Journal Pre-proof

REFUGE Challenge: A Unified Framework for Evaluating Automated Methods for Glaucoma Assessment from Fundus Photographs

José Ignacio Orlando, Huazhu Fu, João Barbossa Breda, Karel van Keer, Deepti R. Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, Joonseok Lee, Xiaoxiao Li, Peng Liu, Shuai Lu, Balamurali Murugesan, Valery Naranjo, Sai Samarth R. Phaye, Sharath M. Shankaranarayana, Apoorva Sikka, Jaemin Son, Anton van den Hengel, Shujun Wang, Junyan Wu, Zifeng Wu, Guanghui Xu, Yongli Xu, Pengshuai Yin, Fei Li, Xiulan Zhang, Yanwu Xu, Hrvoje Bogunović

PII: S1361-8415(19)30110-0
DOI: https://doi.org/10.1016/j.media.2019.101570
Reference: MEDIMA 101570

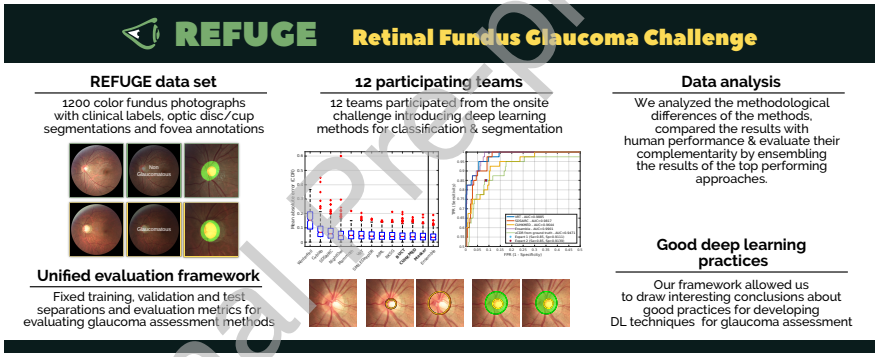
To appear in: *Medical Image Analysis*

Received date: 5 April 2019
Revised date: 26 July 2019
Accepted date: 1 October 2019

Please cite this article as: José Ignacio Orlando, Huazhu Fu, João Barbossa Breda, Karel van Keer, Deepti R. Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, Joonseok Lee, Xiaoxiao Li, Peng Liu, Shuai Lu, Balamurali Murugesan, Valery Naranjo, Sai Samarth R. Phaye, Sharath M. Shankaranarayana, Apoorva Sikka, Jaemin Son, Anton van den Hengel, Shujun Wang, Junyan Wu, Zifeng Wu, Guanghui Xu, Yongli Xu, Pengshuai Yin, Fei Li, Xiulan Zhang, Yanwu Xu, Hrvoje Bogunović, REFUGE Challenge: A Unified Framework for Evaluating Automated Methods for Glaucoma Assessment from Fundus Photographs, *Medical Image Analysis* (2019), doi: https://doi.org/10.1016/j.media.2019.101570

Highlights

- REFUGE, the first challenge on glaucoma assessment from fundus images.
- A data set of 1200 images with reliable ground truth labels is publicly released.
- An evaluation setup for glaucoma classification and optic disc/cup segmentation.
- 12 deep learning methods participating in the challenge are evaluated and discussed.
- Good practices are identified based on the outcomes of the participating teams.



REFUGE Challenge: A Unified Framework for Evaluating Automated Methods for Glaucoma Assessment from Fundus Photographs

José Ignacio Orlando[a,*], Huazhu Fu[b], João Barbossa Breda[c,d], Karel van Keer[d], Deepti R. Bathula[e], Andrés Diaz-Pinto[f], Ruogu Fang[g], Pheng-Ann Heng[h], Jeyoung Kim[i], JoonHo Lee[j], Joonseok Lee[j], Xiaoxiao Li[k], Peng Liu[g], Shuai Lu[l], Balamurali Murugesan[m], Valery Naranjo[f], Sai Samarth R. Phaye[e], Sharath M. Shankaranarayana[n], Apoorva Sikka[e], Jaemin Son[o], Anton van den Hengel[p], Shujun Wang[h], Junyan Wu[q], Zifeng Wu[p], Guanghui Xu[r], Yongli Xu[l], Pengshuai Yin[r], Fei Li[s], Xiulan Zhang[s], Yanwu Xu[t], Hrvoje Bogunović[a]

[a] *Christian Doppler Laboratory for Ophthalmic Image Analysis (OPTIMA), Vienna Reading Center (VRC), Department of Ophthalmology and Optometry, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria.*
[b] *Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates.*
[c] *Surgery and Physiology Department, Ophthalmology Unit, Faculty of Medicine, University of Porto, Porto, Portugal.*
[d] *Research Group Ophthalmology, KU Leuven, Leuven, Belgium*
[e] *Department of Computer Science & Engineering at Indian Institute of Technology (IIT) Ropar, Rupnagar, 140001 Punjab, India.*
[f] *Instituto de Investigación e Innovación en Bioingeniería, I3B, Universitat Politècnica de València, 46022 Valencia, Spain.*
[g] *J. Crayton Pruitt Family Dept. of Biomedical Engineering, University of Florida, 32611 USA.*
[h] *Department of Computer Science and Engineering, The Chinese University of Hong Kong, 999077 Hong Kong.*
[i] *Gachon University, 461-701 Gyeonggi-do, Korea.*
[j] *Samsung SDS AI Research Center, 06765 Seoul, Korea.*
[k] *Yale University, 06510 New Haven, CT USA.*
[l] *Faculty of Science, Beijing University of Chemical Technology, 100029 Beijing, China.*
[m] *Healthcare Technology Innovation Centre, IIT-Madras, India.*
[n] *Department of Electrical Engineering, IIT-Madras, India.*
[o] *VUNO Inc., Seoul, 137-810 Korea.*
[p] *Australian Institute for Machine Learning, Australia.*
[q] *Cleerly Inc. 10022 New York City, NY USA.*
[r] *South China University of Technology, 510006 Guangzhou, China.*
[s] *Zhongshan Ophthalmic Center, Sun Yat-sen University, China.*
[t] *Artificial Intelligence Innovation Business, Baidu Inc., China and Cixi Institute of BioMedical Engineering, Chinese Academy of Sciences, China.*

Abstract

Glaucoma is one of the leading causes of irreversible but preventable blindness in working age populations. Color fundus photography (CFP) is the most cost-effective imaging modality to screen for retinal disorders. However, its application to glaucoma has been limited to the computation of a few related biomarkers such as the vertical cup-to-disc ratio. Deep learning approaches, although widely applied for medical image analysis, have not been extensively used for glaucoma assessment due to the limited size of the available data sets. Furthermore, the lack of a standardize benchmark strategy makes difficult to compare existing methods in a uniform way. In order to overcome these issues we set up the Retinal Fundus Glaucoma Challenge, REFUGE (`https://refuge.grand-challenge.org`), held in conjunction with MICCAI 2018. The challenge consisted of two primary tasks, namely optic disc/cup segmentation and glaucoma classification. As part of REFUGE, we have publicly released a data set of 1200 fundus images with ground truth segmentations and clinical

*Corresponding authors: Yanwu Xu (`ywxu@ieee.org`) and Xiulan Zhang (`zhangxl2@mail.sysu.edu.cn.`).

glaucoma labels, currently the largest existing one. We have also built an evaluation framework to ease and ensure fairness in the comparison of different models, encouraging the development of novel techniques in the field. 12 teams qualified and participated in the online challenge. This paper summarizes their methods and analyzes their corresponding results. In particular, we observed that two of the top-ranked teams outperformed two human experts in the glaucoma classification task. Furthermore, the segmentation results were in general consistent with the ground truth annotations, with complementary outcomes that can be further exploited by ensembling the results.

Keywords: Glaucoma, Fundus photography, Deep Learning, Image segmentation, Image classification

---

List of abbreviations

- abs: Absolute value.
- Acc: Accuracy.
- AMD: Age-related Macular Degeneration.
- ASPP: Atrous Spatial Pyramid Pooling.
- AUC: Area Under the (ROC) Curve.
- CFP: Color Fundus Photograph.
- CLAHE: Contrast Limited Adaptive Histogram Equalization
- CONV: Convolutional layer.
- DR: Diabetic Retinopathy.
- DSC: Dice coefficient.
- FC: Fully Connected layer.
- FCN: Fully Convolutional Network.
- FDA: US Food and Drug Administration
- FN: False Negatives.
- FOV: Field-Of-View.
- FP: False Positives.

- G: Glaucoma.
- HSV: Hue Saturation Value.
- IOP: Intra Ocular Pressure.
- IoU: Intersection over Union / Jaccard index.
- NTG: Normal Tension Glaucoma.
- MAE: Mean Absolute Error.
- MICCAI: Medical Imaging and Computer Assisted Invervention conference.
- OC: Optic Cup.
- OCT: Optical Coherence Tomography.
- OD: Optic Disc.
- ONH: Optic Nerve Head.
- OMIA: Ophthalmic Medical Image Analysis workshop.
- POAG: Primary Open Angle Glaucoma.
- PPA: Peripapillary Atrophy.
- Pr: Precision / Positive predictive value.
- REFUGE: Retinal Fundus Glaucoma challenge.
- RGB: Red Green Blue.
- RNFL: Retinal Nerve Fiber Layer.
- ROC: Receiver-Operating Characteristic curve.
- ROI: Region Of Interest.
- Se: Sensitivity.
- SMOTE: Synthetic Minority Oversampling Technique.
- Sp: Specificity / True negative ratio.
- TN: True Negatives.
- TP: True Positives.
- vCDR: Vertical Cup-to-Disc Ratio.

1. Introduction

Glaucoma is a chronic neuro-degenerative condition that is one of the leading causes of irreversible but preventable blindness in the world (Tham et al., 2014). In 2013, 64.3 million people aged 40-80 years were estimated to suffer from glaucoma, while this number is expected to increase to 76 million by 2020 and 111.8 million by 2040 (Tham et al., 2014). In its many variants, glaucoma is characterized by the damage of the optic nerve head (ONH), typically caused by a high intra-ocular pressure (IOP). IOP is increased as a consequence of abnormal accumulation of aqueous humor in the eye, induced by pathological defects in the eye's drainage system. When the anterior segment is saturated with this fluid, the IOP progressively elevates, compressing the vitreous to the retina. If this remains uncontrolled, it can produce damage in the nerve fiber layer, the vasculature and the ONH, leading to a progressive and irreversible vision loss that can ultimately result in blindness. As this process occurs asymptomatically, glaucoma is frequently referred as the *"silent thief of sight"* (Schacknow and Samples, 2010): patients are not aware of the progressing disease until the vision is irreversibly lost.

Life-long pharmacological treatments based on the regular administration of eye drops are usually prescribed to control the IOP and to temper further damage in the retina. Alternatively, laser procedures and other surgeries can be performed to increase the drainage. In any case, early detection is essential to prevent vision loss (Schacknow and Samples, 2010). Unfortunately, at least half of patients with glaucoma currently remain undiagnosed (Prokofyeva and Zrenner, 2012). Being glaucoma a chronic condition, one of the major challenges is to be able to detect this large number of undiagnosed patients (Prokofyeva and Zrenner, 2012). Generalized screening programs have not been employed because of the large amount of false positives these can generate. These misdiagnoses cannot be absorbed by current healthcare infrastructures and would have an unnecessary negative impact on the patient's quality of life, until it would be recognized that no glaucomatous neuropathy existed (Schacknow and Samples, 2010).

Color fundus photography (CFP, Figure 1) is currently the most economical, non-invasive imaging modality for inspecting the retina (Abràmoff et al., 2010; Schmidt-Erfurth et al., 2018). Its widespread availability makes it ideal for assessing several ophthalmic diseases such as age-related macular degeneration (AMD) (Burlina et al., 2017), diabetic retinopathy (DR) (Gulshan et al., 2016) and glaucoma (Li et al., 2018b). Screening campaigns can be aided by the incorporation of computer-assisted tools for image-based diagnosis. As these initiatives require to manually grade a large number of cases in a short period of time, automated tools can help clinians by providing them with quantitative and/or qualitative feedback (e.g. disease likelihood, segmentations of relevant lesions and pathological structures, etc). These approaches have already been successfully applied for detecting DR, in a FDA-approved autonomous diagnostic system, a first of its kind (Abràmoff et al., 2018). However, the broad application of similar methods for glaucoma detection is still pending. This is partially due to the fact that the earlier signs of glaucoma are not so

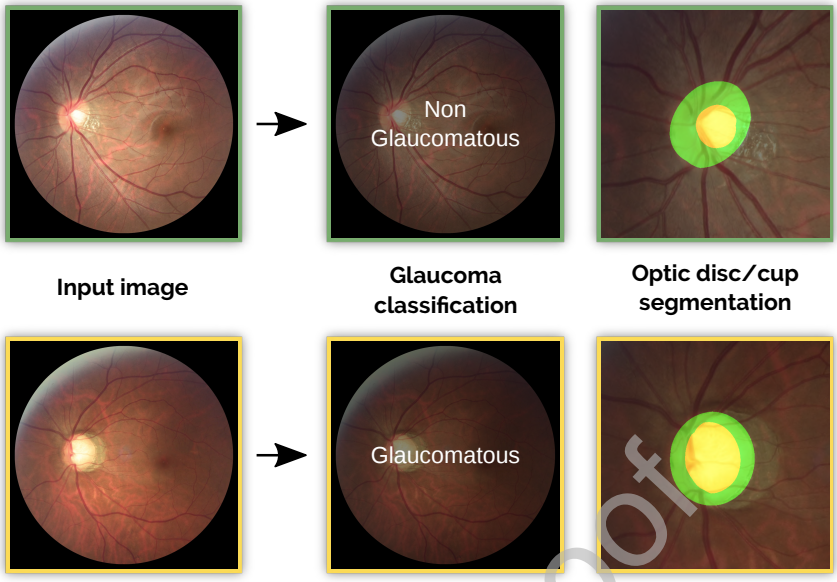


Figure 1: REFUGE challenge tasks: glaucoma classification and optic disc/cup segmentation from color fundus photographs.

easily recognizable in CFP (Lavinsky et al., 2017) (Figure 2). In current best clinical practice, CFPs are complementary to other studies such as IOP measurements, automated perimetry and optical coherence tomography (OCT). This approach is not cost-effective to be applied for large scale population screening for glaucoma (Schacknow and Samples, 2010). Therefore, developing automated tools to better exploit the information in CFP is paramount to reduce this burden and ensure an effective detection of glaucoma suspects.

A significant research effort has been made to introduce automated tools for segmenting the optic disc (OD) and the optic cup (OC) in CFP automatically, or to identify glaucomatous cases based on alternative features (Almazroa et al., 2015; Haleem et al., 2013; Thakur and Juneja, 2018). Nevertheless, these approaches currently cannot be properly compared due to the lack of a unified evaluation framework to validate them. Moreover, the absence of large scale public available data sets of labeled glaucomatous images has hampered the rapid deployment of deep learning techniques for glaucoma detection (Hagiwara et al., 2018). It has been recently shown that image analysis competitions in general can aid to identify challenging scenarios that need further development (Prevedello et al., 2019). Recent grand challenges such as ROC (Niemeijer et al., 2010), Kaggle (Kaggle, 2015) and IDRiD (Porwal et al., 2018), on the other hand, have shown to be useful to address both inconveniences in DR (Schmidt-Erfurth et al., 2018), favoring the deployment of these tools into the daily clinical practice (Abràmoff et al., 2018). Unfortunately, similar initiatives have not been introduced for glaucoma detection and/or assessment yet.

In an effort to overcome these limitations, we introduced the Retinal Fundus Glaucoma Challenge

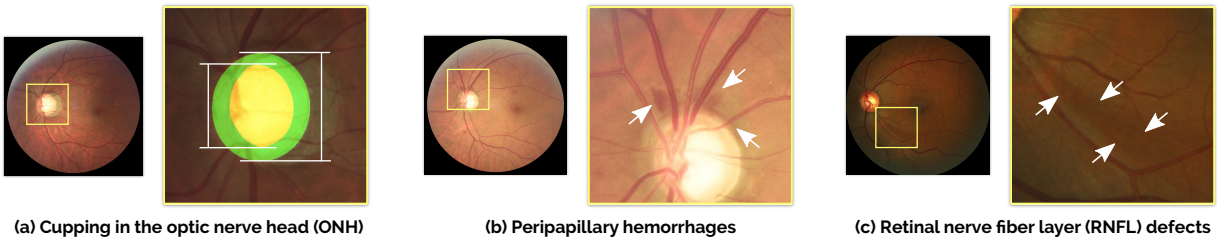


Figure 2: Pathological changes typical from glaucoma, as observed through fundus photography. (a) Neuroretinal rim thinning due to cupping in the optic nerve head (ONH). White lines indicate the vertical diameter of the optic disc (green) and the optic cup (yellow). (b) Peripapillary hemorrhages, observed as flame-shaped bleedings in the vicinity of the ONH. (c) Retinal nerve fiber layer defects are observed as subtle striations spanning from the optic disc border.

