npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-02318-2

Development of deep learning model to screen for primary open-angle glaucoma in African ancestry individuals

Shuo Li[1,4], Rebecca Salowe[2,4], Roy Lee[2], Gui-shuang Ying[3], Insup Lee[1], Joan O'Brien[2] ✉ & Osbert Bastani[1]

Primary open-angle glaucoma (POAG) screening using artificial intelligence (AI) has emerged as a transformative method to identify undiagnosed disease. African ancestry individuals are underrepresented in current datasets for AI models, despite being disproportionally affected by this blinding disease. We developed a deep learning model that screens for POAG using fundus photography from Primary Open-Angle African American Glaucoma Genetics (POAAGG) subjects (n = 64,129 images, including 42,914 images from 1782 cases and 21,215 images from 682 controls). Our final diagnosis pipeline is as follows: (1) select the six most informative images from single timepoint using a Binary Classifier, (2) predict POAG probability from each image using Vision-Transformer, (3) make final POAG predictions by averaging predicted probabilities across selected images (AUC = 0.925). The model was evaluated on the REFUGE-1 dataset of Chinese ancestry individuals (AUC = 0.920). Our model has applications to POAG screening in public settings such as primary care offices, as well as low-resource settings.

Artificial intelligence (AI) is increasingly being utilized as a screening tool for ophthalmic diseases[1]. Some eye diseases, such as diabetic retinopathy, now have several AI-enabled screening programs approved by the Food and Drug Administration (FDA) for usage in primary care clinics[2]. AI can even predict cardiovascular risk factors on images of the optic nerve, such as age, gender, smoking status, and major adverse cardiac events[3]. Despite this exciting progress, the development of AI for glaucoma has advanced at a slower pace and has not yet been fully integrated into screening or clinical workflows.

There is a growing need for AI that accurately screens for primary open-angle glaucoma (POAG)—the most common form of the disease—across diverse populations, expanding access to care without burdening the healthcare system[4]. Early detection and treatment of POAG are imperative as vision loss from the disease is irreversible[5]. Almost half of patients are unaware they have POAG because the disease is asymptomatic in early stages and typically begins with peripheral vision loss[6,7]. POAG screenings thus offer a powerful mechanism to detect early signs of disease in undiagnosed individuals, allowing referral and intervention while there is still vision left to preserve[1]. Currently, the scope and reach of these screenings are limited by the reliance on in-person examinations by eye care professionals. These screenings can be lengthy, labor-intensive, and challenging to practically implement, limiting the number of screened individuals. In developing countries, this challenge is amplified, as there is a high burden of POAG and a very limited number of trained eye professionals[8].

AI that accurately flags POAG on images holds enormous promise to increase access to screenings in an aging population[9]. Such a system could be used to detect and refer possible POAG in real time in settings without an ophthalmologist, such as primary care offices or low-resource settings. Importantly, this system can serve frontline eye care in remote rural areas or countries with a scarcity of ophthalmologists. To this end, several groups have developed deep learning models to screen for POAG, with most approaches utilizing fundus images. Fundus photography, which provides visualization of anatomic changes to the optic nerve head, is ideal for screening as it is low-cost, non-invasive, quick, and portable[10]. Ting et al. developed a deep learning model for diabetic retinopathy in a multiethnic population with diabetes (N = 494,661 images) that successfully detected "referable" glaucoma; however, the 13% false-positive rate could lead to unnecessary referrals[11]. Similarly, Li et al. developed a deep learning algorithm in a Chinese population (N = 48,116 images), showing success in detecting glaucomatous optic neuropathy[12]. The Pegasus system (Visulytix Ltd, London, UK), a cloud-based AI system that evaluates fundus photos

[1]Department of Computer & Information Science, University of Pennsylvania, Philadelphia, PA, USA. [2]Center for Genetics of Complex Disease, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. [3]Center for Preventive Ophthalmology and Biostatistics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA. [4]These authors contributed equally: Shuo Li, Rebecca Salowe.
✉e-mail: joan.obrien@pennmedicine.upenn.edu

using a collection of convolutional neural networks (CNNs), had an overall accuracy of 83.4% in diagnosing glaucoma when compared to eye care professionals[13].

