Pathological myopia classification with simultaneous lesion segmentation using deep learning

Peer-reviewed author version

Hemelings, Ruben; Elen, Bart; Blaschko, Matthew B.; Jacob, Julie; Stalmans, Ingeborg & DE BOEVER, Patrick (2021) Pathological myopia classification with simultaneous lesion segmentation using deep learning. In: Computer Methods and Programs in Biomedicine, 199 (Art N° 105920).

DOI: 10.1016/j.cmpb.2020.105920
Handle: http://hdl.handle.net/1942/33779

Pathological myopia classification with simultaneous lesion segmentation using deep learning

Authors:

Ruben Hemelings[ae*], MS

Bart Elen[e], MS

Matthew B. Blaschko[c], PhD professor

Julie Jacob[b], MD PhD

Ingeborg Stalmans[ab], MD PhD professor

Patrick De Boever[de], PhD professor

Affiliations:

[a] Research Group Ophthalmology, KU Leuven, Herestraat 49, 3000 Leuven, Belgium

[b] Ophthalmology Department, UZ Leuven, Herestraat 49, 3000 Leuven, Belgium

[c] ESAT-PSI, KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium

[d] Hasselt University, Agoralaan building D, 3590 Diepenbeek, Belgium

[e] VITO NV, Boeretang 200, 2400 Mol, Belgium

*corresponding author

Contact details

Affiliation: KU Leuven, VITO

Postal address: Vito Health, Industriezone Vlasmeer 7, 2400 Mol, Belgium

E-mail: ruben.hemelings@kuleuven.be

Phone: +32472748707

Abstract

Background and Objectives: Pathological myopia (PM) is the seventh leading cause of blindness, with a reported global prevalence up to 3%. Early and automated PM detection from fundus images could aid to prevent blindness in a world population that is characterized by a rising myopia prevalence. We aim to assess the use of convolutional neural networks (CNNs) for the detection of PM and semantic segmentation of myopia-induced lesions from fundus images on a recently introduced reference data set.

Methods: This investigation reports on the results of CNNs developed for the recently introduced Pathological Myopia (PALM) dataset, which consists of 1200 images. Our CNN bundles lesion segmentation and PM classification, as the two tasks are heavily intertwined. Domain knowledge is also inserted through the introduction of a new Optic Nerve Head (ONH)-based prediction enhancement for the segmentation of atrophy and fovea localization. Finally, we are the first to approach fovea localization using segmentation instead of detection or regression models. Evaluation metrics include area under the receiver operating characteristic curve (AUC) for PM detection, Euclidean distance for fovea localization, and Dice and F1 metrics for the semantic segmentation tasks (optic disc, retinal atrophy and retinal detachment).

Results: Models trained with 400 available training images achieved an AUC of 0.9867 for PM detection, and a Euclidean distance of 58.27 pixels on the fovea localization task, evaluated on a test set of 400 images. Dice and F1 metrics for semantic segmentation of lesions scored 0.9303 and 0.9869 on optic disc, 0.8001 and 0.9135 on retinal atrophy, and 0.8073 and 0.7059 on retinal detachment, respectively.

Conclusions: We report a successful approach for a simultaneous classification of pathological myopia and segmentation of associated lesions. Our work was acknowledged with an award in the context of the “Pathological Myopia detection from retinal images” challenge held during the IEEE International Symposium on Biomedical Imaging (April 2019). Considering that (pathological) myopia cases are often identified as false positives and negatives in glaucoma deep learning models, we envisage that the current work could aid in future research to discriminate between glaucomatous and highly-myopic

eyes, complemented by the localization and segmentation of landmarks such as fovea, optic disc and atrophy.

Key words: pathological myopia, fovea localization, peripapillary atrophy, retinal detachment, convolutional neural network, fundus image, glaucoma

Introduction

Myopia or nearsightedness currently affects approximately 34% of the world population.[1] High myopia, often defined as a spherical equivalent that exceeds -6.00 diopter or an axial length of 26.5mm or more, has a prevalence ranging from 1% in African Americans[2] and up to 5.5% in the Japanese Tajimi study[3]. Approximately 1-3% of the world population develops vision-impairing macular lesions (lacquer cracks, choroidal neovascularization, and Fuchs spots) as a result of high myopia, referred to as myopic maculopathy.[4,5] Both the presence of myopic maculopathy and posterior staphyloma are used to define pathological myopia (PM), which causes uncorrected and irreversible visual impairment.[6] Other retinal changes due to myopia include: fundus tessellation, (peripapillary) atrophy, optic disc tilting, retinal tear and retinal detachment. Additionally, myopia increases the risk of developing open-angle glaucoma[7], presumably because myopic eyes have thinner and weaker lamina cribrosa tissue[8]. Optic nerve head (ONH) changes such as temporal disc flattening and tilting[9], as a consequence of myopia, hampers glaucoma detection through ONH assessment during fundoscopy or fundus image analysis[10]. Peripapillary atrophy (PPA), being attenuation of retinal pigment epithelium (RPE) neighboring the ONH, is associated with both myopia and glaucoma, and is one of the causes for a high number of myopic patients being diagnosed as glaucoma suspects.

Previous work on automated pathological myopia detection from retinal images is limited. Liu et al described a methodology dubbed PAMELA (Pathological Myopia Detection Through Peripapillary Atrophy), in which a support vector machine (SVM) is trained using exclusively PPA texture features from fundus images.[11] They reported sensitivity and specificity of 0.85 and 0.90, respectively, on 40 test images. As mentioned above, PPA is not unique to pathological myopia, and not the only retinal change induced by the disease.

Zhang et al also employed an SVM, but expanded on the feature set by incorporating additional retinal information such as ONH-related parameters and socio-demographic variables including age and race.[12] 10-fold cross validation led to accuracies ranging from 84.9% to 89.3% on a private data set encompassing imaged eyes of 800 primary school students.

Deep learning-based classification of pathological myopia has not been previously explored, although convolutional neural networks (CNNs) are showing great potential in ophthalmic research for disease identification and staging[13]. Relevant for this manuscript is refraction estimation from fundus images using deep learning by Varadarajan et al., who developed a regression model that estimates refractive error with high accuracy (<1 diopter mean absolute error).[14] Their approach could be useful in stratifying fundus images into emmetropia (normal refraction), hyperopia (farsightedness), myopia (nearsightedness), and high myopia (exceeding -6.0 diopters). The last group could then be further analyzed to detect myopia-induced lesions.

Semantic segmentation or pixel-wise classification has experienced major advances through the introduction of fully convolutional networks (FCN) in 2015.[15] For fundus images, ample FCN-based segmentation networks have been described in popular tasks like vessel extraction[16], artery/vein discrimination[17], and optic cup/disc estimation[18]. Recent work on retinal lesion segmentation in fundus images is dominated by microaneurysms, hard exudates and cotton wool spots induced by diabetic retinopathy.[19] Segmentation of myopia-related lesions (e.g. PPA) from fundus images has been obtained using classic computer vision methods. Lu et al. employed a modified Chan-Vese segmentation tool with shape constraints to delineate both optic disc and PPA, reporting 92.5% accuracy in PPA size estimation on 40 test images.[20]

Here, we report our CNN-based methods and results developed for the classification of (non-)pathological myopia, fovea localization, and semantic segmentation of optic disc, retinal atrophy and detachment on a novel reference data set. The multitude of tasks encouraged us to fuse classification and segmentation tasks when proven to be beneficial on the validation set. Joint disease classification and lesion segmentation systems have been described in deep learning literature, leading to improved

classification performance.[21] We also introduce a novel ONH-based prediction enhancement that results in improved performance for the tasks of lesion segmentation and fovea localization. The latter task is being obtained through a segmentation approach for the first time, improving vastly on coordinate regression. Our results are benchmarked against a holdout validation set, other state-of-the-art methods, and evaluated on external labeled data sets where possible.

