Acta Ophthalmologica

Accurate prediction of glaucoma from colour fundus images with a convolutional neural network that relies on active and transfer learning

Ruben Hemelings,[1,2] Bart Elen,[2] João Barbosa-Breda,[1] Sophie Lemmens,[1] Maarten Meire,[3]
Sayeh Pourjavan,[4] Evelien Vandewalle,[1,5] Sara Van de Veire,[6] Matthew B. Blaschko,[7]
Patrick De Boever[2,8] and Ingeborg Stalmans[1,5]

[1]Research Group Ophthalmology, KU Leuven, Leuven, Belgium
[2]VITO NV, Mol, Belgium
[3]TC CS-ADVISE, KU Leuven, Geel, Belgium
[4]Chirec Hospitals, Brussel, Belgium
[5]Ophthalmology Department, UZ Leuven, Leuven, Belgium
[6]AZ Sint-Jan, Brugge, Belgium
[7]ESAT-PSI, KU Leuven, Leuven, Belgium
[8]Hasselt University, Diepenbeek, Belgium

ABSTRACT.
Purpose: **To assess the use of deep learning (DL) for computer-assisted glaucoma identification, and the impact of training using images selected by an active learning strategy, which minimizes labelling cost. Additionally, this study focuses on the explainability of the glaucoma classifier.**

Methods: **This original investigation pooled 8433 retrospectively collected and anonymized colour optic disc-centred fundus images, in order to develop a deep learning-based classifier for glaucoma diagnosis. The labels of the various deep learning models were compared with the clinical assessment by glaucoma experts. Data were analysed between March and October 2018. Sensitivity, specificity, area under the receiver operating characteristic curve (AUC), and amount of data used for discriminating between glaucomatous and non-glaucomatous fundus images, on both image and patient level.**

Results: **Trained using 2072 colour fundus images, representing 42% of the original training data, the trained DL model achieved an AUC of 0.995, sensitivity and specificity of, respectively, 98.0% (CI 95.5%–99.4%) and 91% (CI 84.0%–96.0%), for glaucoma versus non-glaucoma patient referral.**

Conclusions: **These results demonstrate the benefits of deep learning for automated glaucoma detection based on optic disc-centred fundus images. The combined use of transfer and active learning in the medical community can optimize performance of DL models, while minimizing the labelling cost of domain-specific mavens. Glaucoma experts are able to make use of heat maps generated by the deep learning classifier to assess its decision, which seems to be related to inferior and superior neuroretinal rim (within ONH), and RNFL in superotemporal and inferotemporal zones (outside ONH).**

Key words: artificial intelligence – deep learning – fundus image – glaucoma detection

Acta Ophthalmol. 2020: 98: e94–e100

doi: 10.1111/aos.14193

Introduction

Glaucoma is currently responsible for approximately 12% of all cases of irreversible vision loss (Kapetanakis et al. 2006). The number of patients is expected to increase in our ageing society. Predictions indicate that over 110 million people worldwide may be diagnosed with the disease by 2040 (Tham et al. 2014). Glaucoma is a neurodegenerative disease characterized by retinal ganglion cell loss as a result of multiple factors, including high intraocular pressure, optic nerve ocular blood flow dysregulation and neurotoxicity. Progressive optic nerve fibre damage leads to visual field (VF) loss, which often remains unnoticed by the patient because the initial VF loss is peripheral and is compensated by the overlapping VF of the contralateral eye as well as by a compensatory ‘filling-in’ of these zones by the brain. The resulting lack of early symptoms implies that a significant number of individuals remain undiagnosed, even in high-income countries. Besides VF testing, structural assessment of the optic nerve head (ONH) and retinal nerve fibre layer (RNFL) is crucial in the diagnosis and follow-up of glaucoma. Optical coherence tomography (OCT) and fundus photography are two complementary imaging modalities, with the latter allowing qualitative analysis like disc haemorrhages and colour changes. General population screening for glaucoma is currently not common practice (Ervin et al. 2012), as there is no sufficient evidence of its cost-effectiveness to date (Tuulonen 2011; Burr et al. 2014). With the prospect of a growing population affected by glaucoma, a thorough reassessment of glaucoma care is warranted.

Ophthalmology is pioneering with future possible application of artificial intelligence (AI) (Ting et al. 2019). Gulshan et al. (2016) developed a convolutional neural network (CNN) for the detection of diabetic retinopathy (DR) from fundus images, scoring areas under the receiver operating characteristic curves (AUCs) of 0.991 and 0.990 on two validation sets. More recently, van der Heijden et al. (2018) reported an AUC of 0.94 on a referral task for DR in a prospective study with nearly 900 patients. This pivotal study led to FDA clearance of the first commercial automated grading tool for referable DR using deep learning. Automated detection of age-related macular degeneration from colour fundus images using a pretrained deep learning encoder on the large public AREDS data set was independently described by Burlina et al. (2018) and Grassmann et al. (2018).