(REFUGE), a competition that was held as part of the Ophthalmic Medical Image Analysis (OMIA) workshop at MICCAI 2018. The key contributions of the challenge were: (i) the release of a large database (approximately two times bigger than the largest available so far) of 1200 CFP with reliable reference standard annotations for glaucoma identification, optic disc/cup (OD/OC) segmentation and fovea localization; and (ii) the constitution of a unified evaluation framework that enables a standardized fair protocol to compare different algorithms. To the best of our knowledge, REFUGE is the first initiative to provide these key tools at such a large scale. REFUGE participants were invited to use the data set to train and evaluate their algorithms for glaucoma classification and OD/OC segmentation. Their results were quantitatively evaluated using our uniform protocol, to ensure a fair comparison.

In this paper, we analyze the outcomes and the methodological contributions of REFUGE. We present and describe the challenge, reporting the performance of the best algorithms evaluated in the competition and identifying successful common practices for solving the proposed tasks. The results are contrasted with the outcomes of two glaucoma experts to study their performance with respect to independent human observers. Finally, we take advantage of all these empirical evidence to discuss the clinical implications of the results and to propose further improvements to this evaluation framework. In line with the recommendations of Trucco et al. (2013), REFUGE data and evaluation remain open to encourage further developments and ensure a proper and fair comparison of those new proposals.

2. Automated glaucoma assessment: state-of-the-art and current evaluation protocols

Early attempts for glaucoma classification and OD/OC segmentation were mostly based on hand-crafted methods using a combination of feature extraction techniques and supervised or unsupervised machine learning classifiers (Almazroa et al., 2015; Haleem et al., 2013; Thakur and Juneja, 2018). However, their accuracy was limited due to the application of manually designed features, which are unable to comprehensively characterize the large variability of disease appearance. Deep learning techniques, on the contrary,

automatically learn these characteristics by exploiting the implicit information of large training sets of annotated images (Litjens et al., 2017). In this section we briefly analyze the state-of-the-art techniques for glaucoma classification and OD/OC segmentation and their main evaluation issues. The interested reader could refer to the surveys by Almazroa et al. (2015), Haleem et al. (2013) and Thakur and Juneja (2018) for a comprehensive analysis of the previous non-deep learning based approaches.

2.1. Glaucoma classification

Glaucoma classification consists in categorizing an input CFP into glaucomatous or non-glaucomatous, based on its visual characteristics. A summary table of the most recent deep learning methods introduced for this task is available in the Supplementary Materials. In general, most of the existing approaches are based on adaptations of standard deep supervised learning techniques, customized to deal with small training sets (Section 2.3). Chen et al. (2015a), Chen et al. (2015b) and Raghavendra et al. (2018) proposed to use shallow architectures with a limited number of layers. This is useful to prevent overfitting but limits the ability of the networks to learn rare, specific features. Alternatively, the studies by Christopher et al. (2018), Li et al. (2018a) and Orlando et al. (2017b) used transfer learning methods, based on deeper architectures but pre-trained on non-medical data. Christopher et al. (2018) fine-tuned a network initialized with weights learned from ImageNet (Russakovsky et al., 2015) to detect glaucomatous optic neuropathy. Similarly, transfer learning was shown by Gómez-Valverde et al. (2019) to outperform networks trained from scratch for glaucoma detection. Both studies applied a massive image data set with more than 14.000 images to fine tune these networks. Other works such as those by Orlando et al. (2017b) and Li et al. (2018a) used deep learning features extracted from the last fully connected layers of pre-trained networks. The classification task was then performed using linear classifiers trained with these features (Li et al., 2018a; Orlando et al., 2017b). This allows to use smaller data sets, although at the cost of lower performance.

Another widely used approach is to restrict the area of analysis to the ONH. This region is the one that is mostly affected by glaucoma, and focusing only there allows for a better exploitation of model parameters. This was done by most of the surveyed methods (as observed in Table 1 from the Supplementary Materials) and it resulted in a better performance than when learning from full size images. However, such a strong restriction in the networks' field of view hampers their ability to learn alternative features from other regions (Chen et al., 2015a).

2.2. Optic disc/cup segmentation

Segmenting the OD and the OC from CFPs is a challenging but relevant task that helps to assess glaucomatous damage to the ONH (Haleem et al., 2013). Automated methods have to be robust against complex pathological changes such as peripapillary atrophies (PPA) or hemorrhages (Almazroa et al., 2015; Thakur and Juneja, 2018) (Figure 2 (b)). On the other hand, the accurate delineation of the OC is specially

difficult due to the high vessel density in the area and the lack of depth information in CFP (Miri et al., 2015). Alternative features such as vessels bendings (Joshi et al., 2011) or intensity changes (Xu et al., 2014) have been studied in the past to approximate the ONH depth. The interested reader could refer to Table 2 from the Supplementary Materials for a summary of current deep learning approaches for simultaneous OD/OC segmentation.

Most of existing methods use a surrogate segmentation/detection approach to first localize the ONH area and them crop the images around it (Edupuganti et al., 2018; Fu et al., 2018; Lim et al., 2015; Sevastopolsky, 2017; Zilly et al., 2015). This prevents false positive detections in regions with e.g. severe illumination artifacts and grants a better exploitation of model parameters, as they are only dedicated to characterize the local appeareance of the OD/OC and not to differentiate these structures from other fundus regions. Alternatively, a two-stage approach was followed by Sevastopolsky et al. (2018), using a first neural network to retrieve a coarse segmentation and a second one to refine the results.

Different neural network architectures have been proposed for OD/OC segmentation. Lim et al. (2015) applied a classification network similar to LeNet (LeCun et al., 1998) at a patch level to classify its central pixel as belonging to the OD, the OC or the background. Using patches as training samples artificially increases the available training data, although at the cost of loosing spatial information. Alternatively, Zilly *et al.* proposed to overcome the data limitation issue by training a convolutional neural network using an entropy sampling approach instead of gradient descent. Most of the recent methods (Al-Bander et al., 2018; Edupuganti et al., 2018; Fu et al., 2018; Sevastopolsky, 2017; Sevastopolsky et al., 2018), however, are based on modifications to the original U-Net architecture (Ronneberger et al., 2015). This is due to the fact that this network can achieve good results even when trained using a relatively small amount of images. Architecture changes that heavily increase the capacity of the networks such as those introduced by Edupuganti et al. (2018) usually demand the application of transfer learning in the encoding path. In addition, heavy data augmentation through different combination of image transformations has also been explored (Fu et al., 2018; Sun et al., 2018).

2.3. Evaluation protocols

Large discrepancies in the evaluation protocols were observed in the surveyed literature, regardless of the target task. These differences (summarized in Tables 1 and 2 of the Supplementary Materials), are mostly related with two key aspects: (i) the data sets used for training/evaluation, and (ii) the evaluation metrics.

2.3.1. Data sets

Table 1 summarizes the public available data sets of CFPs for glaucoma classification and/or OD/OC segmentation used by the literature. The REFUGE database (Section 3.1) is included for comparison purposes.

Table 1: Comparison of the REFUGE challenge data set with other publicly available databases of color fundus images. Question marks indicate missing information, and N/A stands for "not applicable".

Dataset	**Num. of images**			**Ground truth labels**			**Different cameras**	**Training & test split**	**Diagnosis from**	**Evaluation framework**
	Glaucoma	**Non glaucoma**	**Total**	**Glaucoma classification**	**Optic disc/cup (assessed on CFP)**	**Fovea localization**				
ARIA (Zheng et al., 2012)	0	143	143	No	Yes/No	Yes	No	No	?	No
DRIONS-DB (Carmona et al., 2008)	-	-	110	No	Yes/No	No	?	No	N/A	No
DRISHTI-GS (Sivaswamy et al., 2014, 2015)	70	31	101	Yes	Yes/Yes	No	No	Yes	Image	No
DR HAGIS (Holm et al., 2017)	10	29	39	Yes	No/No	No	Yes	No	Clinical	No
IDRiD (Porwal et al., 2018)	0	516	516	No	Yes/No	Yes	No	Yes	?	Yes
HRF (Odstrčilík et al., 2013)	15	30	45	Yes	No/No	No	No	No	Clinical	No
LES-AV (Orlando et al., 2018)	11	11	22	Yes	No/No	No	No	No	Clinical	No
ONHSD (Lowell et al., 2004)	-	-	99	No	Yes/No	No	No	No	N/A	No
ORIGA (Zhang et al., 2010)	168	482	650	Yes	Yes/Yes	No	?	No	?	No
RIM-ONE (Fumero et al., 2011) v1	40	118	158	Yes	Yes/No	No	No	No	Clinical	No
RIM-ONE (Fumero et al., 2011) v2	200	255	455	Yes	Yes/No	No	No	No	Clinical	No
RIM-ONE (Fumero et al., 2011) v3	74	85	169	Yes	Yes/No	No	No	No	Clinical	No
RIGA (Almazroa et al., 2018)	-	-	750	No	Yes/Yes	No	Yes	No	?	No
REFUGE	**120**	**1080**	**1200**	**Yes**	**Yes/Yes**	**Yes**	**Yes**	**Yes**	**Clinical**	**Yes**

In general, we observed that a lack of pre-defined partitions into training and test sets has induced a chaotic practical application of the existing data. As discussed by Trucco et al. (2013), this affect the feasibility of directly comparing the performance of existing methods, difficulting to conclude which model characteristics are more appropriate to solve each task. To the best of our knowledge, DRISHTI-GS[1] (Sivaswamy et al., 2014, 2015) is the only existing database for glaucoma assessment that provides a clear training/test split.

Another important aspect is related with the reliability of the assigned diagnostic labels. Some public data sets such as DRISHTI-GS provide glaucoma labels that were assigned based only on image characteristics. This has been also observed in private data sets such as those used by Christopher et al. (2018) and Li et al. (2018b), which were built using images from Internet that were manually graded based on their visual appeareance, without additional clinical information. Surprisingly, no information about the source of the diagnostic labels is provided in most of existing databases (see Table 1). Using images with labels that were not assigned using retrospective analysis of clinical records can be problematic as it might bias automated methods to reproduce wrong labelling practices. On the contrary, clinical labels can aid algorithms to learn and discover other supplemental manifestations of the disease that are still unknown or that are too difficult to distinguish with the naked eye.

The amount of images and their diversity is also an important aspect to consider. In particular, existing databases rarely include images obtained from different acquisitions devices, ethnicities or presenting challenging glaucoma related scenarios. Therefore, the learned models might exhibit a weak generalization ability. To partially bypass this issue, some authors have proposed to train their methods using combinations of different data sets (Cerentinia et al., 2018; Pal et al., 2018).

As indicated in Table 1, all existing data sets with OD/OC annotations contain manually assigned labels obtained from the CFP, without considering depth information and performed by a single reader. Consequently, these segmentations might suffer from deviations that could bias the subsequent evaluations. Incorporating depth information e.g. through stereo imaging or OCT would ensure much trustworthy annotations. On the other hand, providing segmentations obtained by the consensus of multiple readers could better approximate the true anatomy by reducing inter-observer variability.

Finally, it is important to highlight the lack of a large public data set providing both OD/OC segmentations and clinical diagnostics simultaneously. ONHSD[2] (Lowell et al., 2004) and DRIONS-DB[3] (Carmona et al., 2008) only include segmentations of the OD, and no glaucoma labels are given. ARIA[4] (Zheng et al., 2012) provides OD segmentations and incorporates vessel segmentations and annotations of the fovea center.

[1] http://cvit.iiit.ac.in/projects/mip/drishti-gs/mip-dataset2/Home.php
[2] http://www.aldiri.info/Image%20Datasets/ONHSD.aspx
[3] http://www.ia.uned.es/~ejcarmona/DRIONS-DB.html
[4] https://eyecharity.weebly.com/aria_online.html

However, the images correspond to normal subjects and patients with DR and AMD, and no segmentations of the OC are included. DR HAGIS[5] (Holm et al., 2017), HRF[6] (Odstrčilík et al., 2013) and LES-AV[7] (Orlando et al., 2018), on the other hand, include reliable diagnostic labels and vessel segmentations, but no labels for the OD/OC. Moreover, their size is relatively small (39, 45 and 22 images, respectively). RIGA[8] (Almazroa et al., 2018) is a recent data set that contains 750 fundus images with OD/OC segmentations but without glaucoma labels. The three releases of RIM-ONE (v1, v2 and v3) (Fumero et al., 2011) provide image-level glaucoma labels and OD segmentations. RIM-ONE v1 and v2 include CFPs cropped around the ONH. Furthermore, RIM-ONE v1 incorporate OD annotations by five different experts and image level labels for control subjects, ocular hypertensive patients and subjects with early, moderate and deep glaucoma. RIM-ONE v2 and v3, on the contrary, only include OD segmentations by two experts, and the diagnostic labels are classified into normal and glaucoma suspect cases. Moreover, RIM-ONE v3 do not include typical CFPs but stereo images. To the best of our knowledge, only DRISHTI-GS and ORIGA (Zhang et al., 2010) include both glaucoma classification labels and OD/OC segmentations. The diagnostic labels in DRISHTI-GS, however, were assigned solely based on the images (Sivaswamy et al., 2015). ORIGA, on the other hand, is not publicly available anymore.

2.3.2. Metrics

Most of the literature in glaucoma classification uses receiver-operating characteristic (ROC) curves (Davis and Goadrich, 2006) for evaluation, including the area under the curve (AUC) as a summary value (Chen et al., 2015a,b; Christopher et al., 2018; Fu et al., 2018; Gómez-Valverde et al., 2019; Orlando et al., 2017b; Li et al., 2018a,b; Liu et al., 2018; Pal et al., 2018). Sensitivity and specificity (Chen et al., 2015b; Christopher et al., 2018; Fu et al., 2018; Gómez-Valverde et al., 2019; Li et al., 2018a; Liu et al., 2018) are also used in different studies to complement the AUC when targetting binary classification outcomes. Accuracy was reported in (Cerentinia et al., 2018; Raghavendra et al., 2018) as another evaluation metric, although this metric might be biased if the proportion of non-glaucomatous images is significantly higher than the glaucomatous ones (Orlando et al., 2017a). To overcome this limitation, Fu et al. (2018) used a balanced accuracy, consisting on the average between sensitivity and specificity.

Current literature in OD/OC segmentation make use of classical overlap metrics such as the intersection-over-union (IoU, also known as Jaccard index) (Al-Bander et al., 2018; Edupuganti et al., 2018; Fu et al., 2018; Lim et al., 2015; Sevastopolsky, 2017; Sevastopolsky et al., 2018; Sun et al., 2018; Zilly et al., 2015) and the Dice index (Al-Bander et al., 2018; Edupuganti et al., 2018; Sevastopolsky, 2017; Sevastopolsky et al., 2018; Sun et al., 2018; Zilly et al., 2015). Although different by definition, these two metrics can

[5] https://personalpages.manchester.ac.uk/staff/niall.p.mcloughlin/

[6] https://www5.cs.fau.de/research/data/fundus-images/

[7] https://ignaciorlando.github.io/data/LES-AV.zip

[8] https://deepblue.lib.umich.edu/data/concern/data_sets/3b591905z

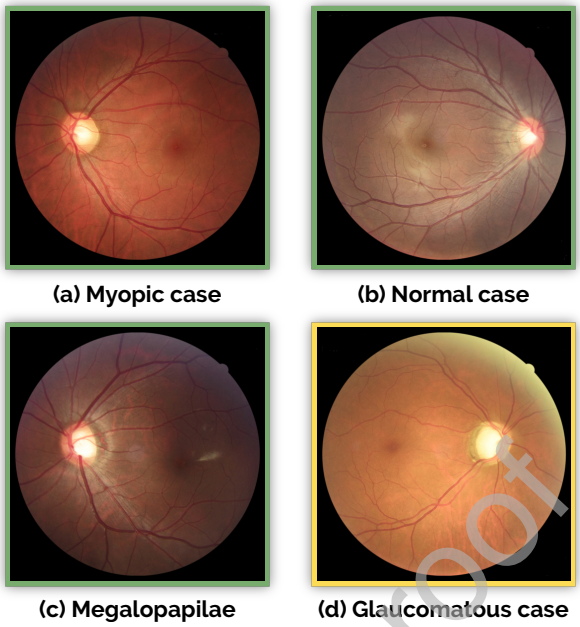


Figure 3: Representative examples of color fundus photographs from the REFUGE data set. Non-glaucomatous (green) and glaucomatous (yellow) groups. (a) Myopic case with enlarged optic cup. (b) Healthy subject. (c) Patient with megalopapilae. (d, yellow) Glaucomatous case with cupping.

be computed from each other, as they are defined as ratios of overlap between the predicted area and the manual reference annotation (Taha and Hanbury, 2015). Pixelwise sensitivity and specificity values have been also reported in (Al-Bander et al., 2018; Fu et al., 2018) to illustrate the behavior in terms of false negatives and false positives, respectively. Finally, the accuracy for segmenting both the OD and the OC has been simultaneously assessed by means of the mean absolute error (MAE) of the estimated vs. manually graded CDR values (Fu et al., 2018; Lim et al., 2015; Sun et al., 2018).