Methodology

Dataset and evaluation

Retinal images were made available in the context of the "Pathological Myopia detection from retinal images" challenge held on the occasion of the IEEE International Symposium on Biomedical Imaging organized in April 2019.[22] The PALM dataset consists of 1200 anonymized color fundus images that were captured with a Zeiss VISUCAM device at a 45° angle with a 2124 x 2156 resolution or 30° with a 1444 x 1444. The images are macula- or optic disc-centered of left eyes with no disclosure of the number of different eyes or patients that were included in the dataset. The 1200 images are split into equally sized train, validation, and test sets sharing the same characteristics. Publicly available labels for the training set of 400 images encompass (1) the binary label for (non-)pathological myopia classification, (2) cartesian coordinates corresponding to the location of the fovea, and (3) semantic segmentation ground truth on pixel level for optic disc, peripapillary/retinal atrophy and retinal detachment. The myopia labels were extracted from the health records of the Zhongshan Ophthalmic Center, Sun Yat-sen University (China) and were determined during an ophthalmic examination, including optical coherence tomography (OCT) and visual field (VF) testing. The fovea coordinates and segmentation masks were generated by seven independent ophthalmologists from the same clinic. The PM detection training labels are balanced (53% PM images), but do not match the prevalence encountered in screening context (up to 3%). Ground truth of optic discs is available for most images, with an empty ground truth mask in case of an absent or partially visible disc. An overview of official training set characteristics is provided in Table 1. Differences in PM and non-PM characteristics were analyzed using a two-tailed t-test.

PM detection was quantified using area under the receiver operating characteristic (AUC), while the fovea localization was evaluated using the average Euclidean distance between the predicted cartesian coordinates and ground truth. The three predicted segmentation masks (optic disc, atrophy, detachment) were evaluated using a weighted combination of Dice[23] similarity coefficient (segmentation) and test's accuracy using the F1 score (detection). See supplementary information for full details on evaluation framework as defined by PALM organizers.

Additional data to evaluate the generalization ability of trained models was included where possible. For PM detection, we evaluated on the recently-introduced Ocular Disease Intelligent Recognition (ODIR) data set aimed at multi-disease classification.[24] The original competition did not include PM detection as task, but structured labels are available in the file with diagnostic keywords. We selected the subset of images having either 'normal fundus' or 'pathological myopia' in the diagnostic keywords (3350 out of 7000 fundus images). Fovea localization was evaluated on Messidor[25], for which 1136 out of 1200 fundus images have official fovea coordinates.

Network architectures and loss functions

UNet++[26], a nested variant of the widely used U-Net[27], was selected for the segmentation tasks because of its reported improved performance. The widely used ResNet[28] encoders were tested as feature extractors to enable transfer learning with pretrained ImageNet[29] weights. We selected a pretrained ResNet-18 encoder as feature extractor as it satisfies our preset conditions of minimizing the amount of trainable weights (there is a limited amount of labeled training images), while maximizing the input size that fits on GPU memory (larger size yields the best performance for segmentation). At the end of the contracting path (ResNet-18), where the input image is converted to a representation in latent space (shape 9x9x512), we added a second output branch for PM classification in light of co-regularization.[30] Figure 1 displays the full architecture, with the contracting path extracting and refining feature maps through convolutional, batch normalization and pooling layers (ResNet-18). In UNet++, these feature maps are connected to a number of dense convolution blocks, before being inserted in the expanding path (decoder). The principle of dense convolution blocks as extended skip connections

is illustrated in Figure 1 as well (highlighted in dark green). The UNet++ with ResNet-18 encoder amounts to a total of 16 million trainable weights, with the detection branch adding 513 trainable weights because of the additional 1x1 convolutional layer.

The employed loss function for PM classification is standard binary cross-entropy. Fovea localization labels are cartesian coordinates, but were converted to filled circles with varying radii to allow for segmentation, as an alternative approach to standard coordinate regression. All segmentation models employed standard binary/categorical cross-entropy as loss function, complemented by Dice similarity coefficient. Finally, we experimented with the Lovàsz-Softmax[31] as third loss component. The latter serves as a tractable surrogate for the optimization of intersection over union (IoU), and has proven itself as finetuning loss in recent semantic segmentation challenges.[32]

Preprocessing, Data augmentation, Training details

Color fundus images are unevenly illuminated due to the curvature of the retina. Local contrast enhancement through background subtraction estimated by a large Gaussian kernel was used to correct this[33]. Data augmentation techniques used throughout all experiments include random cropping, mild elastic deformation, and horizontal flips. Random cropping was performed selecting patches of 288 x 288 within resized images of a random size between half and original image size to teach the model features at multiple resolutions. Data augmentation was not applied to the 40 holdout images used to select the best model weights. The model input of 288 x 288 was selected based on a balance between the merits of pretrained weights (224 x 224) and segmentation output (higher resolution leads to better results).

Due to the severe class imbalance of the retinal detachment segmentation, we adopted a sampling strategy that oversamples images with retinal detachment at earlier stages of the training process to an equal mini-batch distribution, only to gradually slim down to the original data distribution (x 0.75 per five epochs). As such, the model is less likely to treat the detachment label as noise at training start.

Model development was done in Keras v2.2.4 with TensorFlow v1.4.1 backend. All models used Adam[34] optimizer with a default starting learning rate at 0.001. A plateau callback decreased the learning rate by 25% after ten successive epochs of stagnation in validation metric (Dice). To obtain a wider optimum, model weights were averaged over the last twenty epochs when the learning rate reached a value of $1e^{-5}$.[35] Internal validation was performed on a holdout set of 40 images, representing 10% of available training data.

ONH-based prediction enhancement

Theoretically, there should be no overlap between atrophy and optic nerve head (ONH). Peripapillary atrophy represents loss of RPE and choriocapillaris, which ends/starts in Bruch's membrane opening (BMO), and simultaneously delineates the optic disc boundary. Leveraging this domain knowledge, the optic disc and peripapillary/retinal atrophy segmentation tasks were bundled by fusing the two ground masks. Retinal detachment ground truth does overlap with atrophy in certain cases, hence this ground truth was left unprocessed.

In addition to standard coordinate regression, we rebranded the fovea localization task as a segmentation problem. The ground truth masks were generated by drawing filled circles (varying radii between 25 and 75 pixels) based on the official cartesian coordinates as centroids. The optic disc is located on the nasal side of the fovea. Hence, the optic disc segmentation ground truth was added to the fovea ground truth, to implicitly insert this domain knowledge. We also experimented with the implementation of cutout[36], a common regularization technique, to improve the learning of the ONH – fovea relation.