Automated glaucoma detection from fundus imaging has been actively studied prior to deep learning, with the majority of techniques relying on handcrafted features, such as the vertical cup-to-disc ratio, extracted from fundus images. Deep learning architectures for glaucoma have been reported on topics including optic disc and cup segmentation (Fu et al. 2018), VF prediction (Wen et al. 2018), and automated glaucoma detection using small data sets (Matsopoulos et al. 2008; Asaoka et al. 2016; Maheshwari et al. 2017; Muhammad et al. 2017; Ahn et al. 2018; Shibata et al. 2018). In 2015, the first results on glaucoma classification with deep learning were published, using two data sets (<2000 images) (Chen et al. (2015). More recently, Li et al. (2018a,b) described automated glaucoma detection using 48 116 fundus images from an Asian population, reporting high sensitivity (95.6%), specificity (92.0%) and AUC (0.986) on a validation set of more than 8000 images using pretrained deep learning encoders. The main strength of their work is the recruitment of a large number of trained ophthalmologists, who graded the entire set of fundus images for signs of glaucoma.

The current paper reports on the development of a glaucoma prediction model. Optic disc changes are initially subtle and can be challenging to detect by a human grader. Our access to glaucomatous fundus images – labelled based on a complete ophthalmologic examination (tonometry, OCT or confocal scanning laser ophthalmoscopy) – allows the deep learning encoder to learn subtle features in fundus images of early/moderate stage glaucoma patients. Hence, the first objective of this study was to develop and validate a deep learned glaucoma classifier using colour fundus images from a patient population, measured against clinical diagnosis.

The second objective was to explore the added-value of active learning (Settles 2009) on top of deep learning for automated glaucoma detection. Active learning is a special case of semi-supervised learning that aims to leverage uncertainty information from an unlabelled set in order to predict from which unlabelled images the classifier would benefit the most if they would become labelled. True labels, especially in the medical community, can be difficult to obtain. By employing an active learning system that maximizes classification performance, while minimizing the number of required labels, data sets and labelling efforts can be used more efficiently.

The third and final objective was to inspect the trained model's decision process using interpretable heat maps. Deep learning (DL) models learn concepts from the data itself, omitting the need for manual feature extraction, and leading to state-of-the-art results, but lower transparency in understanding the classifier's decision process. Heat maps that visualize the image areas that contributed the most towards glaucoma classification might assist in opening the black box of the trained deep learning system.

Methods

Image and label acquisition

All 30° optic disc-centred colour fundus images of 1620 × 1444 resolution were captured with a Zeiss VISUCAM (Carl Zeiss Meditec, Jena, Germany) and used retrospectively in the current study. The glaucomatous fundus images (6651) originate from 1353 unique patients (±4.9 images per patient) imaged at the glaucoma clinic of the University Hospitals Leuven (Belgium) during several consultations between 2009 and 2017. Over 60% of patients went to follow-up consultations, leading to images taken at different points in time, which can differ due disease progression, hence useful for the model. The vast majority of fundus images (1614) from 403 non-glaucoma (normal) individuals (±4 images per individual) stem from a data set collected at three different locations in the context of an awareness campaign during the World Glaucoma Week 2018 that took place between 11th and 17th March. Screening sessions were organized at different Belgian hospitals (Brussels, Leuven and Bruges) and were aimed at raising public awareness on the disease, with eligible participants restricted to age 40 or above. The images of the healthy subjects at the screening sessions were taken at the same time and show no signs of retinal changes. However, they do hold additional information because of small changes due to focus, lighting, eye movement etc. Additionally, a set of normal fundus images (168) from 88 individuals (±1.9 images per individual, both eyes when applicable) were sourced from a 2016 glaucoma screening program at the University Hospitals Leuven. This resulted in a total set of 1782 images of 491 non-glaucoma individuals. For all images, information provided to data processor was limited to an anonymized patient identifier and glaucoma type. The glaucoma diagnoses (following ICD standards) linked with the fundus images were obtained through a full ophthalmologic examination. Patients were subjected to neuroretinal rim and nerve fibre layer analysis using either OCT (Spectralis OCT; Heidelberg Engineering, Heidelberg, Germany) or confocal scanning laser ophthalmoscopy (Heidelberg retinal tomography; HRT; Heidelberg Engineering, Heidelberg, Germany), tonometry (Goldmann Applanation Tonometry; Haag-Streit AT900; Köniz, Switzerland) and visual field testing (Humphrey Visual Field Analyzer; Carl Zeiss