Though promising, a major challenge faced by these deep learning models (and many AI models for other diseases) is the low representation of Black individuals in training and validation datasets. For a disease such as POAG, this likely stems from the historic under-representation of Black individuals in research studies—and resultant lack of large datasets of images from this ancestry group. African ancestry individuals are more frequently and severely affected by POAG and thus must be represented in datasets used to train deep learning algorithms[14–16]. A lack of diversity in training datasets can result in worse performance of the AI model in other populations, limiting generalizability and exacerbating existing healthcare disparities[17].

In this study, we developed a deep learning model that achieves high accuracy in predicting POAG from fundus photography in a large African ancestry population. We used images from the Primary Open-Angle African American Glaucoma Genetics (POAAGG) study, a large glaucoma genetics study conducted at University of Pennsylvania (UPenn)[18,19]. We compared results from different machine learning models, as well as heuristics to select informative images, to develop a final pipeline that achieves high performance.

Results

A total of 64,129 fundus images from 2464 POAAGG subjects were included in the study, including 42,914 images from 1782 glaucoma cases and 21,215 images from 682 controls. The mean age (± standard deviation) at enrollment of cases and controls was 69.34 and 62.52, respectively, with similar splits across training, validation, and testing datasets (Supplementary Table 1). Cases were 58.6% female, while controls were 68.6% female. The final diagnosis pipeline is shown in Fig. 1.

ResNet and ViT obtain similar performance

First, we made predictions of POAG on individual images in the POAAGG and REFUGE-1 datasets using two state-of-the-art machine learning models: ResNet and Vision-Transformer (ViT). In Table 1, we show the Area Under the Receiver Operating Characteristic (AUC) with 95% confidence interval. We found that ResNet and ViT demonstrated comparable performance. However, ViT offered enhanced interpretability of its predictions through its attention map, which provides a visualization of the parts of the input that the model focuses on when making a prediction. The trained model using ViT pays more attention to the cup/disc region in the original fundus image—the area affected by POAG pathogenesis—and thus can achieve superior performance when integrated with our proposed pipeline (Supplementary Figure 1). Consequently, we selected ViT for our POAG screening pipeline. However, we acknowledge that interpretable machine learning techniques such as Grad-CAM[20] can be employed to enhance the interpretability of CNNs. While our current focus in this paper is analyzing attention maps produced by ViTs, a systematic comparison with interpretability methods applied to CNNs remains an important direction for future work.

Informative image selection improves performance

Next, we assessed if selecting informative images improved model performance. At each subject appointment, a series of images was taken to account for blinking and other confounding factors. We selected informative images from this series using two different methods: a segmentation-based model (Seg) and a Binary Classifier (Binary). After selecting informative images, we made predictions on individual images (i.e., not yet aggregated) and calculated AUCs (Fig. 2).

Compared to including all images (All) (AUC = 0.82), selecting informative images using Seg or Binary significantly improved performance. Specifically, Binary outperformed Seg (AUC = 0.90 versus AUC = 0.83, respectively) and achieved performance on par with images selected by non-physician graders from the Scheie Image Reading Center (Expert) (AUC = 0.90). This demonstrated the capability of the trained Binary Classifier in identifying images suitable for detecting glaucoma. We also showed that selection of six informative images was optimal for our pipeline.

Averaging probabilities from selected images further improves performance

Subsequently, we aggregated predictions on Binary Classifier-selected images from the same timepoint using two different methods: *average* predicted confidence and *maximum* predicted confidence (Fig. 3). Compared to the results in Fig. 2, aggregating the predictions significantly improved the AUCs of both baselines (All and Expert) and our chosen selection method (Binary). Furthermore, we found that aggregating predictions by *averaging* the predicted confidences yielded the best performance (AUC = 0.925), compared to using the maximum predicted probability (AUC = 0.914).