The predicted fovea segmentations required post-processing in case of missing or unlikely predictions. Two sanity checks were performed prior to reconversion to coordinates: (1) whether there is a fovea prediction made, and (2) whether it falls within normal range compared to optic disc location. Normal range was defined as mean ± 2 x standard deviation, with population mean and deviation estimated from the training labels (grouped by image resolution). If the assertions failed, the predicted fovea coordinates were determined based on optic disc centroid and mean distance between optic disc and

fovea. For benchmarking purposes, we also report on experiments without joint optic disc segmentation. Here, the postprocessing was limited to the use of image center coordinates in case of missing fovea prediction.

Ensembling on image and model level

Ensembling on image and model level tend to lead to small performance gains due to its decrease in prediction variance. Hence, final predictions of (non-)pathological myopia classification on the test images were obtained through commonly-used test-time augmentation (TTA) techniques (elastic deformation and horizontal flips). We further enhanced TTA predictions by ensembling on model level through the averaging of predictions obtained on seven separately trained models with different random seed on train/holdout split. Segmentation results were generated using averaged predictions on overlapping 288 x 288 patches from resized images (288 x 288, 294 x 294, and 302 x 302). Overlapping patches were only possible in the last two resolutions.

Results

Table 1 reveals that the largest group of available training images are 45° macula-centered images, whereas its disc-centered variant contains only 3 images. Complete optic discs are missing in all 30° macula-centered images, and in some PM cases imaged at 45° as well. Optic disc area ranged between 1-4%, and was significantly larger in 30° disc-centered PM images. Retinal atrophy was present in almost all PM cases, and in roughly half of non-PM images. The area covered by atrophy was larger in PM images for all modalities. The fovea is visible in nearly all images.

The Dice score on ONH segmentation was found to be the highest in the vanilla setup with a single model (0.9481 Dice). For retinal atrophy however, multi-class segmentation with Lovász as loss component did lead to better performance (0.6948 Dice) when compared to two individual models (0.6210 Dice). The balanced data generator did lead to better performance in segmentation of retinal detachment (0.9998 Dice).

Table 2 summarizes our quantitative results on a holdout validation set, the official test set, obtained through the online competition evaluation server hosted at http://ai.baidu.com/broad/subordinate?dataset=pm, and external data if available. We also provide the official test results obtained by other onsite PALM participants. All PM cases were correctly classified in both experiments on the holdout validation set (n=40), but the validation loss was significantly lower in the setup with combined ONH and atrophy segmentation (0.0824 versus 0.1146). Our trained models for detection of pathological myopia achieve a final AUC value of 0.986 on the test set. There is no statistical significant difference observed between AUC values among PALM participants (range 0.987-0.997). Without having to retrain the model for PM classification with ONH/atrophy segmentation, a high AUC of 0.924 is recorded on fundus images of the ODIR data set, which is significantly higher compared to using a classification-only model (AUC = 0.858). The ROC curves of both PM models on ODIR are plotted in Figure 2. An overview of all PM experiments and results are given in the first section of Table 2.

The move from regression to segmentation for fovea localization seems to be beneficial, with average Euclidean distance at 229 and 129 pixels, respectively recorded on the internal holdout validation set (n=40). The result using a segmentation approach also improved when employing a larger fovea radius of 75 pixels (110 pixels Euclidean distance). Our proprietary ONH-based prediction enhancement led to a major performance gain (87 pixels Euclidean distance). Finally, the post-processing that deals with missing and unrealistic predictions resulted in the best observed performance (62 pixels Euclidean distance). The result on the official PALM set (n=400) is equivalent, with a Euclidean distance of 58.3 pixels. Euclidean distances reported by other PALM participants differed considerably, ranging from 55.7-172.9. Furthermore, our findings are confirmed on the Messidor data set for which the best performance (lowest Euclidean distance) is also obtained using a segmentation approach complemented with our ONH-based prediction enhancement. A complete results overview for fovea localization can be found in section 2 of Table 2.

Table 2 also shows that the Dice score on ONH segmentation was found to be the highest in the vanilla setup with a single model (0.9481 Dice on holdout validation set). For retinal atrophy however (4th section of Table 2), multi-class segmentation with Lovász as loss component did lead to better performance (0.6948 Dice) when compared to two individual models (0.6210 Dice). ONH segmentation on PALM test data achieved a Dice of 0.93. Other participants reported results ranging from 0.91-0.95. The atrophy segmentation Dice result on the PALM test set (0.8001) is considerably higher than the best Dice recorded on the holdout validation set, which is likely caused by the low number of validation images. Again, there existed a small variability in atrophy segmentation Dice results among participants (0.77-0.82).

Finally, the F1 metric for retinal detachment segmentation reveals that the test set contain 11 cases of retinal detachment. The trained deep learning model identified six correct cases. For this subtask, we obtained the highest Dice score (0.8073) among all participants (0.0030-0.7449), as can be retrieved from the last section of Table 2.

Ground truth for validation and test sets on image level will be made publicly available at a later date by the organizers of the PALM challenge. Hence, the qualitative results of four test images displayed in Figure 3 cannot be visually compared to the official ground truth. The optic disc – outlined in green – is detected in both non-pathological (A) and pathological (B,C) fundus images (not present in D), and does not overlap with peripapillary atrophy (B,C). The fovea – indicated by a cross – is localized well in cases of a clear (A) and covered (C,D) macula, or added during postprocessing (B). Atrophy – outlined in white – is segmented at both peripapillary (A,B,C,D) and macular (B) regions. In images where 30% of the image is predicted to be retinal detachment, the prediction is replaced with the size of the image mask (yellow outline of image C).

Figure 4 showcases two examples of bad segmentations for both atrophy and optic disc tasks. These cases were quantitatively selected on the 40 holdout validation images for which the ground truth is publicly available at this time. For atrophy segmentation, we observe the lowest scores in images that feature a small amount of peripapillary atrophy (often healthy eyes). The highest Dice scores are

obtained on images with a lot of retinal atrophy present (eyes with pathological myopia). For optic disc segmentation, the roles are reversed. Lower performance is recorded in challenging cases with atrophy surrounding the disc; while the highest performance is obtained in healthy eyes.

Discussion

This deep learning study on fundus images describes (1) the detection of pathological myopia (PM), (2) the localization of the fovea, and (3) the segmentation of optic disc, retinal atrophy and retinal detachment. The results are obtained after training on 400 labeled fundus images and relies on transfer learning and co-regularization through weight sharing. The methodology described in the manuscript led to a third place in PALM challenge hosted at ISBI 2019. The PALM dataset provides novel challenges to existing research topics, as myopic optic discs are often tilted (optic disc segmentation), and the fovea obscured due to tessellation and macular atrophy in some cases of pathological myopia (fovea localization).

The PM detection task scored an AUC of 0.9867 on the official test set of 400 images. PM detection from fundus images has not been covered in deep learning literature prior to the launch of PALM. The work of Varadajaran et al (2018) comes closest, but employs a whole different setup. Their goal was to develop a data-driven regression model that estimates refractive error (including cases of pathological myopia), using the spherical equivalent as target. In our investigation, the task of PM detection was approached in a different manner, given the different nature of the task and materials. The definition of PM states that a highly myopic case is converting to pathological once a posterior myopia-specific pathology from axial elongation is developing, such as vision-impairing myopia-induced lesions. This is corroborated by the explorative analysis of the training set, given in Table 1. Retinal atrophy, being progressive RPE thinning and attenuation, is present in 98.3% cases of PM, versus 52.6% in non-PM images (restricted to the modality of 45° macula-centered images). By combining atrophy segmentation and PM classification, one forces the model to focus on lesions as main features that contribute to PM classification. This implies a step towards explainable AI or

sufficient transparency to gain clinicians' trust in the future use of deep learning detection systems in ophthalmology.