Meditec, Jena, Germany). The glaucoma experts are aware of the so-called red and green disease surrounding OCT results and do verify the actual images to look for any artifacts or other sources of misinterpretation and check the reliability of the analysis. The transition HRT to Spectralis OCT device rolled out in 2015. Patients that were followed up prior to the switch are still imaged with HRT to ensure consistent progression analysis. Visual field testing at the glaucoma clinic of UZ Leuven is achieved through Humphrey or Octopus standard automated perimetry. Clinicians look for typical glaucomatous visual field defects such as wedge shape defects, steps or nasal breakthrough. The glaucoma experts at UZ Leuven incorporated progression analysis when available, to ensure accurate glaucoma diagnosis. The images sourced from the screening program were evaluated by two glaucoma experts without the aid of OCT and VF tests, but did include a slit lamp biomicroscopic examination including fundoscopy by a glaucoma expert.

Image preprocessing

All 8433 images were manually inspected by two independent retinal image experts to control for quality, omitting images without visible optic disc. Because of the high-quality glaucoma labels based on a full ophthalmological examination, even poor images can be used during training, to increase the robustness of the deep learning model. This quality control does not match the quality that human experts require for diagnosis, hence the task being carried out by retinal image experts with experience in deep learning in ophthalmology. Quality assessed images deemed fit for analysis were initially centre cropped to a square of 1016 × 1016, removing any risk of influence caused by the image border, and subsequently resized to 224 × 224 to match the input layer of the ResNet-50 (He et al. 2016) neural network architecture.

Colour fundus images are characterized by a large intra-image variance in intensity levels mainly due to the curvature of the retina. Therefore, the images were convolved with a Gaussian kernel (30 × 30) to estimate its background, and this was deducted from the original image. The result is a data set of standardized fundus images, as illustrated in the top left of Fig. 2.

In this study, data augmentation was implemented to artificially increase the number of original images used to train the CNN. Augmentation techniques included in the training process of the final model were: horizontal flip, brightness shift and minor elastic deformation. All image augmentations were randomly generated at the start of each mini-batch, as can be seen in the top right of Fig. 2.

Transfer learning

This study used the publicly available Keras (v2.2.0, TensorFlow v1.4.1 backend) ResNet-50 encoder pretrained on ImageNet (Deng et al. 2009), followed by additional layers to increase regularization. The complete deep neural network counted 182 layers of mathematical operations including convolutions and batch normalization (see supplementary material for full network details). During training, all pretrained encoder layers were frozen, except for the last 12 layers, to allow the model to learn features relevant for glaucoma detection. Standard binary cross-entropy was used as cost function, and the Adam (Kingma & Ba 2015) optimizer was used with a constant learning rate of 0.0001.

Active learning

The employed ResNet-50 encoder features over 25 million parameters, requiring a high amount of unique training data to reach its full potential. This study opted for uncertainty sampling as the active learning criterion because of its widespread application in image classification (Joshi et al. 2009).

Uncertainty sampling refers to selecting new samples based on their close distance to the decision boundary set by the classification system, which corresponds to a higher uncertainty. By querying these labels first, the classifier is expected to reduce its uncertainty on these data, more quickly converging to a stable solution. To benchmark the performance of this heuristic, this study also conducted an experiment in which data to be labelled are sampled at random (see Fig. 1 and supplementary material for sampling details).