Dataset size improves transferability

In Table 2, we present the AUCs (with 95% confidence interval) of models trained and tested on different datasets. For each random seed, we partitioned the dataset into training, validation, and test sets, with ratios of 0.8, 0.1, and 0.1, respectively. Firstly, comparing the performance between models trained on POAAGG (first row) and REFUGE (third row) reveals that the model trained on POAAGG exhibited better cross-ancestry transferability than the model trained on REFUGE. This discrepancy could be attributed to at least two factors: differences in ancestry and dataset size. Additional evaluation metrics are shown in Supplementary Table 2.

'While the first factor is challenging to quantify, we investigated the impact of dataset size on transferability. Specifically, by comparing the second row to the third row, we observed that when the size of POAAGG is reduced to match that of REFUGE, the transferability of the model trained on the smaller dataset was similarly poor compared to the model trained on REFUGE. Consequently, we concluded that dataset size is a critical factor in determining the transferability of models across different ancestries.

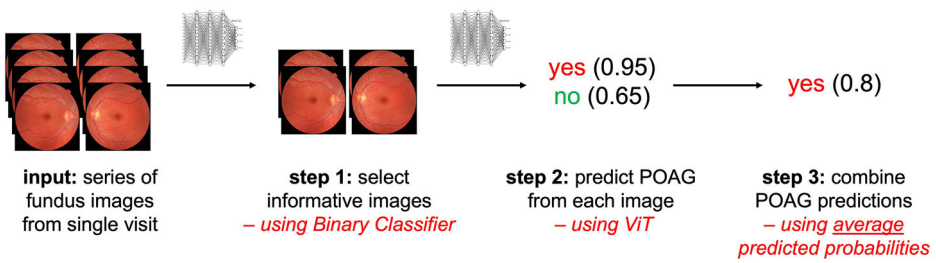

Fig. 1 | Final diagnosis pipeline. *Input*: Series of fundus images taken during a single visit for subjects enrolled in the Primary Open-Angle African American Glaucoma Genetics (POAAGG) study (N = 64,129 images, including 42,914 case and 21,215 control images). *Step 1*: Select the six most informative images from single timepoint using a Binary Classifier (in the visualization, only two images are selected). *Step 2*: Predict probability of POAG from each individual fundus image, using Vision-Transformer (ViT). *Step 3*: Make final POAG predictions, using the average of predicted probabilities across selected images.

Table 1 | AUCs (standard deviation [SD]) from ResNet101 and ViT when tested in the POAAGG and REFUGE datasets

	POAAGG	REFUGE
ResNet	0.831 +/- 0.009	0.920 +/- 0.004
ViT	0.830 +/- 0.009	0.917 +/- 0.006

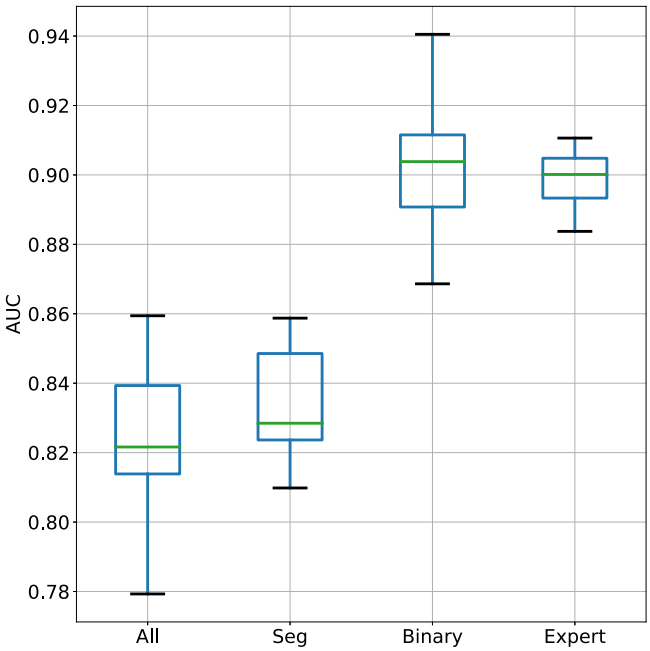

Fig. 2 | Evaluation of model performance when using informative images. We compared the median AUCs when using all images/timepoint versus individual informative images, which were selected via the segmentation-based model (Seg), binary classifier (Binary), and expert-selected images (Expert).