All these metrics are well-known and were previously used in several domains. However, it is still necessary to come up with a uniform evaluation criteria to assist method comparison and prevent the usage of potentially biased metrics.

3. The REFUGE challenge

This section briefly describes REFUGE challenge, introducing the released data set (Section 3.1) and the proposed evaluation procedure (Section 3.2).

3.1. REFUGE database

The REFUGE challenge database consists of 1200 retinal CFPs stored in JPEG format, with 8 bits per color channel, acquired by ophthalmologists or technicians from patients sitting upright and using one of

two devices: a Zeiss Visucam 500 fundus camera with a resolution of 2124×2056 pixels (400 images) and a Canon CR-2 device with a resolution of 1634×1634 pixels (800 images). The images are centered at the posterior pole, with both the macula and the optic disc visible, to allow the assessment of the ONH and potential retinal nerve fiber layer (RNFL) defects. These pictures correspond to Chinese patients (52% and 55% female in offline and online test sets, respectively) visiting eye clinics, and were retrieved retrospectively from multiple sources, including several hospitals and clinical studies. Only high-quality images were selected to ensure a proper labelling, and any personal and/or device information was removed for anonymization.

Each image in the REFUGE data set includes a reference, trustworthy glaucomatous / non-glaucomatous label. These diagnostics were assigned based on the comprehensive evaluation of the subjects' clinical records, including follow-up fundus images, IOP measurements, optical coherence tomography images and visual fields (VF). The glaucomatous cases correspond to subjects with glaucomatous damage in the ONH area and reproducible glaucomatous VF defects. This last characteristic was defined as a reproducible reduction in sensitivity compared to the normative data set, in reliable tests, at: (1) two or more contiguous locations with p-value < 0.01 and (2) three or more contiguous locations with p-value < 0.05. ONH damage was defined as a vCDR > 0.7, thinning of the RNFL, or both, without a retinal or neurological cause for VF loss. Notice, then, that instead of using labels assigned based on a single CFP at a specific timepoint, the labels were retrieved from examinations of follow-up medical records using a pre-determined criterion, to ensure the reliability of the classification labels. 10% of the dataset (120 samples) corresponds to glaucomatous subjects, including Primary Open Angle Glaucoma (POAG) and Normal Tension Glaucoma (NTG). This proportion of diseased cases deviates from the global prevalence of glaucoma (≈ 4 % for populations aged 40-80 years (Tham et al., 2014)). However, reducing the size of the glaucoma set would have negatively affected the ability of the classification approaches to learn features from the diseased cases. Furthermore, in an effort to model a more representative clinical scenario, the non-glaucomatous set was designed to include not only normal healthy cases but also patients with non-glaucomatous conditions such as diabetic retinopathy, myopia and megalopapilae. Myopic and megalopapilae cases were included as subjects suffering from them can easily be missclassified as glaucomatous due to their aberrant ONH appeareance (Figure 3).

Manual annotations of the OD and the OC were provided by seven independent glaucoma specialists from the Zhongshan Ophthalmic Center (Sun Yat-sen University, China), with an average experience of 8 years in the field (ranging from 5 to 10 years). All the ophthalmologists independently reviewed and delineated the OD/OC in all the images, without having access to any patient information or knowledge of disease prevalence in the data. The annotation procedure consisted in manually drawing a tilted ellipse covering the OD and the OC, separately, by means of a free annotation tool with capabilities for image review, zoom and ellipse fitting. A single segmentation per image was afterwards obtained by taking the majority voting of the anotations of the seven experts. A senior specialist with more than 10 years of experience in glaucoma performed a quality check afterwards, analyzing the resulting masks to account for

Table 2: Summary of the main characteristics of each subset of the REFUGE data set.

Characteristics	**Subset**		
	Training	**Offline test set**	**Online test set**
Acquisition device	Zeiss Visucam 500	Canon CR-2	
Resolution	2124×2056	1634×1634	
Num. images	400	400	400
Glaucoma/Non glaucoma	40/360	40/360	40/360
Public labels?	✓	✗	✗

potential mistakes. When errors in the annotations were observed, this additional reader analyzed each of the seven segmentations, removed those that were considered failed in his/her opinion and repeated the majority voting process with the remaining ones. Only a few cases had to be corrected using this protocol.

Manual pixel-wise annotations of the fovea were also assigned to the images to complement the data set. The fovea position was fixed by the seven independent glaucoma specialists, and a reference standard was created taking the average of these annotations.

The entire set was divided into three fixed subsets: training, offline and online test sets, each of them stratified in such a way that they contain an equal proportion of glaucomatous (10%) and non-glaucomatous (90%) cases. Table 2 summarize the main characteristics of each subset. The training set contains all the images acquired with the Zeiss Visucam 500 camera, while the offline and online test sets include the lower resolution images captured with the Canon CR-2 device. This was made on purpose to encourage the teams to develop tools with enough generalization ability to deal with images acquired with at least using two different devices and at two different resolutions.

Figure 4 represents the distribution of vCDR and OD and OC areas of the images within each subset. To account for the differences in the field-of-view (FOV) of acquisitions from the Zeiss and Canon devices, the areas (in pixels) were normalized as a proportion of the FOV area (in pixels). The differences between groups were statistically assessed using Kruskal-Wallis tests with $\alpha = 0.01$. Statistical significant differences were only observed for the OD area ($p = 1.4 \times 10^{-7}$, explained by the training set having larger values than the offline and online test sets ($p < 0.0091$, two-tailed Wilcoxon rank sum tests with a Bonferroni corrected significance $\alpha = 0.025$ to account for the two comparisons).

3.2. Challenge Setup, Evaluation Metrics and Ranking Procedure

REFUGE was held in conjunction with the 5th Ophthalmic Medical Image Analysis (OMIA) workshop, during MICCAI 2018 (Granada, Spain). The challenge proposal was accepted after assessing the compliance to good practices proposed in (Maier-Hein et al., 2018; Reinke et al., 2018). Thereafter, REFUGE was

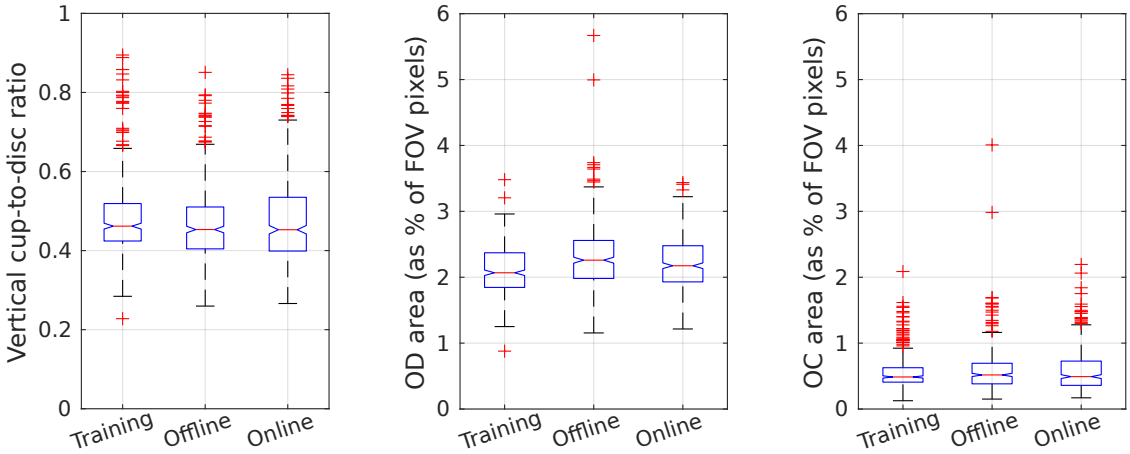


Figure 4: REFUGE data set characteristics in each of the challenge partitions (training set, offline test set and online test set). From left to right: vertical cup-to-disc ratio (vCDR) values, and optic disc and cup areas, as percentages of the field-of-view area.

announced in several platforms to maximize its visibility, including the MICCAI website, its associated mailing lists and on the Grand Challenges in Biomedical Image Analysis website. [9] The challenge was officially launched in June 2018 by releasing the training set (images and labels) on a dedicated website (`https://refuge.grand-challenge.org/Home/`). The registered teams were allowed to use the training set to learn and adjust their proposed algorithms for glaucoma classification, OD/OC segmentation and, optionally, for fovea detection. We will not focus on this last task as it was not mandatory for participating on the challenge, and therefore no team submitted results for it on the test set. The registered teams were allowed to use any other public data set for developing their methods, provided that they were easily accessible by everyone.

The offline test set set (only the images, without labels) was released on July 2018, and the participants were invited to submit their results for an offline validation. Each participant could receive a maximum of five evaluations on this set. Each task was evaluated separately according to a uniform criteria. In particular:

3.2.1. Glaucoma classification:

The teams submitted a table with a glaucoma likelihood per each image on the set. A receiver operating characteristics (ROC) curve was created based on the gold standard glaucoma diagnostic, and the area under the curve (AUC) was used as a ranking score for the classification task, S_{class} (the higher, the better). Additionally, a reference sensitivity $\text{Se} = \frac{\text{TP}}{\text{TP+FN}}$ value at a specificity $\text{Sp} = \frac{\text{TN}}{\text{TN+FP}}$ of 0.85 was also reported, with TP, FP, TN and FN standing for true/false positives and true/false negatives, respectively. This value

[9] `grand-challenge.org`

was not taken into account for the ranking, but allowed each team to assess the overall performance of the classification algorithm in a setting when a low number of false positives is tolerated.

3.2.2. OD/OC segmentation:

The teams submitted one segmentation file for each image. These files were encoded in grayscale BMP format where 0 corresponded to the optic cup, 128 to the optic disc and 255 elsewhere. The results were compared with the gold standard segmentation using the Dice index (DSC) for OD/OC separately, and the mean absolute error (MAE) of the vertical cup-to-disc ratio (vCDR) estimations. In particular, DSC define the overlap between two binary regions:

$$\mathrm{DSC}_k = 2\frac{|Y_k \cap \hat{Y}_k|}{|Y_k \cup \hat{Y}_k|} \tag{1}$$

where Y_k and $\hat{Y}_k$ are the ground truth and predicted segmentations of the region of interest k, respectively (with $k =$ OD or OC). On the other hand, MAE is defined as:

$$\mathrm{MAE} = \mathrm{abs}(\mathrm{vCDR}(\hat{Y}_{\mathrm{OC}}, \hat{Y}_{\mathrm{OD}}) - \mathrm{vCDR}(Y_{\mathrm{OC}}, Y_{\mathrm{OD}})) \tag{2}$$

where $\mathrm{vCDR}(\mathrm{OD}, \mathrm{OC}) = \frac{d(\mathrm{OC})}{d(\mathrm{OD})}$ is a function that estimates the vCDR based on the vertical diameter d of the segmentations of the OD and the OC, respectively. Each team was ranked using the average value of each of these metrics separately, resulting in three rank values $R_{\mathrm{segm}}^{\mathrm{DSC_{OD}}}$, $R_{\mathrm{segm}}^{\mathrm{DSC_{OC}}}$ and $R_{\mathrm{segm}}^{\mathrm{MAE}}$, and an overall segmentation score S_{segm} was assigned to each team based on the following weighted average:

$$S_{\mathrm{segm}} = 0.35 \times R_{\mathrm{segm}}^{\mathrm{DSC_{OD}}} + 0.25 \times R_{\mathrm{segm}}^{\mathrm{DSC_{OC}}} + 0.4 \times R_{\mathrm{segm}}^{\mathrm{MAE}}. \tag{3}$$

Notice that in this case, a lower S_{segm} value is better than a higher one. Since the MAE of the vCDR is calculated based on the segmentation of OC and OD, we set a larger weight for vCDR than to each individual segmentation term. Moreover, it is standard in the literature (Section 2) to first segment the OD region and then extract the OC from the cropped OD area. Hence, we assigned a larger weight to the OD segmentation results than to the OC.

An overall offline score was assigned to each method based on:

$$S_{\mathrm{val}} = 0.4 \times R_{\mathrm{class}} + 0.6 \times R_{\mathrm{segm}} \tag{4}$$

where R_{class} and R_{segm} are the team rank positions based on the classification and segmentation scores S_{class} and S_{segm}, respectively. A larger weight was assigned to the ranking for the segmentation task as the vCDR, derived from OD/OC segmentation, can be used as a primary score for glaucoma classification. An offline test set based leaderboard was created by setting a rank position R_{val} for each team, based on S_{val}.

Only those teams that submitted reports describing their proposed approaches were taken into account for this leaderboard. These reports can be easily accessed from the challenge website. [10]

The first 12 teams according to S_{val} were invited to attend to the on-site challenge, that was held in person at MICCAI. The test set (only the images) was released during the workshop, and the 12 teams had to submit their results before a time deadline (3 hours). The last submission of each team was taken into account for evaluation. Both an on-site rank and a final rank were assigned to each team. The on-site rank R_{test} was created using the scoring described in Eq. 4, while the final rank R_{final} was based on a score S_{final} calculated as the weighted average of the off-line and on-site rank positions:

$$S_{\text{final}} = 0.3 \times R_{\text{val}} + 0.7 \times R_{\text{test}}. \tag{5}$$

Notice that a higher weight was assigned to the results on the test set. In this paper we only focus on the results obtained on the test set, during the on-site challenge.

The evaluation was performed using a Python 3.6 open-source framework that was specially developed for the challenge and is publicly available. [11]

4. Results

This section presents the results on the REFUGE test set of the 12 teams that participated in the on-site challenge. The official final rankings according to the offline and online test set performances can be accessed on the REFUGE website.

4.1. Glaucoma classification

The participating methods for glaucoma classification are summarized in Table 3. Further details about each method are provided in the appendix. The evaluation of the classification task, in terms of AUC and the reference sensitivity at 85% specificity, is presented in Table 4. We also included an additional approach based on using the ground truth vCDR values as a glaucoma likelihood for classification. Figure 5 presents the ROC curves of the three top-ranked teams and the ground truth vCDR values. The curves for each participating method are available for downloading in the challenge website. Matt-Whitney U hypothesis tests (DeLong et al., 1988) with $\alpha = 0.05$ were performed using Vergara et al. (2008) tool, to compare the statistical significance of the differences in the AUC values of these top-ranked teams. VRT reported the best classification performance, achieving significantly better results that the ground truth vCDR ($p = 0.006$). Compared with SDSAIRC and CUHKMED–the second and third teams, respectively–the differences were only significant with respect to CUHKMED (CUHKMED: $p = 0.007$, SDSAIRC: $p = 0.187$). Both SDAIRC

[10] https://refuge.grand-challenge.org/Results-Onsite_TestSet/

[11] https://github.com/ignaciorlando/refuge-evaluation

Table 3: Summary of the glaucoma classification methods evaluated in the on-site challenge, in alphabetical order using the teams names.