Saliency maps

The Keras Visualization Toolkit (Kotikalapudi 2017) was used to

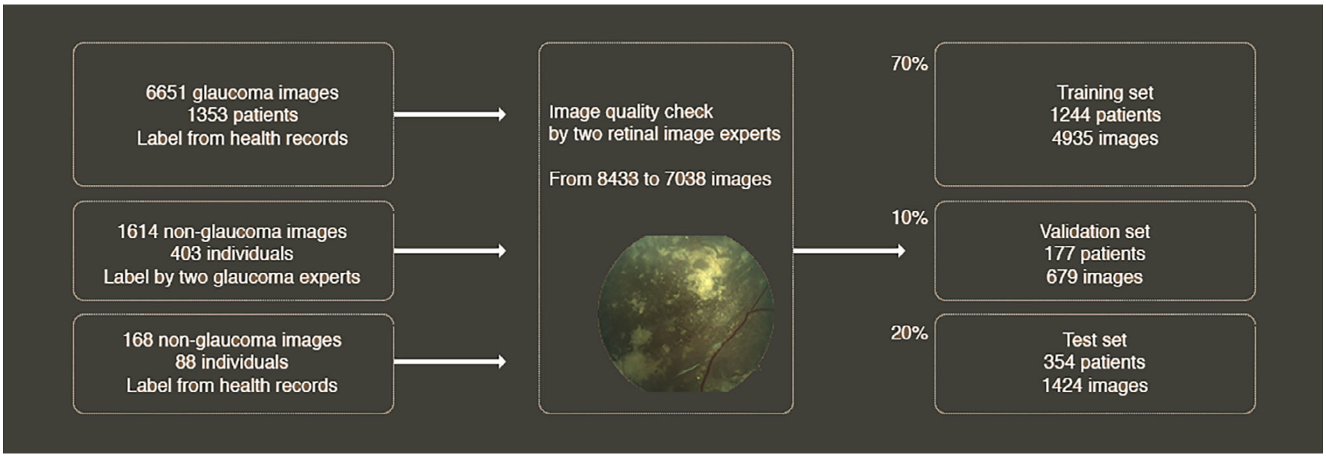

Fig. 1. Top: Overview of data used, effect of image quality control, and subdivision in training, validation and test set.

generate saliency maps. Saliency maps for deep learning accentuate the pixels (coloured reddish) that contribute the most to the classification output, that is if that pixel were to change, the classification output would be likely to change as well (Simonyan et al. 2014). The generated saliency maps of (i) randomly selected images classified correctly by the trained model, and (ii) the false positives (FP) and false negatives (FN) were subsequently examined by two blinded glaucoma experts.

In order to reveal a pattern, saliency maps of thirty oculus dextrus fundus images were manually aligned and averaged. The average saliency map was divided into six zones commonly used in ONH analysis, with differences in saliency intensity quantified.

Evaluation metrics

All predictions by the deep learning models were evaluated against the ground truth label provided by the University Hospitals Leuven. Area under the receiver operating characteristic curve (AUC) was selected as main performance metric, with specificity and sensitivity also reported. The evaluation phase was conducted using the SciPy Python library (Jones et al. 2001).

Results

A total number of 7038 images (83.5% of originally pooled number of images) of 1775 patients passed the manual image quality assessment and were further used in this study. Selected images of 1775 patients were allocated to training (70%; 1244 patients; 4935 images), validation (10%; 177 patients, 679 images), and test set (20%; 354 patients, 1424 images), based on anonymized patient identifier, ensuring that all images from the same patient were to be found in the same class. All glaucoma detection experiments were evaluated on the validation set of 679 images as proxy to select the optimal state of trainable parameters.

Final results are reported on the independent test set of 1424 unique images, corresponding to 354 individuals. For patient level prediction, all glaucoma predictions of images belonging to the same patient are averaged and then classified based on the 0.5 cut-off. Results on the patient level were considered more appropriate for interpretation of the results, as referral decisions would be made on the patient level. Table 1 outlines classification results for glaucoma detection (glaucoma vs non-glaucoma, abbreviated by **GLC** and **NO**) for the active learning experiments and baseline model with all training images and labels included at start of the training process. Confusion matrix and performance metrics are given, computed over the original image with test-time augmentation (TTA). The latter corresponds to randomly augmenting the image tenfold, using the same techniques as in training, followed by averaging the prediction probabilities in order to decrease prediction uncertainty. The use of TTA led to reductions in AUC error up to 14%.

Table 1. Glaucoma detection with transfer and active learning – quantitative results.

Model		Confusion matrix				Metrics	
Uncertainty sampling (2072 training images; 42% of training set)	Image	True	Predicted				
				NO	GLC	Sensitivity	96%
			NO	324	47	Specificity	87%
			GLC	42	1011	AUC	0.983
	Patient	True	Predicted				
				NO	GLC	Sensitivity	98%
			NO	91	9	Specificity	91%
			GLC	5	249	AUC	0.995
Random sampling (2072 training images; 42% of training set)	Image	True	Predicted				
				NO	GLC	Sensitivity	96%
			NO	305	66	Specificity	81%
			GLC	46	1007	AUC	0.972
	Patient	True	Predicted				
				NO	GLC	Sensitivity	98%
			NO	84	16	Specificity	84%
			GLC	5	249	AUC	0.986
Baseline ResNet-50 CNN (4935 training images, complete training set)	Image	True	Predicted				
				NO	GLC	Sensitivity	96%
			NO	346	25	Specificity	93%
			GLC	42	1011	AUC	0.986
	Patient	True	Predicted				
				NO	GLC	Sensitivity	99%
			NO	93	7	Specificity	93%
			GLC	2	252	AUC	0.996

Final models for the two active learning experiments were selected at 2072 training images, due to the marginal improvements when using additional data (Fig. 2, graph bottom right). After seeing 42% (2072 images) of the training data, the model following the active learning strategy achieved an AUC of 0.995, with sensitivity at 98% and specificity at 91% on the test set, clearly benefitting from the employed heuristic that leveraged uncertainty information (Table 1). The performance gap is the most prominent when comparing the specificity of both models, with the random sampling technique yielding a modest 84% on patient level.