All valid PM detection results at the onsite PALM challenge scored above 0.98 AUC. Although an official rank is maintained, there exists no statistical significant difference in results between teams, due to the low amount of test images (at 95% confidence interval). Other participants also relied on transfer learning, but not in combination with segmentation. For example, team Vistalab employed a ResNet-50 pretrained on ImageNet, reporting an AUC of 0.998.[37] Their data augmentation strategy included Gaussian noise addition and random rotations.

Our PM detection model trained on PALM also generalizes well to images captured with multiple fundus cameras (Figure 2). On data from the recent ODIR challenge, we obtain AUC values of 0.858 and 0.924 using a standard classification model and using a combined lesion segmentation branch, respectively. This further illustrates that segmentation on related tasks (myopia-induced lesions) can augment classification performance.

For fovea localization, we obtained the 2nd place among all PALM participants. We initially considered adding a regression branch to the segmentation model for optic disc and retinal atrophy. However, due to subpar performance (229 pixels Euclidean distance), this idea was discarded and replaced by a standalone segmentation model. One potential explanation for poor regression performance could be the combination of scarcity in available regression labels (1 per image) when compared to segmentation labels (1 per pixel), and low variance in coordinate values (the fovea is centrally located in macula-centered images). However, the winning submission by Vistalab did follow a regression approach, using a modified pretrained VGG19[38] model. The main disadvantage of a segmentation approach is the loss of direct optimization on the competition metric. Thoughtful post-processing that relies on domain knowledge further enhanced our final predictions. Fovea localization in fundus images has been investigated with deep learning prior to PALM, but primarily in clean datasets with clear macular depression.[39] To illustrate this, we evaluated on diabetic retinopathy cases from the Messidor data, without retraining. The significantly lower Euclidean distance obtained on this data

emphasizes the difficulty aspect introduced by the novel PALM data. Our domain knowledge insertion – combined optic disc and fovea localization – is considered useful in the move towards general deep learning models that can process large amounts of fundus images with unclear macular regions. Such fovea localizing models can assist future big data research. One application would be the automated image cropping of the macula area to facilitate diabetic retinopathy screening.

The optic disc segmentation model obtained a Dice similarity coefficient of 0.9303, scoring in line with relevant work.[40] Due to axial elongation, myopia induces anatomical changes to the optic nerve head, resulting in tilted and oval-shaped optic discs, often surrounded by peripapillary atrophy. These alterations are significant, as a pretrained optic disc segmentation model on non-myopic fundus images failed to properly delineate the discs in the PALM dataset. Another factor could be the larger optic disc size observed in myopic eyes[41,42]. From Table 1, there is a moderate significance ($P < 0.01$) found between PM and non-PM (which also includes high myopia) in the 30° disc-centered images. Hence, optic disc size is unlikely to be an informative predictor in PM detection.

This original investigation also introduces a pioneering result of 0.8001 Dice on the segmentation of retinal atrophy (PPA, lacquer cracks and Fuch's spots) in fundus images. This type of segmentation can support future research in discriminating between myopia- and glaucoma-induced peripapillary atrophic changes. This is relevant because in previous work it has been observed that false positive and negative predictions in glaucoma classification models are often due to cases of high/degenerative myopia. For example, Liu et al (2019) observed that the most common reason for both false-negative and false-positive grading by their DL model (46.3% and 32.3%) and manual grading (44.2% and 34.0%) was pathological or high myopia.[10] Several studies investigated the discriminatory properties of beta- (area with intact Bruch's membrane) and gamma-PPA (lacking Bruch's membrane) for myopia and glaucoma using OCT, but report contradictory findings and low discriminatory power.[43,44] Another recent study discovered a relationship between PPA shape and glaucoma progression, stating that progression is more correlated with eccentric PPA than concentric PPA.[45] DL may assist in analyzing PPA in a larger set of patients than previous investigations.

The fusion of optic disc and atrophy segmentation tasks ensured no overlap in final predictions. This form of joint prediction increases the odds of generalization to unseen samples (in this case, 800 images split in validation and test set of equal size). Ground truth fusion did lead to better performance for atrophy segmentation, but not for ONH segmentation. Another important motivation for joint training is explainable artificial intelligence, as previously discussed.

Finally, this study reports a top-ranked Dice score of 0.8073 on the task of retinal detachment segmentation. The high performance is mainly due to the correct predictions of empty masks in the high number of cases (~97% of images) without retinal detachment. The actual performance would be much lower when the images with retinal detachment would be isolated. In most cases, retinal detachment covers more than half of the field of view (FOV) in the fundus. Hence, one could question the added value of segmentation over a classification approach.

Other participating teams also heavily relied on the combination of FCN architectures and existing feature extractors pretrained on ImageNet for the segmentation tasks. For optic disc segmentation, Vistalab (2nd place) combined ResNet-34 followed by an Atrous Spatial Pyramid Pooling (ASPP)[46(p)] operator in a U-Net architecture. The winning submission in all segmentation tasks is obtained using a lesion-aware segmentation network described by team PingAn Smart Health.[47] They introduce three innovations: an additional classification branch to aid the network in becoming better aware of lesion presence in images; a custom feature fusion module, and lastly a loss function dubbed edge overlap rate that boosts the accuracy of lesion edge segmentation.

The strengths of our work are significant. We describe a CNN architecture that bundles classification and segmentation tasks when deemed relevant (domain knowledge) and when empirically proven on the validation set. Next, we introduce a new approach to obtain fovea localization in fundus images through the reformulation as a segmentation problem. Further domain knowledge is inserted through a custom ONH-based post-processing scheme that leverages anatomical properties of the retina. We describe and compare our state-of-the-art results on a novel reference data set that is expected to be

widely used. Finally, our models on PM detection and fovea localization generalize well to unseen heterogeneous data sets without recalibration to the target domain.

This study also suffers from several limitations. The ground truth on image level for PALM validation and test sets are currently unavailable, hampering the qualitative comparison of semantic segmentation results, and the calculation of specificity and sensitivity. On the other hand, the introduction of medical labeled datasets and robust online evaluation server should be encouraged, as they allow the objective comparison of innovations in deep learning for medical imaging.

Conclusions

We report a successful approach for a simultaneous classification of pathological myopia and segmentation of associated lesions. These award-winning results were obtained in the context of the "Pathological Myopia detection from retinal images" challenge held on the occasion of the IEEE International Symposium on Biomedical Imaging organized in April 2019. Considering that (pathological) myopia cases are often found as false positives in glaucoma deep learning models, we envision that the current work could aid in future research to discriminate between glaucomatous and highly-myopic eyes, complemented by the localization and segmentation of landmarks such as fovea, optic disc and atrophy.

Acknowledgements

The first author is jointly supported by the Research Group Ophthalmology, KU Leuven and VITO NV. This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. 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.

References

1. Holden BA, Fricke TR, Wilson DA, et al. Global Prevalence of Myopia and High Myopia and Temporal Trends from 2000 through 2050. *Ophthalmology*. 2016;123(5):1036-1042. doi:10.1016/j.ophtha.2016.01.006

2. Katz J, Tielsch JM, Sommer A. Prevalence and risk factors for refractive errors in an adult inner city population. *Invest Ophthalmol Vis Sci*. 1997;38(2):334-340.