Team	**Inputs**	**Architectures**	**Training set**	**Methodology**	**Post-processing**
AIML	Full image / ONH area	ResNet-50, -101, -152 (He et al., 2016), 38 (Wu et al., 2019)	REFUGE training set	Ensemble of glaucoma likelihoods from multiple networks pre-trained on ImageNet and fine-tuned on REFUGE training set	Ensemble by averaging
BUCT	ONH area, grayscale	Xception (Chollet, 2017)	REFUGE training set	Training from scratch on grayscale images	None
CUHKMED	OD/OC segmentation	None	None	vCDR values computed from ellipses fitted to automated OD/OC segmentations	None
Cvblab	Full image	VGG19 (Simonyan and Zisserman, 2014), Inception V3 (Szegedy et al., 2016), ResNet-50 (He et al., 2016), Xception (Chollet, 2017)	REFUGE training set, DRISHTI-GS, HRF, ORIGA and RIM-ONE r3	Ensemble of glaucoma likelihoods from multiple networks pre-trained on ImageNet and fine-tuned, classes in REFUGE training set balanced using SMOTE (Chawla et al., 2002)	Ensemble by averaging
Mammoth	ONH area with CLAHE	ResNet-18 (He et al., 2016) and CatGAN (Wang and Zhang, 2017)	Sample from REFUGE training set	Ensemble of ResNet models pre-trained on ImageNet and fine-tuned using REFUGE data and synthetic images generated with CatGAN	None
Masker	Full image	ResNet (He et al., 2016)	REFUGE training set and ORIGA	Linear combination of vCDR and predictions of multiple ResNet networks	Ensemble with vCDR
NightOwl	ONH area with/without exp. transform	Custom	REFUGE training set (10-fold cross-validation)	Ensemble of classification networks trained to predict glaucoma from features produced by the encoders of the segmentation networks	Ensemble by maximum
NKSG	Full image	SENet (Hu et al., 2018)	REFUGE training set (5-fold cross-validation)	SE-Net pretrained on images from Kaggle DR challenge (Kaggle, 2015) and fine-tuned on REFUGE data, best model from cross-validation taken for final prediction	None
SDSAIRC	Crop with ONH in upper-left corner	ResNet-50 (He et al., 2016)	REFUGE training set	Logistic regression classifier trained with vCDR values from OD/OC segmentation and output of ResNet-50 model fine-tuned from ImageNet	None
SmileDeepDR	ONH area	DeepLabv3+ (Chen et al., 2018)	REFUGE training set	Adaptation of a segmentation network to predict a glaucoma likelihood	None
VRT	Full image with custom mask for attention	Custom (Son et al., 2018)	Kaggle (Kaggle, 2015), MESSIDOR (Decencire et al., 2014) and IDRiD (Porwal et al., 2018)	Attention guided model trained on public data sets of DR images, weakly labelled using pre-trained models for glaucoma classification, RNFL defects detection and segmentation of ONH pathological changes	None
WinterFell	ONH area	ResNet-101, -152 (He et al., 2016), DensNet-169, -201 (Huang et al., 2017)	ORIGA	Ensemble of glaucoma likelihoods from multiple networks pre-trained on Image-Net and fine-tuned on ORIGA	Ensemble by mode, max. and min.

Table 4: Classification results of the participating teams in the REFUGE test set. The last row corresponds to the results obtained using the ground truth vertical cup-to-disc ratio (vCDR).

Rank	Team	AUC	Reference sensitivity
1	**VRT**	**0.9885**	0.9752
2	SDSAIRC	0.9817	**0.9760**
3	CUHKMED	0.9644	0.9500
4	NKSG	0.9587	0.8917
5	Mammoth	0.9555	0.8918
6	Masker	0.9524	0.8500
7	SMILEDeepDR	0.9508	0.8750
8	BUCT	0.9348	0.8500
9	WinterFell	0.9327	0.9250
10	NightOwl	0.9101	0.9000
11	Cvblab	0.8806	0.7318
12	AIML	0.8458	0.7250
Ground truth vCDR		0.9471	0.8750

and CUHKMED achieved also higher AUC values than the ground truth vCDR, although the differences were not statistically significant ($p > 0.05$). If the results of the best three teams are combined e.g. by normalizing their likelihoods and taking the average as a glaucoma score, the AUC is only marginally improved, with no significant differences with respect to the results of the best team ($p = 0.576$).

In order to understand the relevance of the classification results, a comparison with glaucoma experts was performed. To this end, two independent ophtalmologists visually graded the test set images and assigned a binary glaucomatous/non-glaucomatous label to each of them. These two glaucoma specialists were not part of the group of experts that provided the ground truth labels and did not take part of any discussion regarding data collection/preparation or the organization of the challenge. Notice that no clinical information but only the fundus image was used in this case to perform the annotation. This criteria was followed in order to ensure the same inputs to both the experts and the networks. The sensitivity and specificity values obtained by each human reader are included as expert operating points in Figure 5. The two points are close to each other due to a high level of agreement between the two experts (96.25% of the cases). The experts graded with the same sensitivity (85%) and slightly different specificity (91.11% and 91.39%) and accuracy (90.50% and 90.75%). If only the cases with their consensus are considered, then their

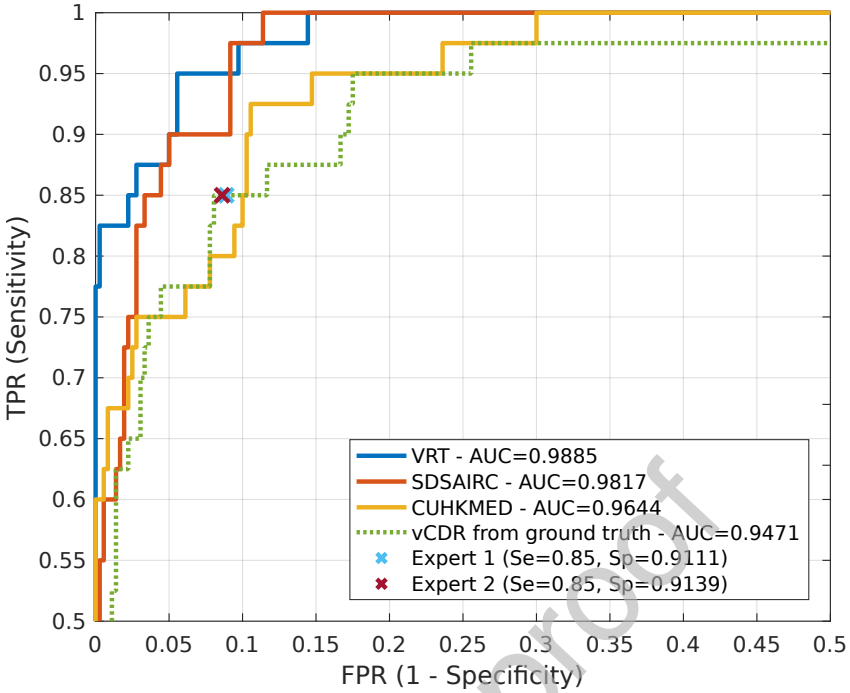


Figure 5: ROC curves and AUC values corresponding to the three top-ranked glaucoma classification methods (solid lines) and the vertical cup-to-disc ratio (vCDR) (green dotted line). Crosses indicate the operating points of two glaucoma experts.

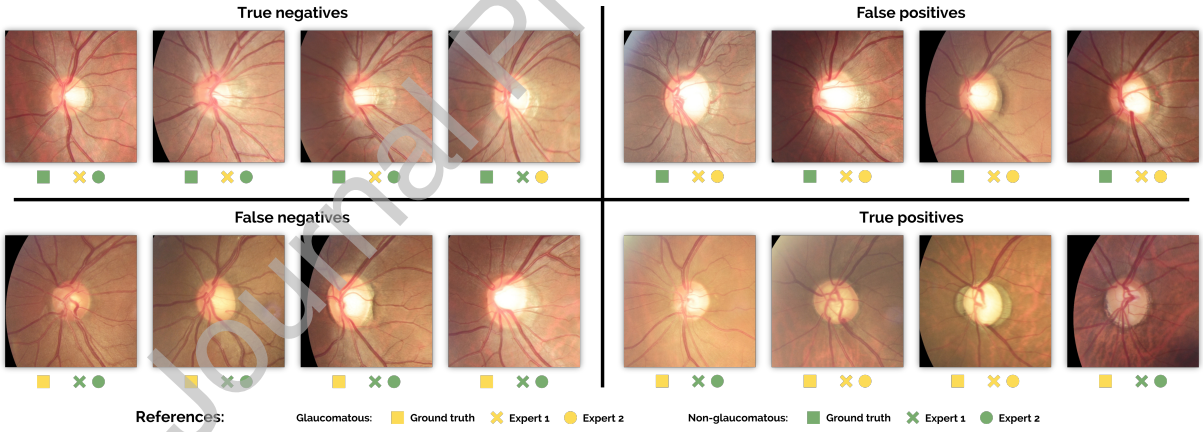


Figure 6: Qualitative results for glaucoma classification. Images are zoomed in the ONH area for better visualization. True positives (negatives) correspond to cases in which the ensemble of the three top-ranked methods reported a high (low) score. False positives (negatives) are images for which the ensemble returned a low (high) score. Ground truth and two experts' labels for glaucomatous (yellow) and non-glaucomatous (green) cases are included as colored squares, crosses and circles, respectively.

joint accuracy increases to 92.21%, while their joint sensitivity remains the same (85%) and the specificity reaches 93.04%. Despite the fact that both readers agreed with the vCDR curve in terms of sensitivity and specificity, this is pure coincidence as they did not take part of the OD/OC annotation procedure and did not have access to any segmentation.

Figure 6 illustrates a sample of true negatives, false positives, false negatives and true positive glaucoma detections from the REFUGE test set. The results correspond to the classification performed by the two additional experts and the average of the normalized glaucoma likelihoods of the three top-ranked teams. Since these values are not binary decisions but glaucoma scores, the false positive (negative) images were selected such that their assigned value was higher (lower) when the ground truth label was negative (positive). Similarly, the true positive (negative) images correspond to cases in which the joint likelihood is high (low).

4.2. Optic disc/cup segmentation

The evaluated methods for OD/OC segmentation are briefly described in Table 5. The interested reader could refer to the appendix for further details. The distribution of DSC and MAE values obtained by each of the participating teams in the REFUGE test set are represented as boxplots in Figure 7. Table 6 summarizes the final ranking, based on the average performance of each team. The statistical significance of the differences in performance of the top-ranked teams was assessed by means of Wilcoxon signed-rank tests ($\alpha = 0.05$). CUHKMED reported the highest DSC values for OD segmentation, significantly outperforming all the alternative approaches ($p < 1.41 \times 10^{-7}$). VRT and BUCT achieved the second and third higher average DSC values, respectively. However, their performance was not statistical significantly different with respect to each other ($p = 0.734$). For OC segmentation, Masker reported the highest average DSC value, followed by CUHKMED and BUCT. The differences in the DSC values achieved by Masker were statistically significant with respect to every other team ($p < 1 \times 10^{-4}$), except to CUHKMED ($p = 0.387$). When evaluating in terms of MAE of the vCDR estimation, Masker also reported the best results, consistently outperforming every other method ($p < 0.014$). CUHKMED retained the second place, although with no significant differences with respect to the BUCT ($p < 0.403$), which was ranked in the third place.

To study the complementarity of the three top-ranked methods according to the final leaderboard (CUHKMED, Masker and BUCT), a majority voting segmentation was obtained from their results, both for OD and OC. By quantitatively evaluating the resulting segmentations, and comparing to the constitutive models, we observed significant improvement in DSC values for OC (mean $\pm$ std $= 0.8922 \pm 0.0551$, Wilcoxon signed rank test, $p < 1.91 \times 10^{-7}$) and OD (mean $\pm$ std $= 0.9626 \pm 0.0196$, Wilcoxon signed rank test, $p < 1.07 \times 10^{-7}$). When the estimated vCDR values were analyzed in terms of MAE (mean$\pm$std $= 0.0398 \pm 0.0313$), the improvements were statistically significant compared to CUHKMED and BUCT ($p < 1.27 \times 10^{-4}$) but not to Masker ($p = 0.148$).

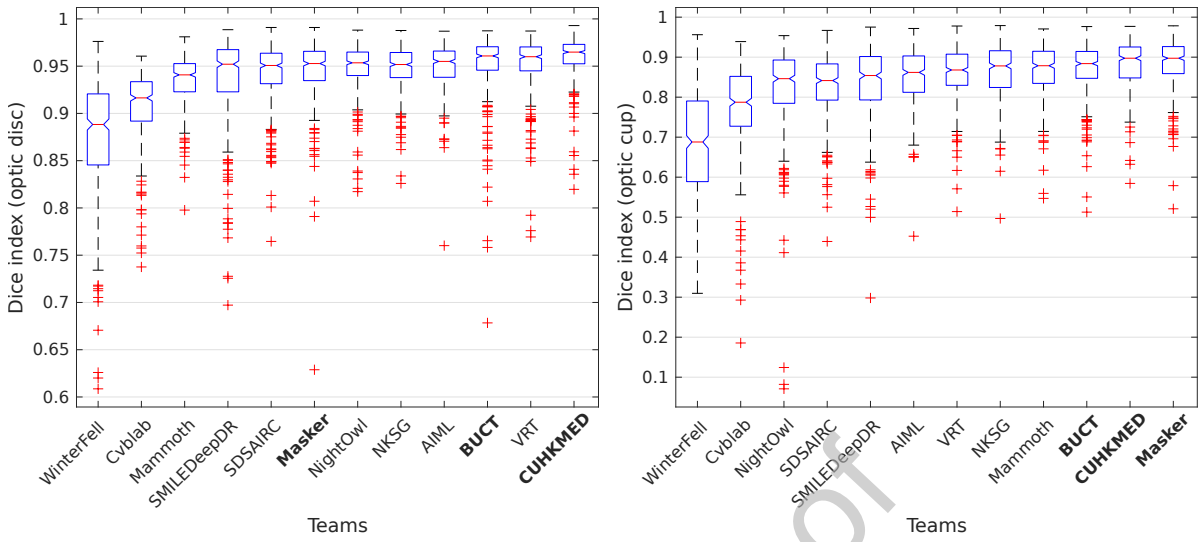
Figure 8 presents the distribution of DSC and MAE values stratified according to the glaucomatous/non-glaucomatous ground truth labels of the images. These metrics were calculated from the majority voting segmentations obtained from the three winning teams (CUHKMED, Masker and BUCT), although an analogous behavior was observed when stratifying the individual results of the methods. The statistical significance of the differences between groups was assessed using a Wilcoxon rank-sum test due to the

Table 5: Summary of the glaucoma classification methods evaluated in the on-site challenge, in alphabetical order using the teams names. FCN(s) stands for fully convolutional network(s).

Team	**Inputs**	**Architectures**	**Training set**	**Methodology**	**Post-processing**
AIML	Full image	FCNs ResNet-50, -101, -152 (He et al., 2016) and -38 (Wu et al., 2019)	REFUGE training set	Two stages: (i) Coarse ONH segmentation with ResNet-50, cropping, (ii) Fine-grain OD/OC segmentation with multi-view ensemble of networks	Ensemble by averaging
BUCT	Full image	U-Net (Ronneberger et al., 2015)	REFUGE training set	Two stages: (i) OD segmentation with a U-Net, postprocessing, cropping (ii) OC segmentation with U-Net and postprocessing	OD/OC: largest area element. OD: ellipse fitting.
CUHKMED	Full image	U-Net (Ronneberger et al., 2015) and DeepLabv3+ (Chen et al., 2018)	REFUGE training set and validation set (without labels)	U-Net used for cropping, DeepLabv3+ with geometry aware loss and domain shift adaptation via adversarial learning used for final segmentation	Ensemble by averaging
Cvblab	Full image with CLAHE	Modified U-Net (Sevastopolsky, 2017)	DRIONS-DB, DRISHTI-GS, RIM-ONE r3 and REFUGE training set	Two stages: (i) OD segmentation with a modified U-Net, cropping, (ii) OC segmentation with a modified U-Net from cropping	None
Mammoth	Full image	Mask-RCNN (He et al., 2017) and U-shaped dense network	Sample from REFUGE training set	Two stages: (i) OD segmentation with Mask-RNN and cropping, (ii) OC segmentation with dense U-Net. Resolution restored with spline interpolation	Ensemble of outputs, spline interpolation
Masker	Full image	Mask-RCNN (He et al., 2017)	REFUGE training set and ORIGA	Two stages: (i) Mask-RCNN to identify the ONH area, cropping, (ii) Ensemble by bootstrap voting of multiclass Mask-RCNN networks	Ensemble by voting
NightOwl	Full image	U-shaped dense network	REFUGE training set	Two stages: (i) C-Net for ONH detection, matching filter and cropping, (ii) OD/OC segmentation using two F-Nets	Opening and closing, Gaussian smoothing
NKSG	ONH area	DeepLabv3+ (Chen et al., 2018)	REFUGE training set	Multiclass segmentation using DeepLabv3+ on cropped images pre-processed with pixel quantization	None
SDSAIRC	Full image	M-Net (Fu et al., 2018)	REFUGE training set	Two stages: (i) OD segmentation with M-Net, cropping, (ii) OC segmentation with M-Net and postprocessing	Ellipse fitting
SmileDeepDR	Full image	U-shaped network with squeeze-and-excitation blocks (X-Unet)	REFUGE training set	X-Unet pre-trained for predicting ground truth labels, and fine-tuned separately for segmenting OD/OC using L1 regression loss	None
VRT	Full image	U-Net (Ronneberger et al., 2015) and vessel-based network (Son et al., 2017)	IDRiD and RIGA data sets	Two different U-Nets were applied for OD/OC segmentation, respectively. An auxiliary CNN using vessel segmentations as inputs was connected to the U-Nets to aid in the segmentation	Holes filling, convex-hull
WinterFell	Full image	Faster R-CNN (Girshick, 2015) and ResU-Net (Shankaranarayana et al., 2017)	ORIGA	Two stages: (i) ONH detection with Faster R-CNN, (ii) OD/OC segmentation in multiple color spaces with ResU-Net	None

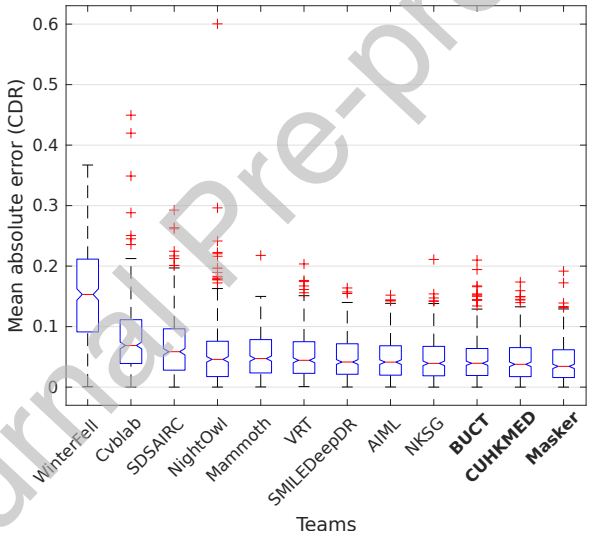
Table 6: Optic disc/cup segmentation results in the REFUGE test set. Average Dice (Avg. DSC) index for optic cup and disc and mean absolute error (MAE) of the vertical cup-to-disc ratio (vCDR). Teams are sorted by their final rank.

