AUTOMATED DETECTION OF GLAUCOMA AND DIAGNOSTIC FEATURES FOR JUSTRAIGS CHALLENGE

Tomasz Kubrak

Tilburg University, The Netherlands

ABSTRACT

Glaucoma encompasses a set of ocular conditions that jeopardize vision and can lead to blindness by harming the optic nerve at the rear of the eye. While early detection is crucial to halt further vision impairment, it is often challenging due to the scarcity of medical experts. To mitigate this limitation, various deep learning architectures have been developed. However, due to their importance, further improvement of these models is crucial. They also tend to operate as "black boxes", which makes their decision process opaque. To address these issues, our work not only establishes a new state-of-the-art (SOTA) glaucoma classification pipeline on JustRAIGS dataset but also detects ten additional features commonly used by ophthalmologists to explain their diagnostic decisions.

Index Terms— Glaucoma Detection, Diagnostic Features Classification, JustRAIGS Challenge

1. INTRODUCTION

Detecting glaucoma through color fundus photographs (CFPs) remains a widely accessible method, yet it depends on relatively costly fundus cameras and the expertise of medical professionals for image analysis. To decrease the costs of this process, various artificial intelligence (AI) architectures have been developed to automate image classification, thus reducing the need for expert analysis. However, despite their advantages these AI systems predominantly operate as “black box” frameworks, making their decision-making process impossible to understand. Furthermore, although current models already outperform ophthalmologists, due to their importance, enhancing their performance remains vital.

In response to these challenges, this research proposes a state-of-the-art, deep learning pipeline to classify glaucoma alongside with ten diagnostic features. Inspired by the winning architecture of the AIROGS challenge [1] (previous challenge on this dataset without additional features), the pipeline employs significant improvements. It substitutes ResNet101-UperNet with YOLOv8 for optic disc (OD) segmentation, maintains regions of interest (ROI) crop in diagnostic pipeline, incorporates CLAHE contrast enhancement [2] and removes black edges of CFP without center cropping. Notably, the input image size has been increased to 512x512, and the model now benefits from weighted loss functions, weight decay, and differential learning rates during its training phase.

The pipeline has been trained on a novel dataset from the JustRAIGS challenge [3]. This dataset comprises CFPs of 113,893 eyes labeled for the presence of glaucoma as well as ten extra features that explain classification decisions of ophthalmologists. By advancing the interpretability and accuracy of AI diagnostics, this research aims to bridge the gap between AI tools and clinical usability, ensuring more reliable and accessible glaucoma detection.

2. DATASET DESCRIPTION

The JustRAIGS challenge dataset [3] contains 113,893 labeled color fundus photographs of eyes from 60,357 individuals collected during the EyePACS diabetic retinopathy screening program. To ensure variability in the data, around 500 screening centers have participated in the program, utilizing a large set of fundus cameras. To provide the highest quality of labels, a careful selection of graders was conducted. Several hundred European ophthalmologists were trained using the European Optic Disc Assessment Trial (EODAT). After EODAT, the top three graders were selected to annotate all images, and a consensus between their judgments was made. They had to choose between "Referable glaucoma" (RG), "No referable glaucoma" (NRG), or "Ungradable" (UG). In the final consensus, 111,183 images were labeled as gradable (RG + NRG), of which 4.38% were considered as RG. If RG was selected, ophthalmologists had to choose between ten typical glaucomatous features of the eye, namely: ANRI (appearance of the neuroretinal rim inferiorly); ANRS (appearance of the neuroretinal rim superiorly); BCLVI (baring of the circumlinear vessel(s) inferiorly); BCLVS (baring of the circumlinear vessel(s) superiorly); DH (disc hemorrhages); LC (large cup); LD (laminar dots); NVT (nasalization (nasal displacement) of the vessel trunk); RNFLDI (retinal nerve fiber layer defect inferiorly); RNFLDS (retinal nerve fiber layer defect superiorly).

3. METHODS

The core of our classification pipeline comprises four Vision Transformers (ViTs) [4], each pre-trained on ImageNet and fine-tuned on different subsets of the dataset (Table 1, Fig. 1). They operate with the same hyperparameters and preprocessing steps, except for the region of interests cropping component, which is contingent on the optic disc segmentation results (Table 1, Fig. 1).