3. Sawada A, Tomidokoro A, Araie M, Iwase A, Yamamoto T. Refractive Errors in an Elderly Japanese Population: The Tajimi Study. *Ophthalmology*. 2008;115(2):363-370.e3. doi:10.1016/j.ophtha.2007.03.075

4. Vongphanit J, Mitchell P, Wang JJ. Prevalence and progression of myopic retinopathy in an older population. *Ophthalmology*. 2002;109(4):704-711. doi:10.1016/S0161-6420(01)01024-7

5. Liu HH, Xu L, Wang YX, Wang S, You QS, Jonas JB. Prevalence and Progression of Myopic Retinopathy in Chinese Adults: The Beijing Eye Study. *Ophthalmology*. 2010;117(9):1763-1768. doi:10.1016/j.ophtha.2010.01.020

6. Ohno-Matsui K. WHAT IS THE FUNDAMENTAL NATURE OF PATHOLOGIC MYOPIA?: *Retina*. 2017;37(6):1043-1048. doi:10.1097/IAE.0000000000001348

7. Marcus MW, de Vries MM, Montolio FGJ, Jansonius NM. Myopia as a Risk Factor for Open-Angle Glaucoma: A Systematic Review and Meta-Analysis. *Ophthalmology*. 2011;118(10):1989-1994.e2. doi:10.1016/j.ophtha.2011.03.012

8. Yun S-C, Hahn IK, Sung KR, Yoon JY, Jeong D, Chung HS. Lamina cribrosa depth according to the level of axial length in normal and glaucomatous eyes. *Graefes Arch Clin Exp Ophthalmol*. 2015;253(12):2247-2253. doi:10.1007/s00417-015-3131-y

9. Mitchell P, Hourihan F, Sandbach J, Jin Wang J. The relationship between glaucoma and myopia: The blue mountains eye study. *Ophthalmology*. 1999;106(10):2010-2015. doi:10.1016/S0161-6420(99)90416-5

10. Liu H, Li L, Wormstone IM, et al. Development and Validation of a Deep Learning System to Detect Glaucomatous Optic Neuropathy Using Fundus Photographs. *JAMA Ophthalmol*. Published online September 12, 2019. doi:10.1001/jamaophthalmol.2019.3501

11. Liu J, Wong DWK, Lim JH, et al. Detection of Pathological Myopia by PAMELA with Texture-Based Features through an SVM Approach. Journal of Healthcare Engineering. doi:https://doi.org/10.1260/2040-2295.1.1.1

12. Zhang Z, Jun Cheng, Liu J, Yeo Cher May Sheri, Chui Chee Kong, Saw Seang Mei. Pathological Myopia detection from selective fundus image features. In: *2012 7th IEEE Conference on Industrial Electronics and Applications (ICIEA)*. ; 2012:1742-1745. doi:10.1109/ICIEA.2012.6361007

13. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol*. 2019;103(2):167-175. doi:10.1136/bjophthalmol-2018-313173

14. Varadarajan AV, Poplin R, Blumer K, et al. Deep Learning for Predicting Refractive Error From Retinal Fundus Images. *Invest Ophthalmol Vis Sci*. 2018;59(7):2861-2868. doi:10.1167/iovs.18-23887

15. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ; 2015:3431-3440. doi:10.1109/CVPR.2015.7298965

16. Liskowski P, Krawiec K. Segmenting Retinal Blood Vessels With Deep Neural Networks. *IEEE Trans Med Imaging*. 2016;35(11):2369-2380. doi:10.1109/TMI.2016.2546227

17. Hemelings R, Elen B, Stalmans I, Van Keer K, De Boever P, Blaschko MB. Artery–vein segmentation in fundus images using a fully convolutional network. *Comput Med Imaging Graph*. 2019;76:101636. doi:10.1016/j.compmedimag.2019.05.004

18. Fu H, Cheng J, Xu Y, Wong DWK, Liu J, Cao X. Joint Optic Disc and Cup Segmentation Based on Multi-Label Deep Network and Polar Transformation. *IEEE Trans Med Imaging*. 2018;37(7):1597-1605. doi:10.1109/TMI.2018.2791488

19. Orlando JI, Prokofyeva E, Del Fresno M, Blaschko MB. An ensemble deep learning based approach for red lesion detection in fundus images. *Comput Methods Programs Biomed*. 2018;153:115-127. doi:10.1016/j.cmpb.2017.10.017

20. Lu C-K, Tang TB, Laude A, Deary IJ, Dhillon B, Murray AF. Quantification of Parapapillary Atrophy and Optic Disc. *Investig Opthalmology Vis Sci*. 2011;52(7):4671. doi:10.1167/iovs.10-6572

21. Xie Y, Zhang J, Xia Y, Shen C. A Mutual Bootstrapping Model for Automated Skin Lesion Segmentation and Classification. *IEEE Trans Med Imaging*. 2020;39(7):2482-2493. doi:10.1109/TMI.2020.2972964

22. Huazhu Fu FL José Ignacio Orlando, Hrvoje Bogunović, Xu Sun, Jingan Liao, Yanwu Xu, Shaochong Zhang, Xiulan Zhang. PALM: PAthoLogic Myopia Challenge. Published online 2019. 10.21227/55pk-8z03

23. Dice LR. Measures of the Amount of Ecologic Association Between Species. *Ecology*. 1945;26(3):297-302. doi:10.2307/1932409

24. introduction - Grand Challenge. grand-challenge.org. Accessed November 23, 2020. https://odir2019.grand-challenge.org/

25. Decencière E, Zhang X, Cazuguel G, et al. FEEDBACK ON A PUBLICLY DISTRIBUTED IMAGE DATABASE: THE MESSIDOR DATABASE. *Image Anal Stereol*. 2014;33(3):231-234. doi:10.5566/ias.1155

26. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In: Stoyanov D, Taylor Z, Carneiro G, et al., eds. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Lecture Notes in Computer Science. Springer International Publishing; 2018:3-11. doi:10.1007/978-3-030-00889-5_1

27. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, eds. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Lecture Notes in Computer Science. Springer International Publishing; 2015:234-241. doi:10.1007/978-3-319-24574-4_28

28. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. *ArXivorg Ithaca*. Published online December 10, 2015. Accessed November 20, 2019. http://search.proquest.com/docview/2083823373?rfr_id=info%3Axri%2Fsid%3Aprimo

29. Deng J, Dong W, Socher R, Li L-J, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. ; 2009:248-255. doi:10.1109/CVPR.2009.5206848

30. Cai Z, Fan Q, Feris RS, Vasconcelos N. A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection. In: Leibe B, Matas J, Sebe N, Welling M, eds. *Computer Vision – ECCV 2016*. Lecture Notes in Computer Science. Springer International Publishing; 2016:354-370. doi:10.1007/978-3-319-46493-0_22

31. Berman M, Triki AR, Blaschko MB. The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. *ArXivorg Ithaca*. Published online April 9, 2018. Accessed November 20, 2019. http://search.proquest.com/docview/2071981122?rfr_id=info%3Axri%2Fsid%3Aprimo

32. Babakhin Y, Sanakoyeu A, Kitamura H. Semi-Supervised Segmentation of Salt Bodies in Seismic Images using an Ensemble of Convolutional Neural Networks. *CoRR*. 2019;abs/1904.04445. http://arxiv.org/abs/1904.04445

33. Hemelings R, Elen B, Barbosa-Breda J, et al. Accurate prediction of glaucoma from colour fundus images with a convolutional neural network that relies on active and transfer learning. *Acta Ophthalmol (Copenh)*. n/a(n/a). doi:10.1111/aos.14193

34. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *ArXivorg Ithaca*. Published online January 30, 2017. Accessed November 20, 2019. http://search.proquest.com/docview/2075396516?rfr_id=info%3Axri%2Fsid%3Aprimo

35. Izmailov P, Podoprikhin D, Garipov T, Vetrov D, Wilson AG. *Averaging Weights Leads to Wider Optima and Better Generalization*.; 2018.