The baseline model trained with all original 4935 training images (accompanied by a large set of artificial images following data augmentation) obtained an AUC of 0.996 on patient referral level. Sensitivity and specificity reach 99.2% and 93%, respectively, corresponding to a low number of false negatives (2) and false positives (7). The grouping of images at the patient level led to a reduction in misclassification. Images of misclassified patients were reviewed by two ophthalmologists specialized in glaucoma. False positives could be grouped into (1) subpar image quality due to blurriness or artefacts like eyelashes (n = 4) and (2) signs of other ocular diseases like macular drusen (n = 1) and (3) peripapillary atrophy (n = 2). The fundus image of one false negative patient did not display any clear signs of glaucoma onset, while the other one was a true FN.

The saliency analysis, aimed at explaining the classifier's decision process, is given in Fig. 3. Careful analysis of over 500 saliency maps by two glaucoma experts revealed a recurrent pattern of elevated saliency in

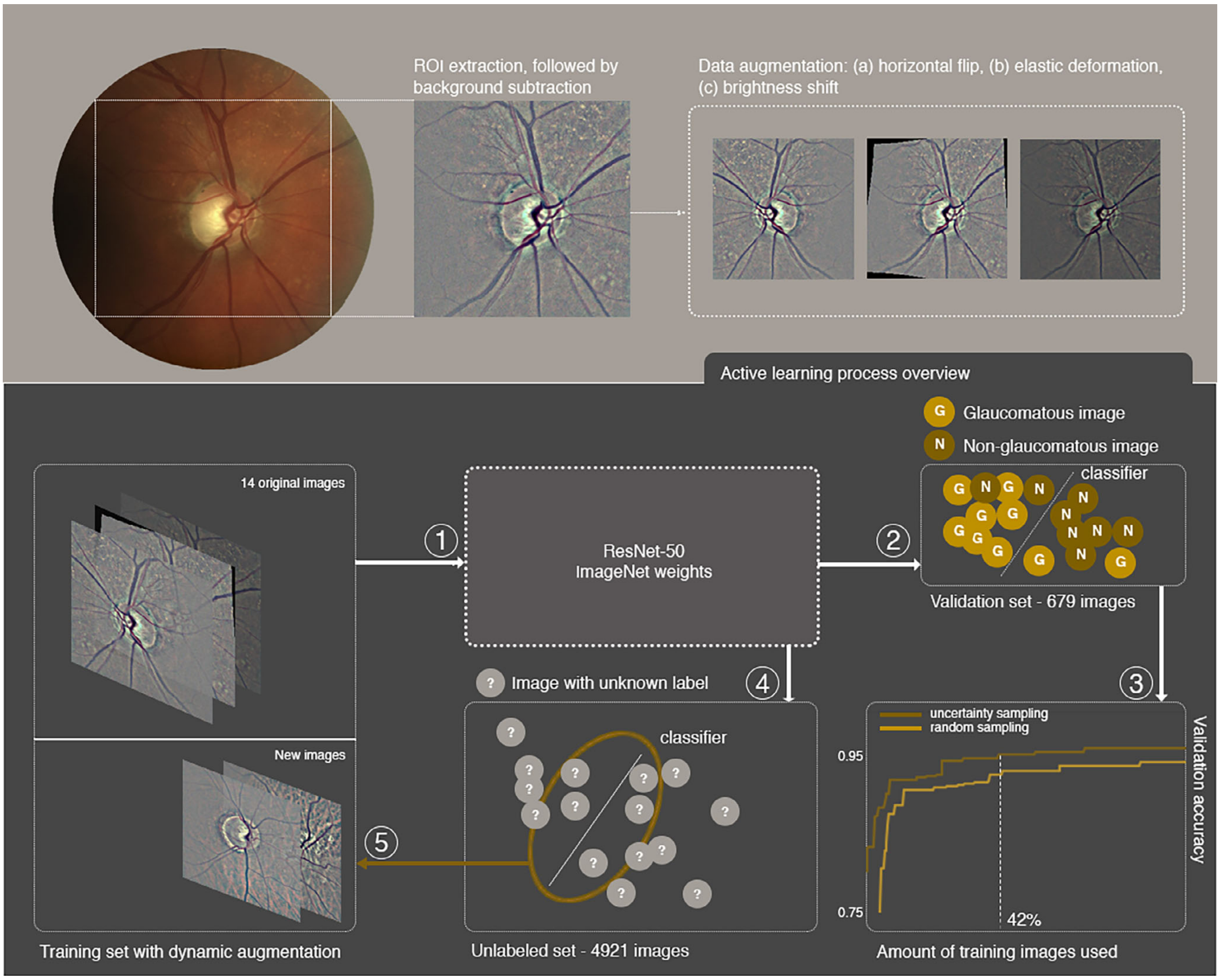
ROI extraction, followed by background subtraction
Data augmentation: (a) horizontal flip, (b) elastic deformation, (c) brightness shift
Active learning process overview
G Glaucomatous image
N Non-glaucomatous image
14 original images
New images
1
ResNet-50
ImageNet weights
2
classifier
Validation set - 679 images
3
? Image with unknown label
4
classifier
5
uncertainty sampling
random sampling
0.95
0.75
42%
Validation accuracy
Training set with dynamic augmentation
Unlabeled set - 4921 images
Amount of training images used