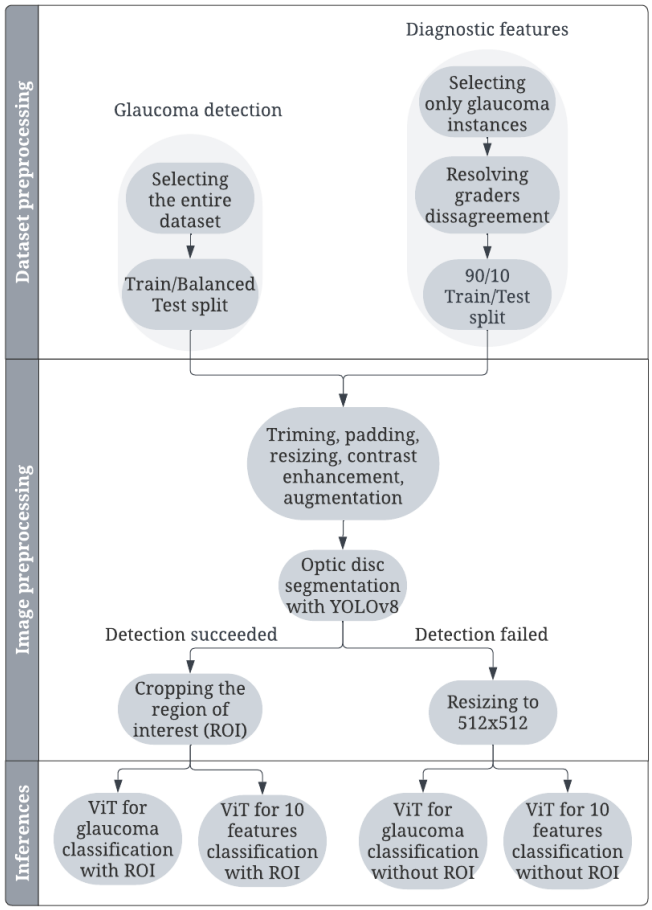

Fig. 1 Pipeline of glaucoma detection and diagnostic features classification.

3.1. Dataset preprocessing

For training of each ViT, a different subset of the dataset was used. To classify the diagnostic features, we only used these images where RG was identified (3,270 images). This subset was then split into a training and test set, reflecting a 90/10 split (Table 1, Fig. 1). The entire dataset was employed for the detection of glaucoma. It was split into two parts: a training set comprising 100,769 images (2,943 RG and 97,827 NRG instances), and a balanced test set consisting of 654 images (327 images per class) (Table 1). For training and testing of these ViTs that were relying on ROI preprocessing, images without successful OD detection were discarded (Table 1).

Three graders were involved in labeling the dataset. The first two graders independently annotated all images. In cases where these two graders disagreed on the presence of glaucoma, the third grader intervened to resolve the conflict. However, this conflict resolution process was not applied when labeling the additional features.

In cases where the initial two graders disagreed on any additional features and the annotations from a third grader were not provided, we defaulted to the most frequently observed label for the contested feature within the dataset. As a result, when labels for ANRI, ANRS, and LC were missing, they were assigned as present. For any other unmarked labels among the 10 features, we classified them as absent (Fig. 1).

Table 1. Training settings and hyperparameters of the ViTs.

	Glaucoma detection		**Features detection**	
ROI cropping	Yes	No	Yes	No
Train set size	100,420	100,769	2,935	2,943
Test set size	653	654	326	327
Batch size	20		11	
Learning rates	Network: 1.0e-4, Head: 1.0e-5.			
Loss function	Weighted cross-entropy with class inverse distribution weights.			
Weight decay	1.0e-4			

3.2. Image preprocessing

Before being processed for optic disc segmentation (Fig. 2a), each image was trimmed of excessive black space and then padded to maintain a 1:1 aspect ratio between image dimensions. Furthermore, it was resized to 2,000x2,000 pixels, and CLAHE [2] contrast enhancement was applied (Fig. 2b). If the ViT was using ROI cropping as preprocessing step, YOLOv8 was run to detect OD (Fig. 1). In the case of YOLOv8 detecting multiple bounding boxes, the encompassing bounding box was selected. If this bounding box exceeded the size of the ROI (518x518), segmentation was discarded. Finally, a square region of interest of size 518x518 was cropped around the center of the bounding box (Fig. 2c). If the desired ROI exceeded the image dimensions, the largest fitting ROI was chosen. As a final step, all images were resized to 512x512, normalized, and augmented with AutoAugment.