Rank	**Team**	**Score**	**Optic cup**		**Optic disc**		**vCDR**	
			Rank	**Avg. DSC**	**Rank**	**Avg. DSC**	**Rank**	**MAE**
1	**CUHKMED**	**1.75**	2	0.8826	**1**	**0.9602**	2	0.0450
2	Masker	2.5	**1**	**0.8837**	7	0.9464	**1**	**0.0414**
3	BUCT	3	3	0.8728	3	0.9525	3	0.0456
4	NKSG	4.6	5	0.8643	5	0.9488	4	0.0465
5	VRT	5.4	6	0.8600	2	0.9532	7	0.0525
6	AIML	5.45	7	0.8519	4	0.9505	5	0.0469
7	Mammoth	7.1	4	0.8667	10	0.9361	8	0.0526
8	SMILEDeepDR	7.45	4	0.8367	10	0.9386	8	0.0488
9	NightOwl	8.6	10	0.8257	6	0.9487	9	0.0563
10	SDSAIRC	9.15	9	0.8315	8	0.9436	10	0.0674
11	Cvblab	11	11	0.7728	11	0.9077	11	0.0798
12	WinterFell	12	12	0.6861	12	0.8772	12	0.1536



(a) Optic disc segmentation

(b) Optic cup segmentation



(c) MAE for vCDR

Figure 7: Box-plots illustrating the performance of each optic disc/cup segmentation method in the REFUGE test set. Distribution of Dice (DSC) values for (a) optic disc and (2) optic cup, and (c) mean absolute error (MAE) of the estimated vertical cup-to-disc-ratio (vCDR). The three top-ranked teams in the final leaderboard (CUHKMED, Masker and BUCT) are highlighted in bolds.

unpaired nature of the two sets (360 vs. 40 samples, respectively). For OD segmentation, the differences in performance between the two groups were not statistically significant ($p = 0.3435$). Higher values were obtained for OC segmentation in the glaucomatous group ($p = 2.09 \times 10^{-5}$), while the MAE values were significantly smaller in the positive set ($p = 0.023$).

Finally, Figure 9 presents some qualitative examples of the segmentations of the top-three ranked methods

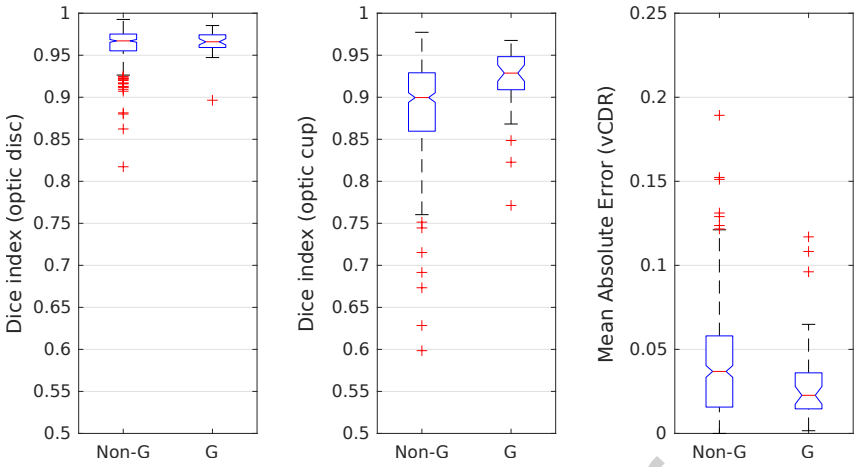


Figure 8: Segmentation metrics stratified for the glaucomatous (G) and non-glaucomatous (Non-G) subsets in the REFUGE test set. From left to right: Dice values for optic disc and optic cup segmentation, and mean absolute error of vertical cup-to-disc ratio (vCDR) estimates. The performance values were computed from segmentations obtained by majority voting of the top-three methods (CUHKMED, Masker and BUCT).

and those obtained by majority voting: (a) and (d) present some degree of peripapillary atrophy (PPA), (b) and (c) correspond to cases with ambiguous edges and (c) and (e) are the worst performing cases as measured in terms of DSC for the OD and the OC, respectively. The general behavior of each of the methods is rather stable compared with each other for most of the cases (Figure 9 (a), (d) and (e)). In challenging scenarios such as those observed in Figure 9 (b-e), where the edges of the ONH structures are difficult to assess, majority voting between methods resulted in more accurate segmentations. However, the voting only made a significant difference when the methods were complementary (Figure 9 (b) and (c) vs. (d) and (e)).

5. Discussion

The key methodological findings concluded after analyzing the challenge results are discussed in Section 5.1. Subsequently, Section 5.2 covers challenge strengths and limitations that might be taken into account in future editions. Finally, Section 5.3 covers the clinical implications of the results.

5.1. Findings

Our unified evaluation framework allowed us to draw some technical findings that might be useful for future developments in the field. Section 5.1.1 and Section 5.1.2 describe our findings in the classification and segmentation tasks, respectively. A special analysis of ensemble methods is provided in Section 5.1.3.

5.1.1. Classification methods

In line with the evolution of the literature in the field, we observed that the proposed solutions for glaucoma detection were generally based on state-of-the-art convolutional neural networks for image classi-

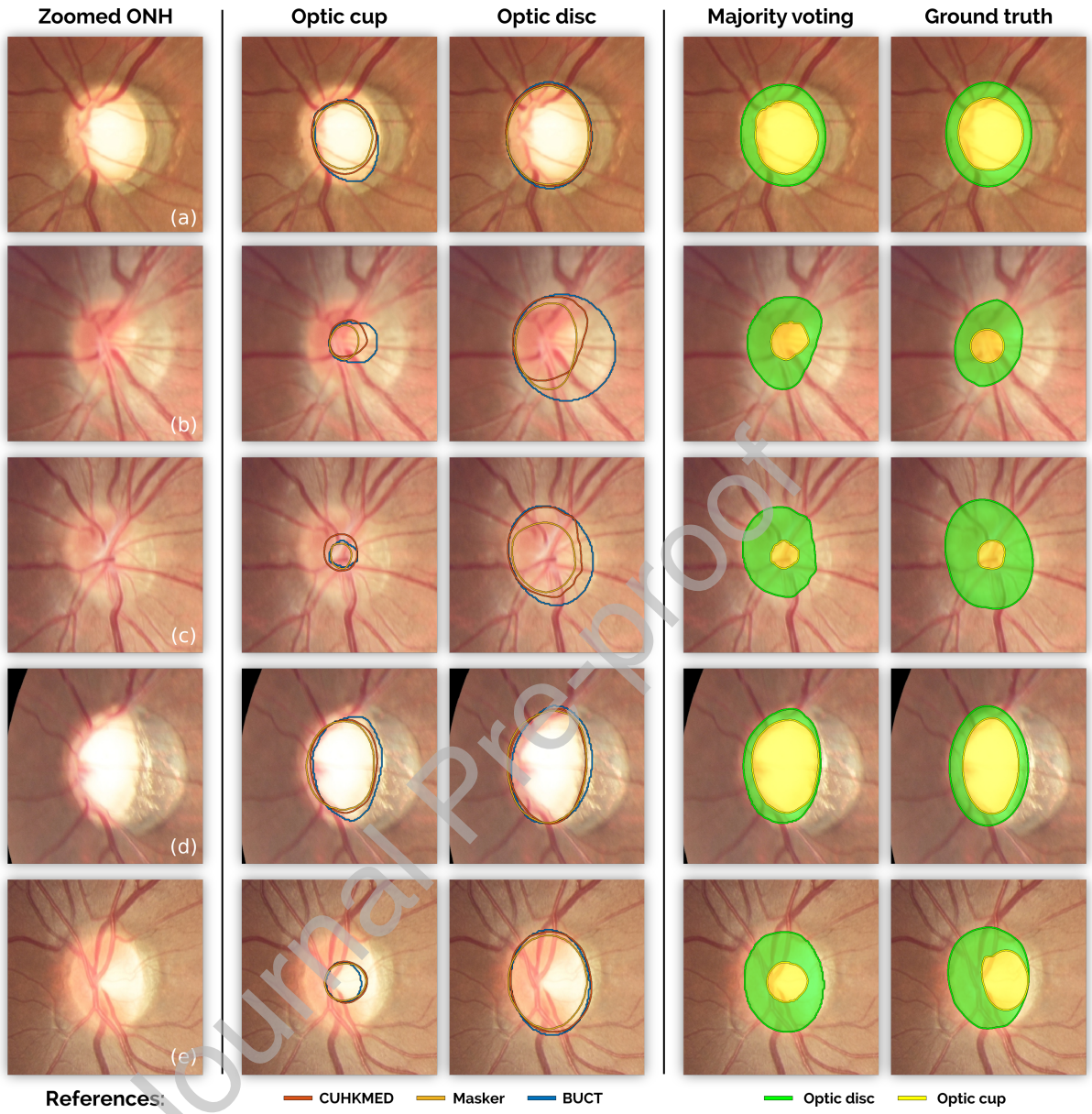


Figure 9: Optic disc/cup segmentation results in the REFUGE test set. From left to right: zoomed ONH area, segmentation results from the three top-ranked teams (BUCT, Masker and CUHKMED) for optic cup and disc segmentation, majority voting of these methods and ground truth segmentations.

fication, with the only exception of SMILEDeepDR and CUHKMED (Table 3). SMILEDeepDR adapted a segmentation network to predict both the OD/OC regions and a glaucoma likelihood, based on the intermediate feature representation generated by the architecture. CUHKMED, on the other hand, proposed to use a normalized vCDR predicted from the OD/OC segmentations.

The classification networks comprised of general-purpose image classification models that were top-ranked

in ImageNet Large Scale Visual Recognition Competition (Russakovsky et al., 2015), such as VGG19 (Simonyan and Zisserman, 2014), ResNets (He et al., 2016), DenseNets (Huang et al., 2017), Inception V3 (Szegedy et al., 2016) or Xception (Chollet, 2017), among others. Since training such deep architectures from scratch on a training set with only 400 images might be prone to overfitting, most of the teams initialized the CNNs with pre-trained weights from ImageNet and fine-tuned them afterwards using the CFPs. Alternatively, NKSG team used pre-trained weights from the Kaggle DR data set (Kaggle, 2015). This eases the fine-tuning task as the transition from natural images to fundus photographs is less smooth than the one from images of DR to glaucoma. Only BUCT trained its networks from scratch, although using the ONH area and not the full images. Nevertheless, we observed that the best solutions were based not only on the application of an existing classification network but also using domain-specific heuristics as discussed next.

CUHKMED achieved the third place by relying only on the prediction of the vCDR. Deep learning was in this case used indirectly, as it was applied for segmenting the OD/OC areas. Exploiting a well-known, clinical parameter such as the vCDR allowed them to identify most of the cases with cupping, which usually correspond to advanced glaucomatous damage. SDSAIRC (second place), on the other hand, obtained better results by combining vCDR estimates with glaucoma likelihoods provided by different CNNs. Team Masker (sixth place) followed a similar idea but using a network trained on full images. Instead, SDSAIRC trained the CNNs using a cropped version of each image in which the ONH is observed at the upper-left corner. We hypothesize that this configuration indirectly forces the network to identify other complementary signs that are not captured by the vCDR, such as the presence of peripapillary hemorrhages–which appear in the border of the OD (Figure 2 (b))–or RNFL defects–observed as striated patterns spanning from the ONH (Figure 2 (c)). Similarly, the winning team, VRT, further improved this idea by using an attention-guided network (Son et al., 2018). This approach takes as input both a fundus image and a region mask covering the optic disc and the RNFL area. By means of a structural region separation model (Park et al., 2018), the network is driven to analyze regions in which disease-specific biomarkers may occur. In principle, a classification network with enough capacity would learn to identify abnormal image patterns by itself, without needing an attention mask, although this is highly dependent on the size of the training set (Poplin et al., 2018). VRT team instead restricted the field-of-view of the method by focusing on disease-relevant areas. This attention mechanism might help to learn more accurate classification models that does not require manual annotations of glaucoma-related abnormalities such as RNFL defects or peripapillary hemorrhages. On the other hand, VRT increased REFUGE training set by incorporating images from other public data sets, assigning to them image-level classification labels using a pre-trained model. Using additional public data with weak labels was accepted by the organizers as the resulting increased data set has annotations that are still prone to errors. Hence, it was possible to evaluate the contribution of a weak training signal to the proposed approach. The results of VRT seems to empirically show that increasing the training set

with further scans is beneficial even if the training labels were obtained automatically.

5.1.2. Segmentation methods

The proposed solutions for OD/OC segmentation were all based on at least one fully convolutional neural network (Table 5). U-shaped networks inspired by the U-Net (Ronneberger et al., 2015) were the prevalent solutions, although incorporating recent technologies such as residual connections (AIML), atrous convolutions (BUCT) or multiscale feeding inputs (SDSAIRC), among others. Most of the strategies were also based on the two stage approach described in Section 2 of first roughly identifying the ONH and then performing the OD/OC segmentation on a cropped version of the original image. The three top-ranked teams followed this principle. CUHKMED and BUCT used a classical U-Net (Ronneberger et al., 2015) to localize the ONH area, while Masker applied a Mask-RCNN (He et al., 2017). Once this area was localized, CUHKMED segmented the OD/OC using a DeepLabv3+ (Chen et al., 2018) architecture, a recently published approach based on atrous separable convolutions that is able to capture multiscale characteristics. Masker, on the other hand, used an ensemble of Mask-RNNs trained with bootstrap, while BUCT used a classical U-Net. NKSG was ranked fourth using the same architecture as CUHKMED, but normalizing image appeareance between training and offline test sets using a pixel quantization technique. CUHKMED, on the other hand, accounted for this domain shift using adversarial learning, which could explain its better performance.

Interestingly, we noticed that the three top-ranked methods and their ensemble by majority voting achieved consistently better segmentation results in the subset of glaucomatous subjects than in the non-glaucomatous cases. This can be linked with the fact that advanced glaucoma cases with severe cupping usually present more clear interfaces between the OD and the OC. Such a characteristic would explain why the improvement is more evident in the Dice index obtained for the OC than in the performance for OD segmentation. On the other hand, the segmentation models showed a slightly worst accuracy in challenging scenarios with unclear transitions between the OD/OC, such as those illustrated in Figure 9 (b), (c) and (e). The lack of depth information in monocular color fundus photographs turns this task significantly difficult. Research in developing automated methods for predicting depth maps from CFPs is currently ungoing, trying to correlate image features with ground truth labels obtained from other imaging modalities such as stereo fundus photography (Shankaranarayana et al., 2019) or OCT (Thurtell et al., 2018). These techniques might aid to solve ambiguities in these scenarios.

If the segmentation results are analyzed separately, BUCT and CUHKMED achieved the second and the third place for OC segmentation and the first and third places for OD segmentation, respectively (Table 6). Using the same criteria, Masker achieved the first place for OC segmentation but the seventh for OD segmentation. Surprinsingly, the team reported the lowest MAE of the vCDR estimation. This indicates that most of their errors in the OD prediction occurs horizontally, and therefore not affect the prediction of

its vertical diameter.