36. Devries T, Taylor GW. Improved Regularization of Convolutional Neural Networks with Cutout. *CoRR*. 2017;abs/1708.04552. http://arxiv.org/abs/1708.04552

37. Xie R, Liu L, Liu J, Qiu CS. Pathological Myopic Image Analysis with Transfer Learning. In: ; 2019. Accessed October 3, 2020. https://openreview.net/forum?id=BkeLp6mTFE

38. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv14091556 Cs*. Published online April 10, 2015. Accessed May 15, 2020. http://arxiv.org/abs/1409.1556

39. Babu SC, Maiya SR, Elango S. *Relation Networks for Optic Disc and Fovea Localization in Retinal Images*.; 2018.

40. Orlando JI, Fu H, Barbossa Breda J, et al. REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Med Image Anal*. 2020;59:101570. doi:10.1016/j.media.2019.101570

41. Wu R-Y, Wong T-Y, Zheng Y-F, et al. Influence of Refractive Error on Optic Disc Topographic Parameters: The Singapore Malay Eye Study. *Am J Ophthalmol*. 2011;152(1):81-86. doi:10.1016/j.ajo.2011.01.018

42. Ramrattan RS, Wolfs RCW, Jonas JB, Hofman A, Jong PTVM de. Determinants of optic disc characteristics in a general population: The Rotterdam study1. *Ophthalmology*. 1999;106(8):1588-1596. doi:10.1016/S0161-6420(99)90457-8

43. Dai Y, Jonas JB, Huang H, Wang M, Sun X. Microstructure of Parapapillary Atrophy: Beta Zone and Gamma Zone. *Invest Ophthalmol Vis Sci*. 2013;54(3):2013-2018. doi:10.1167/iovs.12-11255

44. Vianna JR, Malik R, Danthurebandara VM, et al. Beta and Gamma Peripapillary Atrophy in Myopic Eyes With and Without Glaucoma. *Invest Ophthalmol Vis Sci*. 2016;57(7):3103-3111. doi:10.1167/iovs.16-19646

45. Song MK, Sung KR, Shin JW, Kwon J, Lee JY, Park JM. Progressive change in peripapillary atrophy in myopic glaucomatous eyes. *Br J Ophthalmol*. 2018;102(11):1527-1532. doi:10.1136/bjophthalmol-2017-311152

46. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, eds. *Computer Vision – ECCV 2018*. Vol 11211. Lecture Notes in Computer Science. Springer International Publishing; 2018:833-851. doi:10.1007/978-3-030-01234-2_49

47. Guo Y, Wang R, Zhou X, et al. Lesion-Aware Segmentation Network for Atrophy and Detachment of Pathological Myopia on Fundus Images. In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. ; 2020:1242-1245. doi:10.1109/ISBI45749.2020.9098669

Figures

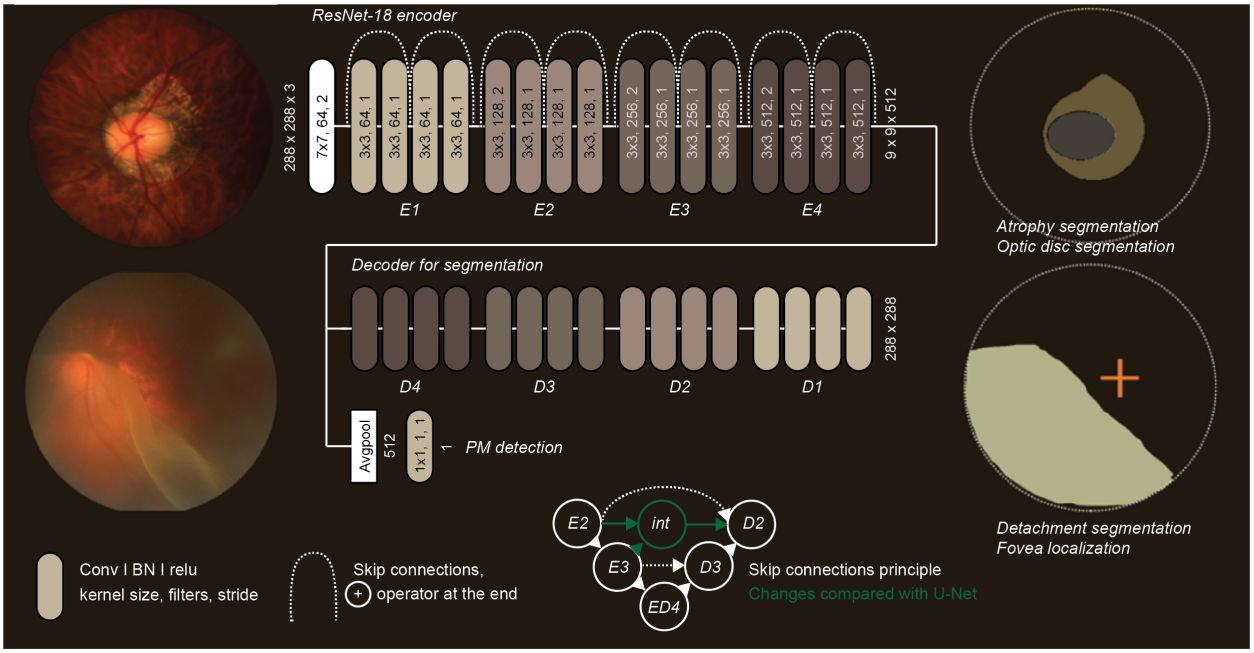

Figure 1: Overview of the final model architecture used for inference on the PALM official validation and test set. Our model is aimed at PM classification with simultaneous segmentation of ONH and retinal atrophy. The ResNet encoder accepts resized fundus images of (288 x 288) and outputs (9 x 9 x 512) at the latent space. The decoder upscales this output back to the original image size, using a plethora of skip connections (principle given in bottom center). The graphic on the upper right represents the generated segmentation map of the ONH (grey) and retinal atrophy (olive). The output of the encoder is also separately transformed to a single prediction for PM classification (through average pooling and a convolution operation). The model for fovea localization employs a similar architecture as for ONH/atrophy segmentation, but generates a circle. This circle is then transformed to coordinates using its centroid (visualized by the orange cross on the right bottom segmentation map). Finally, the UNet++ model for segmentation of retinal detachment is identical to the other models, but outputs detachment.