Discussion

In this paper, we developed a deep learning model that accurately detects the presence of POAG on fundus images from a large African ancestry population. We tested multiple pipelines in order to achieve the highest possible performance. In our final pipeline, we selected the six most informative images for per subject from a single timepoint using a Binary Classifier; used ViT to predict POAG from each image; and averaged the predicted probabilities across selected images to make final POAG predictions. We also tested our model on a Chinese ancestry population and demonstrated that dataset size can critically affect generalizability across different ancestry groups.

There is an unmet need for AI that can accurately detect the presence of POAG on images[9]. Not only is the prevalence of POAG continuing to grow in an aging population, but more than half of POAG patients are unaware that they have the disease or are sub-optimally managed, especially in developing countires[8,21]. An autonomous screening system could be applied in low-resource areas, community outreach events, or primary care offices, allowing the detection and referral of patients with early signs of POAG, before irreversible vision loss has occurred. This will be especially impactful for individuals who live in remote rural areas or underserved areas, with the added benefit of reducing the cost of healthcare by decreasing the numbers of late-stage glaucoma patients[8]. If referrals to ophthalmologists are not possible in such locations, the relevant collected diagnostic data from flagged patients can be sent via encrypted internet connection to an eye care specialist located remotely, or a specialist can provide real-time consultations via an online platform[22]. To achieve such screening purposes, fundus photography is an ideal imaging modality, as it is portable, inexpensive, and quick, allowing easy application in such settings in the future[10]. For this reason, we used fundus images for our AI pipeline in this study.

Though AI for POAG has seen advances in recent years, progress still lags behind other ocular diseases, especially diabetic retinopathy. As of July 2024, there were three algorithms cleared by the FDA for clinical use in diabetic retinopathy screening (IDx-DR, EyeARt, AEYE-DS)[23]. The latter algorithm, from AEYE Health, allows diagnosis of referable diabetic retinopathy using a handheld camera to take retinal images. Several groups have shown exciting progress in developing AI models for POAG screening and diagnosis in recent years, but no AI software is available for real clinical application[11–13,24–29]. This may be partially due to challenges specific to this disease, such as the lack of consistent and objective diagnostic criteria and limited availability of images in public databases in terms of both quantity and diversity[30].

One of our main contributions is the design and evaluation of a novel pipeline for glaucoma screening that includes high-quality image selection. To fully automate glaucoma screening pipelines, the selection of high-quality images must also be automated. We investigated the optimal number of informative images and compared methods of selection and aggregation to determine which process led to the best performance. Ultimately, we found that selecting six informative images using the Binary Classifier, and aggregating the predictions by averaging the predicted confidences, produced the highest AUC. We showed that this selection of informative images outperformed the inclusion of all available images (despite the higher N), and performed on par with inclusion of "expert" images, which were manually chosen by Reading Center graders, a very expensive and time-consuming process. This improvement can be attributed to the selection of high-quality images (i.e., no patient blinking, no confounding factors, etc.), which enabled the trained models to make more accurate predictions. Additionally, the better performance of average versus maximum for aggregation of probabilities is likely due to the reduction of noise in the predicted confidences, which are often poorly calibrated.

Another major strength of this study is the inclusion of images from a large cohort of African ancestry individuals. Most POAG studies have been conducted on European and Asian ancestry individuals, with low representation of African ancestry individuals, leading to a lack of large image datasets for use in AI models. African ancestry individuals are disproportionately affected by POAG and up to 15 times more likely to experience vision loss from the disease compared to European Americans[15,16], so this population is essential to include in training datasets for AI models. A lack of diversity in training datasets—seen in many AI models for diseases beyond POAG—has been proven to lead to worse performance of these models in new data from another ancestry group[31]. Studies show that AI models reflect biases from the training dataset, with variables such as optic disc size or fundus pigmentation affecting accuracy[32]. In this study, we assessed how our trained model transferred to a Chinese ancestry population (REFUGE dataset). Our model exhibited cross-ancestry transferability on the REFUGE dataset, though performance may have been hindered by the small number of positive examples in REFUGE. Additionally, the reverse process (training on REFUGE, evaluating in POAAGG) did not show the same strong performance. Further testing with a "small" POAAGG dataset (reduced to be the same size as REFUGE) exhibited similar poor transferability, leading us to conclude that dataset size is a critical factor in determining the transferability of models across different ancestries. This further reinforces the need for large datasets of diverse populations, in order to facilitate better transferability and to ensure that advances AI benefit all patients, especially the most affected population.