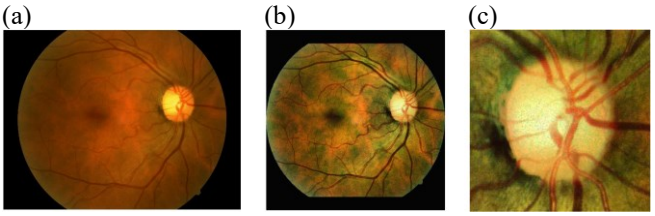

Fig. 2. (a) CFP before preprocessing. (b) CFP after trimming, padding, resizing and CLAHE contrast enhancement. (c) CFP after cropping ROI.

3.3. Optic disc segmentation

We selected YOLOv8 as our segmentation model. The JustRAIGS dataset [3] did not include masks of the OD, therefore manual labeling was needed. For initial training of YOLOv8, we generated 1,000 masks using a U-Net (Encoder-Decoder Network) designed for OD segmentation tasks [5]. Although these masks were only partially accurate, they were sufficient for the preliminary training phase of YOLOv8. Following this initial phase, YOLOv8 was deployed to generate OD masks for the entire dataset, achieving a successful detection rate of about 90%. Approximately 90,000 detected images were then utilized for YOLOv8's last training. The final model successfully segmented ROI in roughly 99.6% of all the images.

3.4. Vision Transformers

As a backbone model for all classifications, a large ViT with a patch size of 16, pre-trained on ImageNet, was chosen [4]. In total, four ViTs were fine-tuned depending on the successfulness of ROI segmentation and the purpose of the task (Fig.1, Table 1). Although each of them used different parts of the dataset, due to task similarity, they shared the same hyperparameters as described in Table 1. To decrease training time, mixed precision was applied.

3.5. Evaluation metrics

In the JustRAIGS challenge, two main metrics are evaluated: sensitivity at 95% specificity for glaucoma detection and modified Hamming distance for extra features classification. For additional insight, we also evaluated the accuracy of models used for detecting glaucoma.

4. RESULTS

Table 2 summarizes the results obtained from this internal test set (Table 1) and from the development phase leaderboard of the competition. During the development phase, the entire pipeline was evaluated collectively (Fig. 1). Consequently, by default, the ViT with ROI cropping was employed. If the optic disc was not detected by YOLOv8, the system then ran inferences using the ViT without ROI preprocessing (Fig. 1).

Table 2. Evaluation metrics with and without ROI cropping across test and development set.

	Test set		**Development set**
ROI cropping	Yes	No	Yes, if OD detected
Hamming distance	0.1414	0.1779	0.1280
Sensitivity at 95% specificity	0.9325	0.8930	0.9090
Accuracy	0.9449	0.9235	

5. DISCUSSION AND CONCLUSION

This work presents a novel approach for the classification of recently released diagnostic features, as well as for glaucoma detection, for the JustRAIGS challenge. Our pipeline achieved second place across all teams during the development phase of competition and demonstrated state-of-the-art performance for glaucoma detection. At the time of writing, the performance of the pipeline on the test phase of the challenge remains unknown.

6. ACKNOWLEDGMENTS

This research was conducted retrospectively using human subject data made available in open access by EyePACS LLC, Santa Cruz, CA, USA. Labels were provided by the Rotterdam Ophthalmic Institute, Rotterdam Eye Hospital, Rotterdam, The Netherlands. Ethical approval was not required as confirmed by the license attached with the open access data.

No funding was received for conducting this study. The author has no relevant financial or non-financial interests to disclose.

7. REFERENCES

[1] C. De Vente et al., "AIROGS: Artificial Intelligence for ROBUST Glaucoma Screening Challenge," IEEE Transactions on Medical Imaging, vol. 43, no. 1, pp. 542–557, Jan. 2024, doi: 10.1109/tmi.2023.3313786.

[2] S. M. Pizer et al., "Adaptive histogram equalization and its variations," Computer Vision, Graphics, and Image Processing, vol. 39, no. 3, pp. 355–368, Sep. 1987, doi: 10.1016/s0734-189x(87)80186-x.

[3] H. G. Lemij, C. De Vente, C. I. SáNchez, and K. A. Vermeer, "Characteristics of a Large, Labeled Data Set for the Training of Artificial Intelligence for Glaucoma Screening with Fundus Photographs," Ophthalmology Science, vol. 3, no. 3, p. 100300, Sep. 2023, doi: 10.1016/j.xops.2023.100300.

[4] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv (Cornell University), Jan. 2020, doi: 10.48550/arxiv.2010.11929.

[5] Paresh-Shahare, "GitHub - Paresh-Shahare/Optic-Disc-Segmentation," GitHub. https://github.com/Paresh-shahare/Optic-Disc-Segmentation [Accessed: 29 April 2024].