Fig. 2. Top: Overview of image preprocessing (ROI extraction, background subtraction) and data augmentation (horizontal flip, elastic deformation and brightness shift). Bottom: Overview of active learning process. 1: 14 preprocessed and augmented fundus images were used to finetune a CNN with pretrained ImageNet weights. 2: After convergence (no improvement of validation accuracy for two epochs), the model was validated on 679 images, with the results of each active learning iteration visualized in 3. 4: The model was also evaluated on an unlabelled set of (non-)glaucomatous images, with the 14 most uncertain samples (or random samples) transferred to the training set (5). This process was repeated until the unlabelled set was depleted.

inferotemporal and superotemporal zones, either within (early/moderate stage, remaining neuroretinal rim) or outside (late stage, complete thinning) the ONH. This recurrent pattern was subsequently confirmed through the averaging of thirty optic disc-aligned saliency maps.

Discussion

This study resulted in an accurate deep learning-based glaucoma classifier, achieving patient referral AUC of 0.995 on 1424 test images from 354 individuals, with only 42% (2072 images) of the complete training set (4935 images) used. The joint forces of transfer and active learning foster potential in the domain of glaucoma classification from fundus images, allowing model training with a 58% reduction in labelling requirements.

The development of a baseline model trained with all available training data (4935 images) and transfer learning yielded an AUC of 0.996, sensitivity and specificity of 99.2% and 93%, on the test set. The merits of transfer learning in the field of automated glaucoma detection using fundus images have been illustrated using both small (Ahn et al. 2018; Shibata et al. 2018) and large (Christopher et al. 2018; Li et al. 2018a,b) (>5000 images) data sets. Li et al. (2018a) trained a CNN for glaucoma classification using a data set of 48116 images, reporting AUC, sensitivity and specificity of 0.986, 95.6% and 92%, respectively. While the efforts to reach a labelled data set of this size are to be commended, one could question whether the same performance can be reached in a more cost-effective manner, with significantly less labelled fundus images used during training. Annotated medical image data are hard to gather, with images and associated glaucoma diagnosis employed in this study generated over several years. The field of active learning encompasses a set of techniques that accelerate training by querying experts for