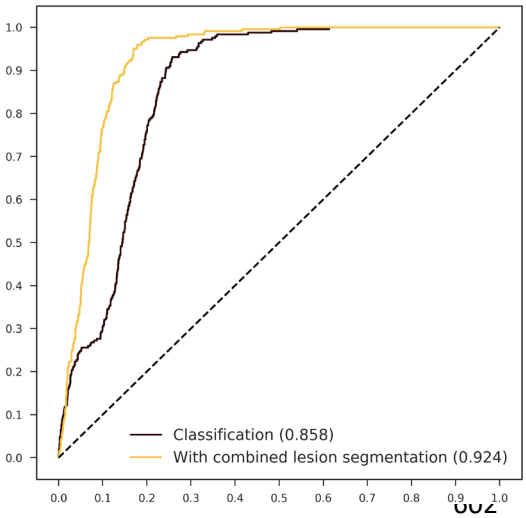

Figure 2: ROC curves of models trained on PALM data, evaluated on 3350 images of ODIR. The model with combined lesion segmentation significantly outperforms the classification-only model.

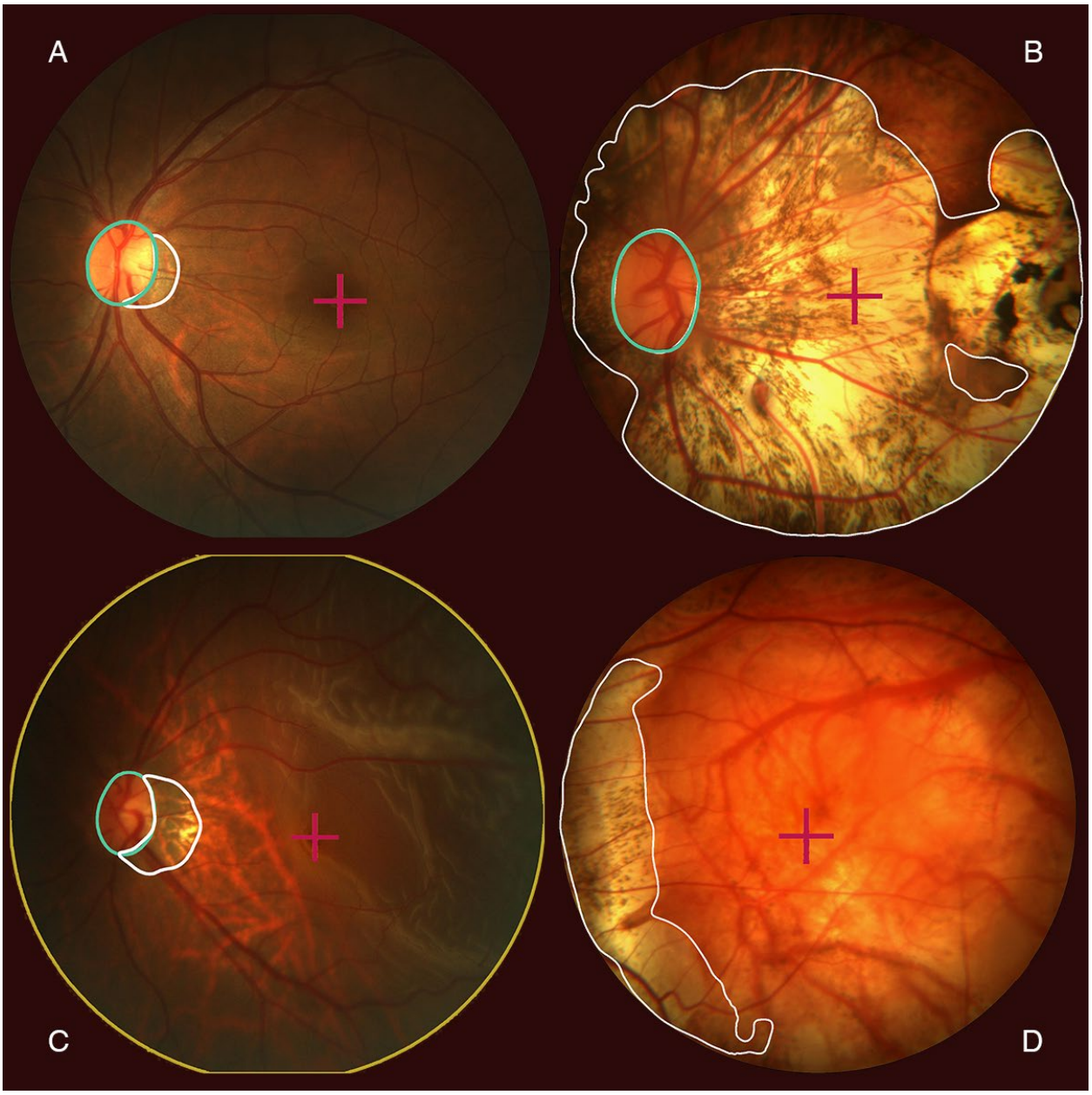

Figure 3: Qualitative results giving four cases of the official test set. The optic nerve head (outlined in green) is detected and segmented in A, B and C. Retinal atrophy is detected and segmented (outlined in white) in B and C. Retinal detachment was detected in C, for which the whole fundus is outlined in yellow. Finally, the fovea is localized in all cases, indicated by a purple cross. Image D does not feature an optic disc, but clear retinal atrophy on the left.