5.1.3. Ensemble methods

Independently of the target task, we noticed that several submissions exploited to some extent the application of ensembles. Combining the outcomes of multiple models is a common practice in challenges as it allows to achieve (sometimes marginal) quantitative improvements that can eventually ensure higher positions in the final rankings (Kaggle, 2015; Kamnitsas et al., 2017). We observed three types of ensembles in REFUGE. Teams AIML, Cvblab and WinterFell, for instance, combined the outputs of multiple architectures trained with the same data set. This approach allows to take advantage of the characteristics of each model without explicitly integrating them into a single network. Alternatively, team Mammoth averaged the outputs of a single architecture trained under different configurations (e.g. images preprocessed with multiple strategies). Under this setting, model selection is bypassed as there is no need to choose a single configuration because a subset or even all of them are exploited in test time. Finally, a similar approach was followed by NightOwl and Masker for classification and segmentation, respectively, although by training the same architecture on different portions of the training data.

Applying majority voting or averaging on the collective responses of multiple models might ensure more reliable results. This has been recently applied by De Fauw et al. (2018) for retinal OCT analysis with a significant success. However, this will strictly depend on the complementarity (and non-redundance) of the ensembled methods. We experimentally assessed how complementary the top-winning methods are by averaging their normalized likelihoods (for the glaucoma classification task) and taking segmentations by majority voting (for the OD/OC segmentation task). In both tasks we have observed increments in performance that in principle indicate that each winning approach is complementary with the others. This was more notorious for the second task (Figure 7), where the segmentations obtained by majority voting of the top-ranked methods were more accurate when the models disagreed the most. This indicates that, despite their impressive but similar performance, the methods are still complementary with each other, and can be integrated to generate a more accurate automated response. This can be qualitatively observed in the segmentation examples in Figure 9, where e.g. BUCT oversegmented the OD and the OC in (b) but achieved more accurate results in (c). On the other hand, cases such as those in Figure 9 (d) and (e) illustrate the need of model diversity to achieve more accurate results under challenging conditions. The improvements in the classification task were only marginal when averaging the top-three models (AUC = 0.9901) and not significant ($p = 0.576$). This is most likely a consequence of the high agreement between the models, indicating that there are still cases that are missclassified. In any case, notice, however, that we cannot argue that the ensemble of these particular approaches is *per-se* the best way to go for performing the individual tasks. To ensure a proper generalization error and avoid any selection bias, an ensemble approach must be based on models that are choosen according to their individual performance on a held-out

validation set.

5.2. Challenge strengths and limitations

REFUGE was the first open initiative aiming to introduce a uniform evaluation framework to assess automated methods for OD/OC segmentation and glaucoma classification from CFPs. To this end, the challenge provided to the community with the largest public available data set of fundus photographs (1200 scans) to date. In addition, it contains gold standard clinical diagnostic labels, and a high quality reference OD/OC masks and fovea positions from a total of nine glaucoma experts. This unique characteristic ensures a more appropriate development of glaucoma classification methods, as it was recently observed that training with fundus-derived labels have a negative impact on performance to detect truly diseased cases (Phene et al., 2018). To the best of our knowledge, the most similar data set to REFUGE was ORIGA (Zhang et al., 2010), which provided 650 images with OD/OC segmentations and glaucoma labels. However, at the time of submitting this manuscript ORIGA was not available anymore[12], while, more than 350 teams have successfully registered to the REFUGE website to access the database, with 183 requests submitted after the on-site challenge. Such a large interest of the scientific community in accessing REFUGE data clearly demonstrates that a quality open glaucoma data set and challenge was needed.

The challenge design matched most of the principles for evaluating retinal image analysis algorithms proposed by Trucco et al. (2013). In particular, REFUGE data set can be easily accessed through a website that is part of the Grand Challenges organization. Furthermore, an automated tool is provided to evaluate the results of any participating team, ensuring a uniform, un-biased criterion for comparing methods, based on trustable and accurate annotations. Furthermore, the data is already partitioned into fixed training, offline and online test sets, with labels publicly available only for the first two sets. Future participants are invited to submit their results to the website to estimate their performance on the test set. By keeping these ground truth annotations private we prevent the teams to overfit on test data, ensuring a fair comparison between models.

As the offline test set set was used to determine which teams were qualified to participate in the on-site challenge, the access to the validation labels was initially restricted. Only five submissions per team were allowed to evaluate the performance on the offline test set, limiting its applicability for design tasks such as model selection. As a consequence of this constrain, most of the teams ended up using the REFUGE training set or other third-party data sets for this purpose, which might have affected their performance. To overcome this issue, we have publicly released the offline test set labels right after the onsite event. We encourage future participants to use this data as a validation set, not only for model selection but also to better explain their models' behavior e.g. through ablation studies, to empirically show the contribution

[12] `http://imed.nimte.ac.cn/origa-650.html`

of each decision in intermediate results. This might help to better identify good practices to follow when designing glaucoma classification and OD/OC segmentation methods.

Another remark regarding the data set organization is that the winners of the challenge were selected according to a weighted sum of their rankings in the offline and the onsite test sets (Eq. 5). This was intentionally done to reward the participants for their efforts in having good results in the offline test set, while preventing dummy submissions with the sole purpose of participating in the onsite event. This last point was also guaranteed by inviting the 12 best performing teams on the offsite test set to participate in the onsite challenge. Each team was allowed to request a maximum number of 5 evaluations in the offline test set, to avoid strong overfitting on it. Nevertheless, and despite the fact that a low weight was assigned to this rank, the final score might be biased due to some form of overfitting on the offline test set. This paper is focused only on the results of the onsite test set, though, which was held out during the entire challenge and for which only a single submission is allowed. Our conclusions remain therefore unbiased by this issue. Future challenges might perhaps consider the posibility of using four splits instead of three: two with public labels for training and model selection/validation, and other two with private labels for offline and onsite evaluations. Hence, if only one submission is allowed for the last two sets, then further conclusions regarding the generalization ability of the methods could be drawn.

One limitation of REFUGE is the lack of diverse ethnicities in its data set, as the images correspond to a Chinese population. Ethnicities manifest differently in CFP due to changes in the pigment of the fundus. Therefore, it cannot be ensured that the best performing models on the REFUGE challenge can be applied to a different population and obtain the same outcomes without retraining. Furthermore, it is worth mentioning that the percentage of glaucoma cases in the REFUGE data set is higher than expected to be encountered in a screening scenario and more representative of a clinical one. Furthermore, despite the fact that REFUGE data set is the largest publicly available image source for glaucoma classification, 1200 images is still not big enough for developing general enough deep learning solutions. Similar initiatives in other diseases have provided larger data sets: the Kaggle challenge in diabetic retinopathy grading, for instance, released more than 80.000 CFPs for training and testing the algorithms (Kaggle, 2015). The high quality of the images also hampers the applicability of the proposed methods in real screening scenarios, where imaging artifacts and low quality scans are expected to appear much more frequently. A representative screening data set should include comorbidities, diverse ethnicities, ages and genders and low quality images with acquisition artifacts. These characteristics should be addressed in future challenges e.g. by multicenter collaboration for data collection, to ensure that the winning models can be applied in a more general environment.

Another potential limitation of REFUGE data sets is that the OD/OC manual segmentations were performed from CFPs, without considering depth information or knowledge about the Bruchs membrane opening. The latter is considered by most as the best OD anatomical delimitation, and serves as reference for one of the most recent measures of the amount of retinal nerve fibers, the Bruchs membrane opening

minimum rim width (BMO-MRW) (Reis et al., 2012). As a consequence, these annotations might be deviated from the real anatomy of both areas. In an effort to alleviate this drawback, our annotations resulted from majority voting of delineations performed first by seven different glaucoma specialists, and then controlled by another independent expert. We believe that these steps ensured much more reliable outcomes than using annotations from a single reader, although further validation would be certainly needed to confirm this hypothesis. Better ground truth labels could be obtained e.g. by delineating OD/OC from OCT scans, which provide cross-sectional images of the retina (and therefore depth information). However, the resulting labels should be afterwards transferred to CFP e.g. via image registration, which might be subject to errors if the registration algorithm fails. In that case, manual correction based on CFP would be still required, and deviation from the true geometry might then still occur.

REFUGE data was prepared with glaucoma status as the main target label. After applying the pre-defined protocol to analyze the follow-up medical records of each CFP, the images were annonimized and it is unfeasible now to link them with their clinical information. As a consequence, additional labels for other co-existing morbities in the non-glaucomatous and glaucomatous sets were lost. Similar initiatives might take this into consideration in the future, and provide not only the target labels for the specific applications of the challenge but also complementary information such as labels for other conditions or functional parameters such as the IOP. This would not only allow further assessment of the challenge results (e.g. the influence of comorbidities or some parameters in the final outcomes) but also indirectly benefit other derived applications (e.g. automated myopia or megalopapilae detection or computerized prediction of IOP from CFP).

Finally, it is important to remark a point regarding the evaluated proposals and their differences in training settings, particularly those related with data availability. Despite the fact that ORIGA is claimed to be publicly available since 2010 (Zhang et al., 2010), by the time of this publication it was not possible to download the images. These kind of fluctuations in data access might have influenced the decisions made by the participating teams about which image sources to use for training their models. Future challenges might address this issue e.g. by providing a curated list of potential sources to retrieve images. In any case, it is worth mentioning that only one of the top-ranked teams (Masker, who achieved the second place in the onsite evaluation of the segmentation task) used ORIGA to train its model, so in principle the access to this data was not per se a guaranty of success.

5.3. Clinical implications of the results and future work

Can we envision automated systems for detecting suspicious cases of glaucoma from fundus photographs? This is still an open question, although REFUGE results might help us to catch a glimpse of a possible answer. With the constant development of much cheaper and easy-to-use fundus cameras, it is expected that this imaging technique will be widespread even more in the decades to follow. Turning it into a cost-effective

imaging modality for glaucoma screening is still pending due to the subtle manifestation of the early stages of the disease in these images. Nevertheless, novel image analysis techniques based on deep learning can pave the way towards computer-aided screening of glaucoma from fundus photographs.

We observed that some of the proposed segmentation models were able to obtain accurate vertical cup-to-disc ratio estimates. The best team in the segmentation task (CUHKMED) achieved the third place in the classification ranking by using the vCDR as a glaucoma likelihood, with sensitivity and specificity values almost in pair with two human experts, and statistically equivalent to those obtained using the ground truth measurements. The best performing teams, however, complemented ONH measurements with the classification outcomes of deep learning based models, and were able to significantly surpass the glaucoma experts, with increments in sensitivity up to a 10%. Although these results are limited to a specific image population, we can still argue that these deep learning models are able to identify complementary features, invisible to the naked eye, that are essential to ensure a more accurate diagnosis of the disease. Representing the activation areas on the images might help to better understand which areas were considered by the automated models to produce their predictions. We believe that these tools might contribute in the future to a better identification of glaucoma suspects based on color fundus images alone.

The challenge results also seem to indicate that vCDR, although being an important risk factor for glaucoma, is not enough for detecting the disease at a single time-point basis. This is likely as a consequence of vCDR ignoring other important features such as ONH hemorrhages or RNFL defects. Other metrics derived from the OD/OC relative shapes were recently observed to outperform vCDR for screening, such as the rim to disc ratio (Kumar et al., 2019). Notice also that some clinical guidelines such as European Glaucoma Society (2017) do not recommend vCDR to classify patients, as several healthy discs might have large vCDR. Attention is instead recommended towards the neuroretinal rim thickness and the degree of vCDR symmetry between eyes (European Glaucoma Society, 2017). In any case, vCDR is still a relevant parameter (it achieved an AUC of 0.9471 in our test set for glaucoma classification) that might help to analyze disease progression (e.g. in a follow-up study in which the evolution of the vCDR is assessed for each visit of the patient). Glaucoma screening tools should certainly not ignore vCDR but should also take other biomarkers into account such as the presence, size and location of ONH hemorrhages or the presence and size of RNFL defects, to ensure more reliable predictions.

The complementarity of CFP and OCT for automated glaucoma screening still needs to be exploited. Although CFP allows a cost-effective assessment of the retina, features such as the damage in the ONH or the RNFL are more evident in optic disc centered OCT. This is due to the fact that OCT provides a three dimensional view of the retina, with a micrometric resolution. Hence, the cross-sectional scans–or B-scans–can be used to quantify the thickness of the RNFL or the degree of cupping in the ONH. Nevertheless, the OCT acquisition devices are more expensive than fundus cameras, and the manual analysis of the volumetric information is costly and time-consuming. Developing deep learning methods to quantify

glaucoma biomarkers from OCT scans is therefore necessary to complement results in fundus images and pave the way towards cost-effective glaucoma screening.

6. Conclusions

We summarized the results and findings from REFUGE, the first open challenge focused on glaucoma classification and optic disc/cup segmentation from color fundus photographs. We analyzed the performance of each of the twelve teams that participated in the on-site edition of the competition, during MICCAI 2018. We observed that the best approaches for glaucoma classification integrated deep learning techniques with well-known glaucoma specific biomarkers such as changes in the vertical cup-to-disc ratio or retinal nerve fiber layer defects. The two top-ranked teams, on the other hand, achieved better results than two glaucoma specialists, a promising sign towards using automated methods to identify glaucoma suspects with fundus imaging. For the segmentation task, the best solutions took into account the domain shift between training and test sets, aiming to regularize the models to deal with image variability. Cases with ambiguous edges between the optic disc and the optic cup showed to be the most challenging ones. Further research should be performed to improve the results in those scenarios. For both tasks of the challenge, we observed that integrating the outcomes of multiple models allowed to improve their individual performance.

REFUGE unified evaluation framework allowed us to identify good common practices based on the results of the twelve proposed approaches. We expect these findings to help in the future to develop strong baselines for comparison and to aid in the design of new automated tools for image-based glaucoma assessment.

REFUGE challenge data and evaluation framework are publicly accessible through the Grand Challenges website at `https://refuge.grand-challenge.org/`. In parallel, a sibling platform has been deployed at `http://eye.baidu.com/` with capabilities to automatically process teams' submissions. Future participants are invited to submit their results in any of these websites. Participation requests have to include all the requested information (full real name, institution and e-mail) to be approved, or will be otherwise declined. The two websites will remain permanently available for submissions, to encourage future developments in the field.

Acknowledgements

This work was supported by the Christian Doppler Research Association, the Austrian Federal Ministry for Digital and Economic Affairs and the National Foundation for Research, Technology and Development, J.I.O is supported by WWTF (Medical University of Vienna: AugUniWien/FA7464A0249, University of Vienna: VRG12-009). Team Masker is supported by Natural Science Foundation of Guangdong Province of China (Grant 2017A030310647). Team BUCT is partially supported by the National Natural Science

Foundation of China (Grant 11571031). The authors would also like to thank REFUGE study group for collaborating with this challenge.

Conflicts of Interest

None.

References

Abràmoff, M.D., Garvin, M.K., Sonka, M., 2010. Retinal imaging and image analysis. IEEE Rev. Biomed. Eng. 3, 169–208.

Abràmoff, M.D., Lavin, P.T., Birch, M., Shah, N., Folk, J.C., 2018. Pivotal trial of an autonomous ai-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit. Med. 1, 39.

Al-Bander, B., Williams, B.M., Al-Nuaimy, W., Al-Taee, M.A., Pratt, H., Zheng, Y., 2018. Dense fully convolutional segmentation of the optic disc and cup in colour fundus for glaucoma diagnosis. Symmetry 10, 87.

Almazroa, A., Alodhayb, S., Osman, E., Ramadan, E., Hummadi, M., Dlaim, M., Alkatee, M., Raahemifar, K., Lakshminarayanan, V., 2018. Retinal fundus images for glaucoma analysis: the RIGA dataset, in: Med. Imaging 2018: Imaging Inform. for Healthc., Res., and Appl., SPIE. p. 105790B.

Almazroa, A., Burman, R., Raahemifar, K., Lakshminarayanan, V., 2015. Optic disc and optic cup segmentation methodologies for glaucoma image detection: a survey. J. Ophthalmol. 2015.

Berman, D., Avidan, S., et al., 2016. Non-local image dehazing, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 1674–1682.

Burlina, P.M., Joshi, N., Pekala, M., Pacheco, K.D., Freund, D.E., Bressler, N.M., 2017. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. JAMA Ophthalmol. 135, 1170–1176.

Carmona, E.J., Rincón, M., García-Feijoó, J., Martínez-de-la Casa, J.M., 2008. Identification of the optic nerve head with genetic algorithms. Artif. Intell. Med. 43, 243–259.

Cerentinia, A., Welfera, D., d'Ornellasa, M.C., Haygertb, C.J.P., Dottob, G.N., 2018. Automatic identification of glaucoma using deep learning methods, in: Proc. World Congress on Med. and Health Informatics, IOS Press. p. 318.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.

Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 .

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, in: Comput. Vis. ECCV, pp. 801–818.

Chen, X., Xu, Y., Wong, D.W.K., Wong, T.Y., Liu, J., 2015a. Glaucoma detection based on deep convolutional neural network, in: Conf. Proc. IEEE Eng. Med. Biol. Soc., IEEE. pp. 715–718.

Chen, X., Xu, Y., Yan, S., Wong, D.W.K., Wong, T.Y., Liu, J., 2015b. Automatic feature learning for glaucoma detection based on deep learning, in: Med. Image Comput. Comput. Assist. Interv.. Springer, pp. 669–677.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. arXiv preprint , 1610–02357.

Christopher, M., Belghith, A., Bowd, C., Proudfoot, J.A., Goldbaum, M.H., Weinreb, R.N., Girkin, C.A., Liebmann, J.M., Zangwill, L.M., 2018. Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. Sci. Rep. 8, 16685.

Davis, J., Goadrich, M., 2006. The relationship between Precision-Recall and ROC curves, in: Proc. Int. Conference on Mach. Learn., ACM. pp. 233–240.

De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., ODonoghue, B., Visentin, D., et al., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. 24, 1342–1350.

Decencire, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., Charton, B., Klein, J.C., 2014. Feedback on a publicly distributed database: the Messidor database. Image Anal. & Stereol. 33, 231–234. doi:10.5566/ias.1155.

DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837–845.

Edupuganti, V.G., Chawla, A., Kale, A., 2018. Automatic optic disk and cup segmentation of fundus images using deep learning, in: Conf. Proc. IEEE Int. Image Processing (ICIP), IEEE. pp. 2227–2231.

European Glaucoma Society, 2017. European glaucoma society terminology and guidelines for glaucoma, 4th edition - part 1 supported by the egs foundation. Br. J. Ophthalmol. 101, 1–72. doi:10.1136/bjophthalmol-2016-EGSguideline.001.

Farbman, Z., Fattal, R., Lischinski, D., Szeliski, R., 2008. Edge-preserving decompositions for multi-scale tone and detail manipulation. ACM Trans. Graph. 27, 67.

Fu, H., Cheng, J., Xu, Y., Wong, D.W.K., Liu, J., Cao, X., 2018. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE Trans. Med. Imaging 37, 1597–1605.

Fumero, F., Alayón, S., Sanchez, J.L., Sigut, J., Gonzalez-Hernandez, M., 2011. RIM-ONE: An open retinal image database for optic nerve evaluation, in: Proc. Int. Symposium on Comp.-based Med. Syst. (CBMS), IEEE. pp. 1–6.

Girshick, R., 2015. Fast r-cnn, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 1440–1448.

Gómez-Valverde, J.J., Antón, A., Fatti, G., Liefers, B., Herranz, A., Santos, A., Sánchez, C.I., Ledesma-Carbayo, M.J., 2019. Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning. Biomed. Opt. Express 10, 892–913.

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316, 2402–2410.

Hagiwara, Y., Koh, J.E.W., Tan, J.H., Bhandary, S.V., Laude, A., Ciaccio, E.J., Tong, L., Acharya, U.R., 2018. Computer-aided diagnosis of glaucoma using fundus images: A review. Comput. Methods Programs Biomed. 165, 1–12.

Haleem, M.S., Han, L., van Hemert, J., Li, B., 2013. Automatic extraction of retinal features from colour retinal images for glaucoma diagnosis: A review. Comput. Med. Imaging Graph. 37, 581–596.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: Proc. IEEE Int. Conf. Comput. Vis., IEEE. pp. 2980–2988.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 770–778.

Holm, S., Russell, G., Nourrit, V., McLoughlin, N., 2017. DR HAGIS: a fundus image database for the automatic extraction of retinal surface vessels from diabetic patients. J. Med. Imaging 4, 014503.

Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E., 2018. Squeeze-and-Excitation Networks, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 7132–7141.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 4700–4708.

Joshi, G.D., Sivaswamy, J., Krishnadas, S., 2011. Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment. IEEE Trans. Med. Imaging 30, 1192–1205.

Kaggle, 2015. Diabetic Retinopathy Detection. https://www.kaggle.com/c/diabetic-retinopathy-detection. [Online; ac-

cessed 10-January-2019].

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., et al., 2017. Ensembles of multiple models and architectures for robust brain tumour segmentation. arXiv preprint arXiv:1711.01468 .

Kumar, J.H., Seelamantula, C.S., Kamath, Y.S., Jampala, R., 2019. Rim-to-disc ratio outperforms cup-to-disc ratio for glaucoma prescreening. Sci. Rep. 9, 7099.

Lavinsky, F., Wollstein, G., Tauber, J., Schuman, J.S., 2017. The future of imaging in detecting glaucoma progression. Ophthalmology 124, S76–S82.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86, 2278–2324.

Li, A., Wang, Y., Cheng, J., Liu, J., 2018a. Combining multiple deep features for glaucoma classification, in: Proc. Int. Conf. on Acoust., Speech and Signal Processing (ICASSP), IEEE. pp. 985–989.

Li, Z., He, Y., Keel, S., Meng, W., Chang, R.T., He, M., 2018b. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. Ophthalmology 125, 1199–1206.

Lim, G., Cheng, Y., Hsu, W., Lee, M.L., 2015. Integrated optic disc and cup segmentation with deep learning, in: Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on, IEEE. pp. 162–169.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context, in: Comput. Vis. ECCV, Springer. pp. 740–755.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Liu, S., Graham, S.L., Schulz, A., Kalloniatis, M., Zangerl, B., Cai, W., Gao, Y., Chua, B., Arvind, H., Grigg, J., et al., 2018. A deep learning-based algorithm identifies glaucomatous discs using monoscopic fundus photographs. Ophthalmology Glaucoma 1, 15–22.

Lowell, J., Hunter, A., Steel, D., Basu, A., Ryder, R., Fletcher, E., Kennedy, L., et al., 2004. Optic nerve head segmentation. IEEE Trans. Med. Imaging 23, 256–264.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., et al., 2018. Why rankings of biomedical image analysis competitions should be interpreted with care. Nat. Commun. 9, 5217.

Miri, M.S., Abràmoff, M.D., Lee, K., Niemeijer, M., Wang, J.K., Kwon, Y.H., Garvin, M.K., 2015. Multimodal segmentation of optic disc and cup from SD-OCT and color fundus photographs using a machine-learning graph-based approach. IEEE Trans. Med. Imaging 34, 1854–1866.

Niemeijer, M., Van Ginneken, B., Cree, M.J., Mizutani, A., Quellec, G., Sánchez, C.I., Zhang, B., Hornero, R., Lamard, M., Muramatsu, C., et al., 2010. Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. IEEE Trans. Med. Imaging 29, 185–195.

Odstrčilík, J., Kolar, R., Budai, A., Hornegger, J., Jan, J., Gazarek, J., Kubena, T., Cernosek, P., Svoboda, O., Angelopoulou, E., 2013. Retinal vessel segmentation by improved matched filtering: evaluation on a new high-resolution fundus image database. IET Image Processing 7, 373–383.

Orlando, J.I., Barbosa Breda, J., van Keer, K., Blaschko, M.B., Blanco, P.J., Bulant, C.A., 2018. Towards a glaucoma risk index based on simulated hemodynamics from fundus images, in: Med. Image Comput. Comput. Assist. Interv.. Springer. volume 11071, pp. 65–73.

Orlando, J.I., Prokofyeva, E., Blaschko, M.B., 2017a. A discriminatively trained fully connected conditional random field model for blood vessel segmentation in fundus images. IEEE. Trans. Biomed. Eng. 64, 16–27.

Orlando, J.I., Prokofyeva, E., del Fresno, M., Blaschko, M.B., 2017b. Convolutional neural network transfer for automated

glaucoma identification, in: Proc. SPIE, pp. 101600U–101600U–10.

Pal, A., Moorthy, M.R., Shahina, A., 2018. G-eyenet: A convolutional autoencoding classifier framework for the detection of glaucoma from retinal fundus images, in: Conf. Proc. IEEE Int. Image Processing (ICIP), IEEE. pp. 2775–2779.

Park, S.J., Shin, J.Y., Kim, S., Son, J., Jung, K.H., Park, K.H., 2018. A novel fundus image reading tool for efficient generation of a multi-dimensional categorical image database for machine learning algorithm training. J. Korean Med. Sci. 33, e239.

Phene, S., Dunn, R.C., Hammel, N., Liu, Y., Krause, J., Kitade, N., Schaekermann, M., Sayres, R., Wu, D.J., Bora, A., et al., 2018. Deep learning to assess glaucoma risk and associated features in fundus images. arXiv preprint arXiv:1812.08911 .

Poplin, R., Varadarajan, A.V., Blumer, K., Liu, Y., McConnell, M.V., Corrado, G.S., Peng, L., Webster, D.R., 2018. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat. Biomed. Eng. 2, 158164.

Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Meriaudeau, F., 2018. Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research. Data 3, 25.

Prevedello, L.M., Halabi, S.S., Shih, G., Wu, C.C., Kohli, M.D., Chokshi, F.H., Erickson, B.J., Kalpathy-Cramer, J., Andriole, K.P., Flanders, A.E., 2019. Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions. Radiology: Artificial Intelligence 1, e180031.

Prokofyeva, E., Zrenner, E., 2012. Epidemiology of major eye diseases leading to blindness in Europe: A literature review. Ophthalmic Res. 47, 171–188.

Raghavendra, U., Fujita, H., Bhandary, S.V., Gudigar, A., Tan, J.H., Acharya, U.R., 2018. Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images. Inf. Sci. 441, 41–49.

Reinke, A., Eisenmann, M., Onogur, S., Stankovic, M., Scholz, P., Full, P.M., Bogunovic, H., Landman, B.A., Maier, O., Menze, B., et al., 2018. How to exploit weaknesses in biomedical challenge design and organization, in: Med. Image Comput. Comput. Assist. Interv., Springer. pp. 388–395.

Reis, A.S., Sharpe, G.P., Yang, H., Nicolela, M.T., Burgoyne, C.F., Chauhan, B.C., 2012. Optic disc margin anatomy in patients with glaucoma and normal controls with spectral domain optical coherence tomography. Ophthalmology 119, 738–747.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Med. Image Comput. Comput. Assist. Interv., Springer. pp. 234–241.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 4510–4520.

Schacknow, P.N., Samples, J.R., 2010. The glaucoma book: a practical, evidence-based approach to patient care. Springer Science & Business Media, New York.

Schmidt-Erfurth, U., Sadeghipour, A., Gerendas, B.S., Waldstein, S.M., Bogunović, H., 2018. Artificial Intelligence in Retina. Prog. Retin. and Eye Res. 67, 1–29.

Sevastopolsky, A., 2017. Optic disc and cup segmentation methods for glaucoma detection with modification of U-Net convolutional neural network. Pattern Recognit. and Image Anal. 27, 618–624.

Sevastopolsky, A., Drapak, S., Kiselev, K., Snyder, B.M., Georgievskaya, A., 2018. Stack-U-Net: Refinement network for image segmentation on the example of optic disc and cup. arXiv preprint arXiv:1804.11294 .

Shankaranarayana, S.M., Ram, K., Mitra, K., Sivaprakasam, M., 2017. Joint optic disc and cup segmentation using fully convolutional and adversarial networks, in: Fetal, Infant and Ophthalmic Med. Image Anal.. Springer, pp. 168–176.

Shankaranarayana, S.M., Ram, K., Mitra, K., Sivaprakasam, M., 2019. Fully convolutional networks for monocular retinal depth estimation and optic disc-cup segmentation. arXiv preprint arXiv:1902.01040 .

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

Sivaswamy, J., Krishnadas, S., Chakravarty, A., Joshi, G., Tabish, A.S., et al., 2015. A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. JSM Biomed. Imaging Data Papers 2, 1004.

Sivaswamy, J., Krishnadas, S., Joshi, G.D., Jain, M., Tabish, A.U.S., 2014. Drishti-gs: Retinal image dataset for optic nerve head (ONH) segmentation, in: Proc. IEEE Int. Symp. Biomed. Imaging, IEEE. pp. 53–56.

Son, J., Bae, W., Kim, S., Park, S.J., Jung, K.H., 2018. Classification of Findings with Localized Lesions in Fundoscopic Images Using a Regionally Guided CNN, in: Comput. Pathol. and Ophthalmic Med. Image Anal.. Springer, pp. 176–184.

Son, J., Park, S.J., Jung, K.H., 2017. Retinal vessel segmentation in fundoscopic images with generative adversarial networks. arXiv preprint arXiv:1706.09318 .

Sun, X., Xu, Y., Tan, M., Fu, H., Zhao, W., You, T., Liu, J., 2018. Localizing optic disc and cup for glaucoma screening via deep object detection networks, in: Comput. Pathol. and Ophthalmic Med. Image Anal.. Springer, pp. 236–244.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 2818–2826.

Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med. imaging 15, 15–29.

Thakur, N., Juneja, M., 2018. Survey on segmentation and classification approaches of optic cup and optic disc for diagnosis of glaucoma. Biomed. Signal Process Control 42, 162–189.

Tham, Y.C., Li, X., Wong, T.Y., Quigley, H.A., Aung, T., Cheng, C.Y., 2014. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. Ophthalmology 121, 2081–2090.

Thurtell, M.J., Kardon, R.H., Garvin, M.K., 2018. Local estimation of the degree of optic disc swelling from color fundus photography. Comput. Pathol. and Ophthalmic Med. Image Anal. 11039, 277.

Trucco, E., Ruggeri, A., Karnowski, T., Giancardo, L., Chaum, E., Hubschman, J.P., Al-Diri, B., Cheung, C.Y., Wong, D., Abramoff, M., et al., 2013. Validating retinal fundus image analysis algorithms: Issues and a proposal. Invest. Ophthalmol. Vis. Sc. 54, 3546–3559.

Vergara, I.A., Norambuena, T., Ferrada, E., Slater, A.W., Melo, F., 2008. StAR: a simple tool for the statistical comparison of ROC curves. BMC Bioinformatics 9, 265.

Wang, S., Yu, L., Yang, X., Fu, C.W., Heng, P.A., 2019. Patch-based output space adversarial learning for joint optic disc and cup segmentation. IEEE Trans. Med. Imaging, In press.

Wang, S., Zhang, L., 2017. CatGAN: coupled adversarial transfer for domain generation. arXiv preprint arXiv:1711.08904 .

Wu, Z., Shen, C., Van Den Hengel, A., 2019. Wider or deeper: Revisiting the ResNet model for visual recognition. Pattern Recognit. 90, 119–133.

Xu, Y., Duan, L., Lin, S., Chen, X., Wong, D.W.K., Wong, T.Y., Liu, J., 2014. Optic cup segmentation for glaucoma detection using low-rank superpixel representation, in: Med. Image Comput. Comput. Assist. Interv., Springer. pp. 788–795.

Zhang, Z., Yin, F.S., Liu, J., Wong, W.K., Tan, N.M., Lee, B.H., Cheng, J., Wong, T.Y., 2010. Origa-light: An online retinal fundus image database for glaucoma analysis and research, in: Conf. Proc. IEEE Eng. Med. Biol. Soc., IEEE. pp. 3065–3068.

Zheng, Y., Hijazi, M.H.A., Coenen, F., 2012. Automated disease/no disease grading of age-related macular degeneration by an image mining approach. Invest. Ophthalmol. Vis. Sc. 53, 8310–8318.

Zilly, J.G., Buhmann, J.M., Mahapatra, D., 2015. Boosting convolutional filters with entropy sampling for optic cup and disc image segmentation from fundus images, in: Int. Worksh. on Mach. Learn. in Med. Imaging, Springer. pp. 136–143.

Appendix A. Participating methods

AIML

OD/OC segmentation. The method was a two-stage approach based on a combination of multiple dilated fully-convolutional networks (FCNs) based on ResNet-50, -101, -152 (He et al., 2016) and -38 (Wu et al., 2019). First, a ResNet-50 FCN was used to coarsely segment the ONH. The corresponding region was afterwards cropped to cover approximately one quarter of the original resolution. These images were used to fed the ResNet-50, -101, -152 (He et al., 2016) and -38 (Wu et al., 2019) models, which produced the final segmentations of the OD/OC. The networks were trained using the REFUGE training set with data augmentation, including rescalings and rotations. The final prediction was obtained by averaging multi-view results produced by all the networks on different augmented versions of each image.

Glaucoma classification. Two sets of classification models were combined. One was trained using the whole fundus images, while the other was trained using only local regions around the ONH. The OD/OC area was detected using the segmentation model described above. Subsequently, the REFUGE training set was used to fine-tune pre-trained ResNet-50, -101, -152 (He et al., 2016) and -38 (Wu et al., 2019) models. The final classification result was assigned by ensembling the outputs of these architectures by averaging.