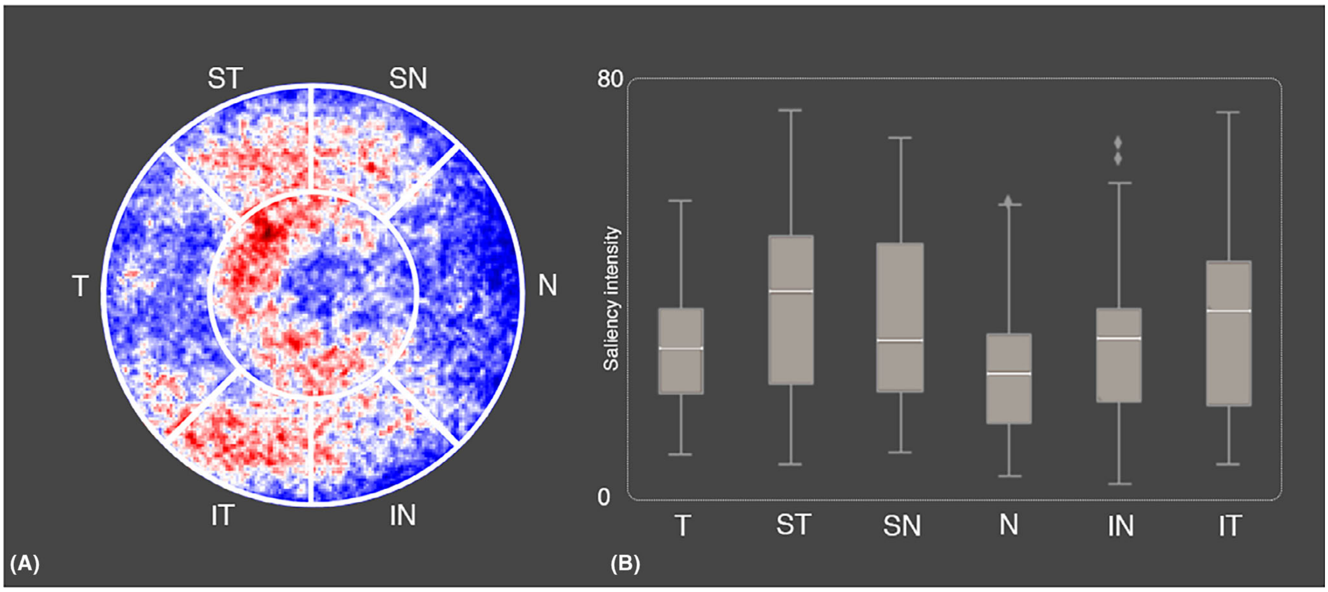

Fig. 3. (A) Average saliency map of thirty aligned oculus dextrus images split into six sectors, with reddish colour corresponding to high saliency (important area for glaucoma classification); optic disc is contained inside the inner circle. (B) Quantification of average saliency in the six zones outside the disc area (complementary to A).

labels that would benefit the classification system the most. In this study, the addition of an active learning component resulted in a model with 42% of the data used, while still attaining an AUC of 0.995 on patient referral.

Two trained glaucoma clinicians analysed more than 500 saliency maps, accompanied by the original glaucomatous images, and indicate a recurrent pattern of salient regions in the inferotemporal and superotemporal zones neighbouring the ONH. These regions likely correspond to the RNFL areas that are affected as a result of glaucoma. The hypothesis of a recurrent pattern of elevated saliency in inferotemporal and superotemporal regions was supported by a statistical analysis using the manually aligned average of 30 randomly selected saliency maps (Fig. 3). The centre part (disc area) of the average saliency map provides additional evidence on a significant concentration of salient regions in the inferior, temporal and superotemporal region of the ONH. The latter partly matches the findings described by Christopher et al. (2018), who used an occlusion-based strategy to reveal salient regions in inferior and superior zones within the disc. This study is the first to indicate that regions outside the ONH could be valuable in glaucoma classification using deep learning. We aim to further investigate the importance of RNFL defects in glaucoma classification from fundus images in future work.

Manual image quality assessment led to 83.5% of available fundus images being actually of sufficient quality for analysis in this original investigation. Two retinal image experts graded each image, omitting those with an excessive presence of camera artefacts or missing optic nerve head. Image quality is essential to ensure proper functioning of the convolutional neural network. In this study, the latter is backed up by the analysis of false positives and false negatives by two ophthalmologists (performed in a blind manner), who indicated subpar image quality to be the culprit in several cases.

This study has several limitations. The class distribution, with over 70% glaucomatous images, is far from the real-life prevalence one would encounter at screening sessions. The selected data imbalance is due to the small availability of non-glaucomatous images, which are often not stored in hospitals. In addition, a large set of the glaucoma images are intermediate or late stage (based on neuroretinal rim assessment), while an important application of glaucoma classification with deep learning could be early detection. Finally, the models trained and validated in this study used images of mainly Caucasian patients that were captured with a fundus camera device from one vendor. To overcome this limitation, we are extending our work by validating and refining our current model using heterogenous data sets obtained through international collaborations, with the goal to develop a model suitable for global screening.

Conclusions

This study achieves state-of-the-art results for automated glaucoma referral with a 60% decrease in labelling cost through the combination of transfer learning, careful data augmentation, and uncertainty sampling, a heuristic commonly used in the domain of active learning. Our iterative sampling process provides novel evidence that deep learning can achieve excellent performance in glaucoma classification, even when using a limited amount of labelled training data. These findings should motivate research groups that have access to less data to help to advance the field of artificial intelligence applied to ophthalmology. Finally, this study provides novel insights into the decision-making process of the trained deep learning glaucoma classifier through the averaging of saliency maps, which seem to be highlighting inferior and superior neuroretinal rim thinning (within ONH) as well as RNFL defects in superotemporal and inferotemporal zones (outside ONH).

References

Ahn JM, Kim S, Ahn KS, Cho SH, Lee KB & Kim US (2018): A deep learning model for the detection of both advanced and early glaucoma using fundus photography. PLoS ONE **13**(11): e0207982.

Asaoka R, Murata H, Iwase A & Araie M (2016): Detecting preperimetric glaucoma with standard automated perimetry using a deep learning classifier. Ophthalmology **123** (9): 1974–1980.

Burlina PM, Joshi N, Pacheco KD, Freund DE, Kong J & Bressler NM (2018): Use of deep learning for detailed severity characterization and estimation of 5-year risk among patients with age-related macular degeneration. JAMA Ophthalmol **136**(12): 1359–1366.

Burr J, Hernández R, Ramsay C et al. (2014): Is it worthwhile to conduct a randomized controlled trial of glaucoma screening in the United Kingdom? J Health Serv Res Policy **19**(1): 42–51.