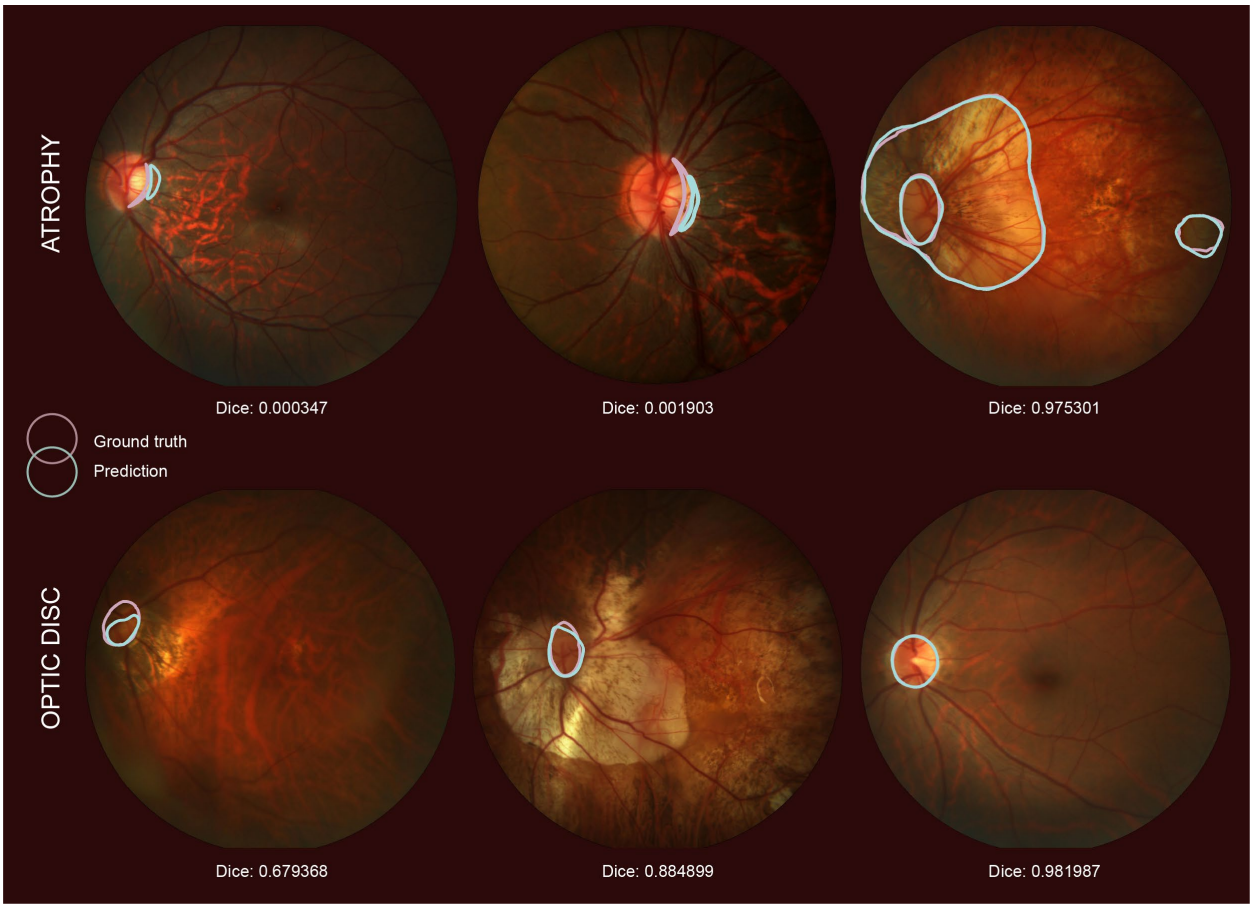

Figure 4: Selected samples of atrophy (top row) and optic disc (bottom row) segmentations. Per row, the images with lowest Dice scores on the holdout set of 40 images are visualized (left and middle column), complemented with the image for which the best prediction was obtained (far right).

Tables

*Table 1: Overview of characteristics of labeled training set of 400 images. Significance level between PM and Non-PM on same camera settings provided with asterisks (where applicable, * <0.05, ** <0.01, *** <0.001, **** <0.0001).*

Centering	Macula				Disc			
Angle	30°		45°		30°		45°	
	PM	Non-PM	PM	Non-Pm	PM	Non-PM	PM	Non-PM
Number of images	6	4	174	173	31	9	2	1
Images with full optic disc	0%	0%	94.8%	100%	100%	100%	100%	100%
Images with atrophy	100%	75%	98.3%	52.6%	100%	77.8%	100%	0%
Images with fovea	100%	100%	99.4%	100%	96.8%	88.9%	100%	100%
Optic disc area	-	-	1.66%	1.72%	3.38%	2.61%**	1.69%	1.15%
Atrophy area	5.93%	0.41%*	11.77%	0.25%****	13.97%	0.70%****	42.37%	-
Fovea x mean	768	758	1236	1102****	1261	1387*	1748	1792
Fovea y mean	713	741	1026	1081****	754	715	1144	1049

Table 2: Results for five tasks, obtained on holdout validation set (PALM holdout), the official PALM test set, and external data sets when available (ODIR and Messidor). PM detection is measured in AUC, fovea localization in Euclidean distance. Dice and F1 are given for the three segmentation tasks (ONH, atrophy, detachment).

PM detection (AUC)	PALM holdout (n=40)	PALM test set (n=400)	ODIR (n=3350)
Classification	1 (loss: 0.1446)	-	0.8584
Classification combined with ONH/atrophy segmentation	1 (loss: 0.0824)	0.9867	0.9245
Fovea localization (Euclidean dist)	PALM holdout (n=40)	PALM test set (n=400)	Messidor (n=1136)
Regression	229.428	-	53.488
Segmentation, radius 25 pixels	129.182	-	25.765
Segmentation, radius 75 pixels	109.770	-	20.220
Segmentation, combined with ONH	86.675	-	18.296
Segmentation, combined with ONH, postprocessing	61.924	58.3	-
ONH segmentation	PALM holdout (n=40)	PALM test set (n=400)	
Metric	Dice	Dice	F1
Segmentation	0.9481	0.9303	0.9869
Segmentation combined with atrophy	0.9462	-	-
Segmentation combined with atrophy, Lovász loss	0.9414	-	-
Atrophy segmentation	PALM holdout (n=40)	PALM test set (n=400)	
Metric	Dice	Dice	F1
Segmentation	0.6210	-	-
Segmentation combined with atrophy	0.6810	-	-
Segmentation combined with atrophy, Lovász loss	0.6948	0.8001	0.9135
Detachment segmentation	PALM holdout (n=40)	PALM test set (n=400)	
Metric	Dice	Dice	F1
Segmentation	0.9500	-	-
Segmentation with a balanced data generator	0.9998	0.8073	0.7059

Supplementary material

Tables 3-7: Results for five tasks, obtained on holdout validation set (PALM holdout), the official PALM test set, and external data sets when available (ODIR and Messidor). Team A-F correspond to Vistalab, Masker, LAIS, PingAn Smart Heatlh, CUHK, and RYE-NUS, respectively.

PM detection	PALM holdout (n=40)	PALM test set (n=400)	ODIR (n=3350)
Classification	1 (loss: 0.1446)	-	0.8584
Classification combined with ONH/atrophy segmentation	1 (loss: 0.0824)	0.9867	0.9245
Team A	-	0.9974	-
Team B	-	0.9960	-
Team C	-	0.9957	-
Team D	-	0.9934	-
Team E	-	-	-
Team F	-	-	-

Fovea localization	PALM holdout (n=40)	PALM test set (n=400)	Messidor (n=1136)
Regression	229.428	-	53.488
Segmentation, radius 25 pixels	129.182	-	25.765
Segmentation, radius 75 pixels	109.770	-	20.220
Segmentation, combined with ONH	86.675	-	18.296
Segmentation, combined with ONH, postprocessing	61.924	58.3	-
Team A	-	55.7	-
Team B	-	172.9	-
Team C	-	71.3	-
Team D	-	66.6	-
Team E	-	-	-
Team F	-	-	-

ONH segmentation	PALM holdout (n=40)	PALM test set (n=400)	
Metric	Dice	Dice	F1
Segmentation	0.9481	0.9303	0.9869
Segmentation combined with atrophy	0.9462	-	-
Segmentation combined with atrophy, Lovász loss	0.9414	-	-
Team A	-	0.9362	0.9909
Team B	-	0.9367	0.9806
Team C	-	0.9093	0.9855
Team D	-	0.9508	0.9974
Team E	-	-	-
Team F	-	0.9288	0.9871

Atrophy segmentation	PALM holdout (n=40)	PALM test set (n=400)	
Metric	Dice	Dice	F1
Segmentation	0.6210	-	-
Segmentation combined with atrophy	0.6810	-	-
Segmentation combined with atrophy, Lovász loss	0.6948	0.8001	0.9135
Team A	-	0.7879	0.8972
Team B	-	0.7702	0.8372
Team C	-	0.7798	0.9091
Team D	-	0.8220	0.9303
Team E	-	0.8183	0.9199
Team F	-	-	-

Detachment segmentation	PALM holdout (n=40)	PALM test set (n=400)	
Metric	Dice	Dice	F1
Segmentation	0.9500	-	-
Segmentation with a balanced data generator	0.9998	0.8073	0.7059
Team A	-	0.1584	0.1667
Team B	-	0.0030	0.0541
Team C	-	0.5546	0.7273
Team D	-	0.6617	0.9091
Team E	-	0.7449	0.8571
Team F	-	-	-