BUCT

OD/OC segmentation. The OD/OC were segmented separately by two different U-Net (Ronneberger et al., 2015) models. First, the images on the REFUGE training set were resized to fit the resolution of those on the validation set and converted to gray scale. Then, for OD segmentation, a square of 817×817 pixels was cropped from the input images, leaving the ONH on the left-hand side, and then resized to 256×256 pixels. A U-Net with less convolutional filters than the original approach (Ronneberger et al., 2015) was applied to retrieve the OD. To remove false positives, the largest connected component was taken, and an ellipse was fitted to the OD segmentation. For OC segmentation, the smallest rectangle containing the OD was clipped out, and each side of the rectangle was extended with 100 pixels to fit a resolution of 128×128 pixels. The same U-Net architecture was retrained then on these images and applied to retrieve the OC. The largest connected component was taken as the final result, too. In both cases, the U-Nets were trained using the REFUGE training set with data augmentation, including rotations and flippings.

Glaucoma classification. The same cropping strategy applied for OD/OC segmentation was used for this task. The resulting CFPs were then transformed into grayscale images. Standard data augmentation techniques such as rotations and shiftings were applied to increase the size of the training set. Then, an

X-ception (Chollet, 2017) network was trained from scratch for glaucoma classification using grayscale versions of the color images on the REFUGE training set and the ground truth annotations.

CUHKMED

OD/OC segmentation. A patch-based Output Space Adversarial Learning framework (pOSAL) (Wang et al., 2019) was introduced for this task. This method enables output space domain adaptation to reduce the segmentation performance degradation on target datasets with domain shift in an unsupervised way. A region of interest (ROI) containing the OD from each original image was first extracted using a U-Net (Ronneberger et al., 2015) model. The DeepLabv3+ (Chen et al., 2018) architecture was afterwards applied for segmentation, using the backbone of MobileNetV2 (Sandler et al., 2018). Considering the shape of the OD and OC, a morphology-aware segmentation loss was designed to force the network to generate smooth predictions. To overcome the domain shift between training and testing datasets, adversarial learning was exploited, encouraging the segmentation predictions in the target domain to be similar to the source ones. During this process, the labelled training images are considered as the source domain, while the unlabelled validation images are from the target domain. Specifically, a patch-based discriminator was introduced to distinguish whether the prediction came from the source or the target domain and the adversarial learning prompts the segmentation network to generate validation predictions similar to predictions of training images (Wang et al., 2019). The final image prediction was acquired by ensembling five models, to further improve the segmentation performance. Further details are provided in (Wang et al., 2019).

Glaucoma classification. This task was tackled without using a dedicated method. Instead, the authors proposed to use the OD/OC segmentation masks–automatically obtained with the method described above–to compute the vertical CDR (vCDR). To this end, two ellipses were fitted to the the OD and OC masks, respectively. The vCDR values were normalized into 0-1 as a final classification probability following: $p_{\text{new}} = \frac{p - p_{\min}}{p_{\max} - p_{\min}}$, where p is the calculated vCDR values, $p_{\min}$ and $p_{\max}$ are the minimum and maximum vCDR values among all the testing images.

Cvblab

OD/OC segmentation. A two-stage process was followed for this task, based on a modified U-Net architecture (Sevastopolsky, 2017). The OD was segmented first and the resulting mask was used to crop the image and segmenting the OC. As a pre-processing technique, the Contrast Limited Adaptive Histogram Equalization (CLAHE) method, was applied. The images were also resized to 256×256 pixels before feeding the network. The models were trained using DRIONS-DB, DRISHTI-GS, RIM-ONE v3 and the REFUGE training set.

Glaucoma classification. An ensemble of VGG19 (Simonyan and Zisserman, 2014), GoogLeNet (InceptionV3) (Szegedy et al., 2016), ResNet-50 (He et al., 2016) and the Xception (Chollet, 2017) architectures was applied for this task. Each network was independently fine-tuned from the weights pre-trained from ImageNet (Russakovsky et al., 2015) to identify glaucomatous images, using DRISHTI-GS1, HRF, ORIGA, RIM-ONE and the training set of the REFUGE databases. Data augmentation was applied in the form of vertical and horizontal flippings, rotations up to 50°, height/width shifts of 0.15 and zooms in a range between 0.7 and 1.3. Prior to fine tunning, the training data was balanced using SMOTE (Chawla et al., 2002) on the REFUGE training set, with the aim of reducing the bias on the prediction model towards the more common class (Normal). All the images were resized to 256×256 pixels before feeding the network. The results were merged together by ensembling the models' outputs taking the average glaucoma likelihood.

Mammoth

OD/OC segmentation. A Mask-RCNN (He et al., 2017) and a Dense U-Net (Ronneberger et al., 2015) were ensembled for this task. For Mask-RCNN, the OD was first segmented. Then, each input image was cropped around its center to retrieve a patch with a size of 512×512 pixels, and the segmentation of the OC was performed on it. For the Dense U-Net, which is a modified U-Net architecture with dense convolutional blocks and dilated convolutions, the OD was first segmented. Then the probability mask was used as additional channel of the input (as attention) to segment OC. Both networks were trained using a linear combination of cross-entropy and Dice losses. The probability outputs of both networks were averaged to generate the final segmentation results. A subsample from the original REFUGE training set was used to learn the models. In particular, it was divided into two new sets, one used for training (32 glaucoma images and 288 non-glaucoma images) and a second for validation (8 glaucoma images and 72 non-glaucoma images). The Mask-RCNN internally used a ResNet-50 (He et al., 2016) model pre-trained in the COCO (Lin et al., 2014) data set and fine-tuned using the above mentioned training set.

Glaucoma classification. The OD/OC segmentation method was used to crop each input image and generate a patch centered in the ONH, covering 1.5 times the radius of the OD. The resulting image was then resized to 448×448, and CLAHE contrast equalization and mean color normalization were subsequently applied to uniform image characteristics across data sets. A combination of a ResNet-18 (He et al., 2016) (supervised) and a CatGAN (Wang and Zhang, 2017) (semi-supervised) classification networks was applied for diagnosis. The CatGAN was used to aid the learning process of the ResNet-18 model in a semi-supervised setting, using fake images generated by the CatGAN to increase the size of the training set. The same training/validation partition used for OD/OC segmentation was applied for this task. A series of ResNet-18 models was trained using 4-fold cross-validation on these training set and a weighted and an

unweighted cross-entropy loss, resulting in $4 \times 2 = 8$ models in total. At inference time, the predictions of all the models were averaged into a final glaucoma likelihood.

Masker

OD/OC segmentation. The first step consisted of localizing the ONH region. A Mask-RCNN (He et al., 2017) architecture was used to this end. Afterwards, the image was cropped around the ONH to build a new training set. This set was divided into 14 partitions based on a bagging principle. Different image preprocessing techniques were applied to each subset, namely image dehazing (Berman et al., 2016) and edge-preserving multiscale image decomposition based on weighted least squares optimization (Farbman et al., 2008). Different networks including Mask-RCNN (He et al., 2017), U-Net (Ronneberger et al., 2015) and M-Net (Fu et al., 2018) were trained on each subset, and the final result was obtained by a voting procedure in which regions predicted by 80% of all the networks were taken as the final segmentation.

Glaucoma classification. The vCDR value was first calculated using the segmentation results obtained with the previously described method. Subsequently, several classification networks based on ResNet (He et al., 2016) were trained from scratch to predict the risk of glaucoma. The REFUGE training set and ORIGA were used to learn the models. The final result was obtained based on a linear combination of the vCDR values and the prediction of the classification networks. We use ResNet-50, ResNet-101 and ResNet-152 as the basic classification models. The final glaucoma risk is:

$$\text{Glaucoma Risk} = 0.8 \times \text{CDR} + 0.2 \times \text{CNets}. \tag{A.1}$$

Here, CDR is the vertical cup to disc ratio and CNets is the final voting of the ensemble classification networks. If 80% of all the networks predict a image with high risk of glaucoma, CNets = 1, otherwise, CNets = 0. In our implementation, we use 14 different networks.

NightOwl

OD/OC segmentation. A coarse to fine approach was proposed for this task, based on two dense U-shaped networks with dense blocks (Huang et al., 2017), namely CoarseNet (C-Net) and FineNet (F-Net), respectively. The C-Net model was used to coarsely localize the ONH region. Then, the F-Net was applied to retrieve the final segmentation of the OD and the OC. A modified version of pooling based on the mean of average and max-pooling was applied for better feature accumulation. The images were preprocessed using histogram matching to normalize the intensities in the sample space–and exponential transformations to enhance the boundaries of the optic cup—. Standard data augmentation techniques were applied to the

REFUGE training set to balance the number of images from each class (glaucomatous / non-glaucomatous). The original inputs, resized to 112×112 pixels, were fed to the C-Net for localizing the ONH region. This area was then extracted from the original input image, resized to 112×112 pixels too, and fed to two different F-Nets for OD/OC segmentation. Outliers were removed using morphological operations (opening and closing) and Gaussian smoothing.

Glaucoma classification. The encoders of each F-Net were used for extracting two vectors of 2048 features each, one for the OD and one for the OC. Dimensionality reduction via convolutions was applied to retrieve two new vectors with 64 features each. The concatenation of these two vectors was used to feed a neural network with 4 fully connected layers, trained to predict the glaucoma likelihood. The weights of the F-Net encoders were not adjusted for glaucoma classification, only the weights used for dimensionality reduction and those of the fully connected layers. 10-fold cross-validation was applied to retrieve 10 different models, and 7 of them were retrieved based on their confusion matrices. The final glaucoma likelihood was obtained by taking the maximum likelihood from all the models.

NKSG

OD/OC segmentation. The DeepLabv3+ (Chen et al., 2018) architecture was used for this task, based on the assumption that atrous spatial pyramid pooling (ASPP) is effective to segment objects at multiple scales. The network was trained using cross-entropy as the loss function. The images were pre-processed using pixel quantization to reduce the sensitivity of the model to changes in color and to improve its robustness. Moreover, the segmentation approach was applied on cropped versions of the input images. These were obtained by extracting a bounding box surrounding the ONH area.

Glaucoma classification. This task was performed using a SENet (Hu et al., 2018) architecture. This network has large capacity, as it has 154 layers in total. Instead of using fully connected layers, it uses 1×1 convolutions. The images were preprocessed by applying the same strategy used for segmentation. The glaucomatous/non-glaucomatous classes were balanced using re-sampling. By means of data augmentation using rotations and stretching, the REFUGE training set was increased to a total of 2000 images.

SDSAIRC

OD/OC segmentation. A method inspired by the M-Net (Fu et al., 2018) was applied for this task. An area of 480×480 pixels size was defined and prepared as the segmentation ROI for each image, centered on the OD and transformed to polar coordinates afterwards. The histogram of the test images were matched to the average histogram of the REFUGE training set to compensate image variance per camera vendor.

The segmentation task was divided into OD segmentation from the segmentation ROI and OC segmentation from the bounding box of the OD. This box was tightly cropped to contain the entire OD. This two stage separation helped to tackle the difficulty in finding the ideal weights for the M-Net (Fu et al., 2018). The segmentation accuracy was further improved by post-processing the resulting masks using ellipse fitting.

Glaucoma classification. A ResNet-50 (He et al., 2016) network with pre-trained weights from ImageNet (Russakovsky et al., 2015) was fine-tuned on the REFUGE training for glaucoma classification. Histogram matching was applied to uniform the appeareance of images with respect to the training set. The CFPs were also cropped in such a way that the OD was positioned in the upper-left corner. This setting allows to capture RNFL defects in more detail than cropping a square centered in the ONH. The final glaucoma likelihood was obtained by averaging the classification score predicted by the network with the resulting score of a logistic regression which takes advantage of vCDR value, estimated from the OD/OC segmentation, as an input. To this end, the logistic regression classifier was trained separately using the transformed vCDR value.

SMILEDeepDR

OD/OC segmentation. A modified U-Net (Ronneberger et al., 2015) architecture, namely X-Unet, was applied for this task. It used 3 inputs so that it was able to receive more original raw pixel information during training. This strategy was used to reduce the risk of overfitting while enhancing the network's learning capability. Moreover, squeeze-and-excitation blocks were embedded into this U-Net variant to weight the features from different convolutional layers' channels. Such a mechanism was able to selectively amplify the valuable channel-wise features and suppress the useless feature from global information. In addition, deconvolution were used in the network decoder to refine the decoding capability by refusing the features between different level encoded features and the corresponding level decoded features. The segmentation task was also posed as a linear regression task instead of a typical pixel classification problem, using L_1 loss for training. A *split-copy-merge* strategy was followed: a X-Unet network was trained first to predict the ground labels. Secondly, two X-Unets were separately fine-tuned using the learned weights, only to predict the OD and the OC, respectively. Then, the predictions of both networks were merged to get the final result.

Glaucoma classification. The Deeplabv3+ (Chen et al., 2017) was modified and used as a classifier. Its last layer was replaced by a global average pooling layer followed by a fully connected layer. The model was trained on the REFUGE training set using the cross-entropy loss. Instead of using the full images, a pre-processing stage based on cropping the regions around the ONH was followed.

VRT

OD/OC segmentation. A U-Net (Ronneberger et al., 2015) based architecture was used, complemented by an auxiliary CNN (Son et al., 2017) that took a vessel segmentation mask and generated a coarse mask with the estimated OD/OC location. The output of the second network was concatenated to the bottleneck layer of the U-Net to generate the final segmentation mask. A combined loss $L_{\text{total}} = L_{\text{main}} + \lambda * L_{\text{vessel}}$ was applied, where L_{main} and L_{vessel} are pixel-wise binary cross entropy for the U-Net and the auxiliary CNN. The values for λ, the depth of U-Net and the number of filters at the last layer of the auxiliary CNN were experimentally selected using a hill-climbing approach. The OD and the OC were segmented separately using two different U-Net architectures. Holes in the final segmentations were filled, and the OD/OC areas were converted to convex-hulls to ensure a single binary mask per regions.

Glaucoma classification. The method was based on three architectures as described in (Son et al., 2018),[13] each of them targetting glaucoma classification or the detection of glaucomatous disc changes and RNFL defects. The three models were trained using images from three public data sets, namely Kaggle (Kaggle, 2015), MESSIDOR (Decencire et al., 2014) and IDRiD (Porwal et al., 2018). Since these databases do not have labels for any of these tasks, a semi-supervised learning approach was followed. Models pre-trained on a private data set were used to assign labels to the images on each of the public sets. Given that the data sets used are still public and the assigned labels are not gold standard annotations but automated and therefore prone to errors, the organizers decided that this proposal is still in accordance with the participation rules. The same architectures used for assigning the automated labels were then trained from scratch on the combined data set to produce final predictions. The final glaucoma likelihood was assigned by doing: max{glaucomatous disc change, RNFL defect + glaucoma suspect/2}.

WinterFell

OD/OC segmentation. The ONH was initially detected using a Faster R-CNN (Girshick, 2015). This area was cropped in all the images, and two image processing techniques were applied on the outputs. The first approach consisted of selecting a standard image and then normalize the remaining ones using it as a reference. The second image version was the inverted green channel of the original RGB cropped image. Finally, a ResU-Net (Shankaranarayana et al., 2017) model was applied on the resulting images for OD/OC segmentation.

Glaucoma classification. An ensemble of ResNets (He et al., 2016) (101 and 152) and DensNets (Huang et al., 2017) (169 and 201) was used for classification. The networks were pre-trained on ImageNet and

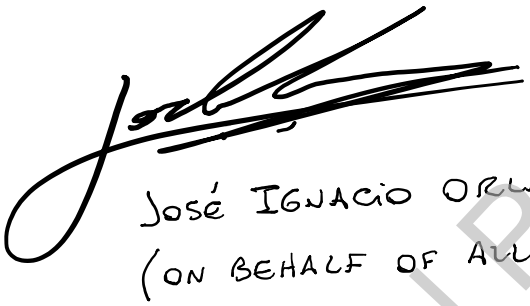
[13] `https://bitbucket.org/woalsdnd/refuge/src`

separately fine-tuned using ORIGA, based on the log-likelihood loss. Each model was trained on cropped versions of the inputs images, centered in the ONH and on three different color spaces (RGB, HSV and the inverted green channel). Hence, $4 \times 3 = 12$ different models were produced. The final result was obtained by taking the mode of the binary decisions of each network. If the predicted label was glaucoma, the maximum confidence score was used as a final likelihood. On the contrary, if the image was labeled as non-glaucomatous, then the minimum score was applied.

1177 **Declaration of interests**

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:



JOSÉ IGNACIO ORLANDO
(ON BEHALF OF ALL THE AUTHORS)