 (0·869–0·920)	..
Sex								
Male	0·973 (0·962–0·984)	0·54	0·911 (0·867–0·955)	0·35	0·913 (0·862–0·963)	0·62	0·880 (0·850–0·910)	0·22
Female	0·967 (0·95–0·983)	..	0·881 (0·835–0·927)	..	0·894 (0·836–0·951)	..	0·903 (0·881–0·926)	..
Signal strength								
<8	0·859 (0·825–0·893)	0·042	0·915 (0·854–0·977)	0·42	0·848 (0·734–0·961)	0·36	0·882 (0·850–0·914)	0·33
≥8	0·912 (0·874–0·951)	..	0·885 (0·846–0·924)	..	0·905 (0·863–0·946)	..	0·901 (0·879–0·923)	..
Severity of glaucomatous optic neuropathy								
Mild	0·957 (0·940–0·974)	<0·0001	0·841 (0·791–0·891)	<0·0001	0·884 (0·840–0·929)	0·043	0·861 (0·836–0·885)	<0·0001
Moderate or severe	0·982 (0·975–0·989)	..	0·919 (0·890–0·947)	..	0·943 (0·908–0·978)	..	0·926 (0·908–0·943)	..
Disc area								
Small	0·915 (0·888–0·942)	<0·0001	0·917 (0·880–0·953)	0·064	0·905 (0·855–0·955)	0·56	0·947 (0·929–0·965)	<0·0001
Large	0·794 (0·737–0·852)	..	0·848 (0·786–0·910)	..	0·880 (0·715–0·946)	..	0·900 (0·876–0·924)	..
Race or ethnicity								
Asian	..	..	..	..	..	..	0·905 (0·880–0·930)	0·49
Non-Asian*	..	..	..	..	..	..	0·892 (0·867–0·918)	..

Figure 6: *Comparisons of AUROCs for the three-dimensional deep-learning system in the datasets. The Z test was used to calculate p values for comparison of AUROCs between groups. Non-Asian ethnicity covered African-American, white, and Hispanic people.*

Remark on hypothesis tests

Two sample z-tests were used to compare the AUROC

- Note that the ROCs may not be independent
  - e.g., the 3D and 2D ROCs
  - then problem with variance estimation
- See this link for a short discussion

Remark 3

Some statistical tests in the paper may be improvable.

Section 4

Discussion

Implications

Automated detection of GON in SDOCT volumes is possible

- Note that glaucoma should not be diagnosed solely based on SDOCT
  - can detect glaucomatous structural changes
  - can provide preliminary detection
- Performance was similar to experienced glaucoma specialists
- Heatmaps highlighted other potentially useful structures, e.g., choroid
- Some additonal risk factors seem to make no difference in detection
  - age, eye (ie, left vs right), sex, signal strength and ethnicity
  - performance was worse for mild than moderate or severe GON
  - but the situation is same for experienced ophthalmologists

Implications

Performance in external validations was slightly reduced. Possible reasons:

- inter-grader and intra-grader variability in assessments of GON
- difference in glaucoma-related features (e.g., severity) between datasets
- different variances in SDOCT raw images among the datasets

2D deep-learning system that was substantially outperformed by the 3D one

- difference in input compared with literature
  - this paper: line-scanning ophthalmoscope images
  - literature: paired colour fundus photographs
- AUROC better than most models in the literature
  - note that it is problematic to directly compare across studies
  - however fundus photography missed features like inner retinal neuronal layers and lamina cribrosa morphology

Remark on 2D vs 3D

The benchmark comparison is not surprising

- Same data source, only differs in dimension
- 3D images may contain more (hidden) information
- Similar models are used
- Result thus directly reflect data quality

Remark 4

The benchmark comparison is not surprising. However it did illustrate the usefulness of 3D images.

Strengths

1. External validation datasets were collected from eye clinics in different geographical locations
2. Visual field reports were available for labeling of GON
3. Heatmaps were generated to visualize the discriminative image regions among the SDOCT volumes

Idea 7

As in the papers we read previously, visualization is important for real applications. Practitioners may not believe the system works without it as deep learning models somehow work like a black box.

Limitations

1. Only gradable images were included for training and validation
   - the authors are working on a separate deep-learning algorithm for automated filtering of ungradable SDOCT volumes
2. Only cases of GON that had visual field defects were included
   - plan to include eyes with suspected glaucoma in the next version
3. Only one type of SDOCT device was used
4. Data was collected mostly Chinese participants
   - although diagnostic performance did not differ between Asian and non-Asian patients in the external validation
5. Number of participants without GON was low in the validation datasets
6. Inter-grader and intra-grader variability in ground truth labeling
7. Tried only in clinic-based samples to replicate ophthalmologists' grading