Chen X, Xu Y, Kee Wong DW, Wong TY & Liu J (2015): Glaucoma detection based on deep convolutional neural network. 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan 715–718. https://doi.org/10.1109/embc.2015.7318462

Christopher M, Belghith A, Bowd C et al. (2018): Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. Sci Rep **8**(1): 16685.

Deng J, Dong W, Socher R, Li L-J, Li K & Fei-Fei L (2009): ImageNet: A Large-Scale Hierarchical Image Database. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL 248–255. https://doi.org/10.1109/cvpr.2009.5206848

Ervin AM, Boland MV, Myrowitz EH et al. (2012): Screening for Glaucoma: Comparative Effectiveness. Rockville (MD): Agency for Healthcare Research and Quality (US). (Comparative Effectiveness Reviews, No. 59.)

Fu H, Cheng J, Xu Y, Wong DWK, 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**(7): 1597–1605.

Grassmann F, Mengelkamp J, Brandl C et al. (2018): A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. Ophthalmology **125**(9): 1410–1420.

Gulshan V, Peng L, Coram M et al. (2016): Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA **316**(22): 2402–2410.

He K, Zhang X, Ren S & Sun J (2016): Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Piscataway, NJ 771–778.

van der Heijden AA, Abramoff MD, Verbraak F, Hecke MV, Liem A & Nijpels G (2018): Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. Acta Ophthalmol **96**: 63–68.

Jones E, Oliphant E & Peterson P (2001): SciPy: Open Source Scientific Tools for Python. http://www.scipy.org/

Joshi AJ, Porikli F & Papanikolopoulos N (2009): Multi-class active learning for image classification. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, 2372-2379. https://doi.org/10.1109/cvpr.2009.5206627

Kapetanakis VV, Chan MPY, Foster PJ et al. (2006): Global variations and time trends in the prevalence of primary open angle glaucoma (POAG): a systematic review and meta-analysis. Br J Ophthalmol **100**: 86–93.

Kingma DP & Ba J. (2015): Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR).

Kotikalapudi R (2017): keras-vis. https://github.com/raghakot/keras-vis.

Li Z, He Y, Keel S, Meng W, Chang R & He M (2018a): Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. Ophthalmology **125**(8): 1199–1206.

Li Z, Keel S, Liu C & He M. (2018b): Can artificial intelligence make screening faster, more accurate, and more accessible? Asia Pac J Ophthalmol (Phila) **7**: 436–441.

Maheshwari S, Pachori RB & Acharya UR (2017): Automated diagnosis of glaucoma using empirical wavelet transform and correntropy features extracted from fundus images. IEEE J Biomed Health Inform **21** (3): 803–813.

Matsopoulos GK, Asvestas PA, Delibasis KK, Mouravliansky NA & Zeyen TG (2008): Detection of glaucomatous change based on vessel shape analysis. Comput Med Imaging Graph **32**(3): 183–192.

Muhammad H, Fuchs TJ, De Cuir N et al. (2017): Hybrid deep learning on single wide-field optical coherence tomography scans accurately classifies glaucoma suspects. J Glaucoma **26**(12): 1086–1094.

Settles B (2009): Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Shibata N, Tanito M, Mitsuhashi K et al. (2018): Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. Sci Rep **8**(1): 14665.

Simonyan K, Vedaldi A & Zisserman A (2014): Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. Proceedings of the 2014 International Conference on Learning Representations (ICLR).

Tham Y, Li X, Wong T, Quigley H, Aung T & Cheng C (2014): Global Prevalence of Glaucoma and Projections of Glaucoma Burden through 2040. Ophthalmology **121** (11): 2081–2090.

Ting DSW, Pasquale LR, Peng L et al. (2019): Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol **103** 167–175.

Tuulonen A. (2011): Cost-effectiveness of screening for open angle glaucoma in developed countries. Indian J Ophthalmol **59** (Suppl1): S24–S30.

Wen JC, Lee CS, Keane PA et al. (2018): Forecasting Future Humphrey Visual Fields Using Deep Learning. arXiv e-prints.

Received on February 5th, 2019.
Accepted on June 26th, 2019.

Correspondence:
Ruben Hemelings, MS
KU Leuven
VITO
Vito Biologie
Industriezone Vlasmeer 7
2400 Mol
Belgium
Tel: +32472748707
Email: ruben.hemelings@kuleuven.be

The first author is jointly supported by the Research Group Ophthalmology, KU Leuven and VITO NV. No outside entities have been involved in the study design, in the collection, analysis and interpretation of data, in the writing of the manuscript, nor in the decision to submit the manuscript for publication. Thus, the authors declare that there are no conflicts of interest in this work.