There are several limitations to our study. First, our training dataset included images from solely African ancestry individuals and thus may reflect biases from the training dataset, potentially leading to worse performance in new data from other ancestry groups. Though the inclusion of an African ancestry population is also a strength, and we showed that our model translated well to a Chinese ancestry population, further testing is required to ensure that our model transfers successfully to other ancestry groups, such as the Los Angeles Latino Eye Study. Variables such as fundus

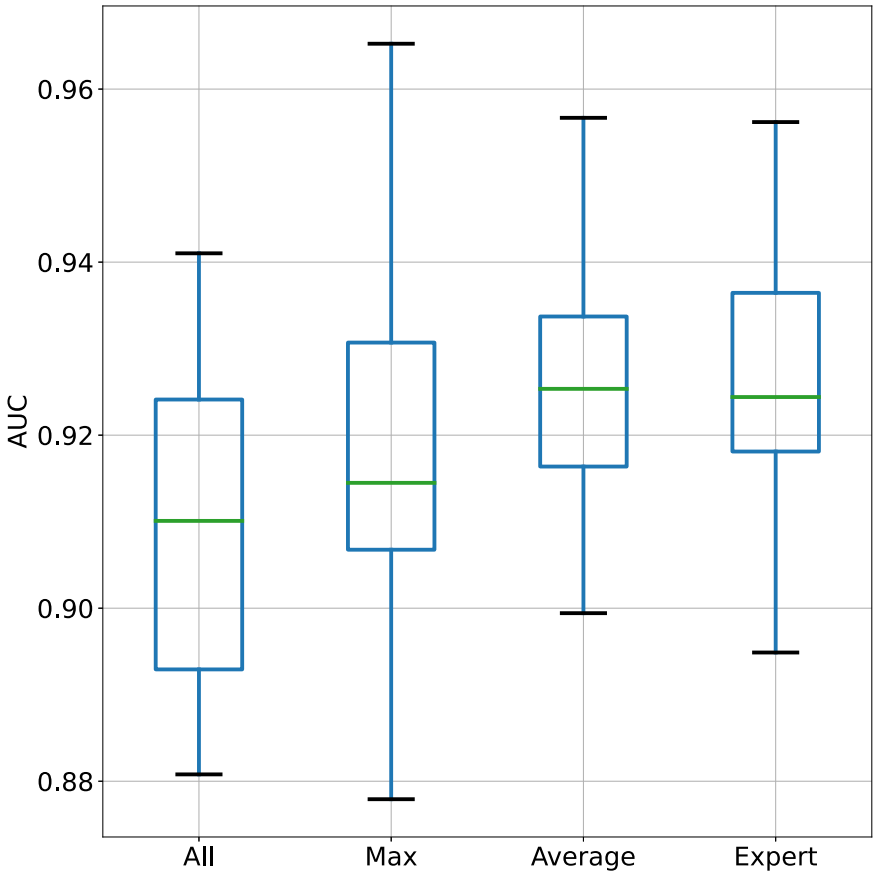

Fig. 3 | Evaluation of model performance when using different methods of aggregating binary-selected informative images. To aggregate these images, we used the average predicted confidence (Average) and maximum predicted confidence (Max) among selected images. We compared the median AUCs to all images/timepoint (All) and expert-selected images (Expert) calculated based on the average probability.

Table 2 | Model performance across ancestry

	AUC (SD) in Testing dataset	
Trained dataset	**POAAGG**	**REFUGE-1**
POAAGG	0.831 +/- 0.009	0.920 +/- 0.004
POAAGG Small	0.604 +/- 0.018	0.579 +/- 0.048
REFUGE	0.611 +/- 0.018	0.825 +/- 0.033

The rows correspond to models trained on POAAGG, $POAAGG_{small}$, and REFUGE, while the columns correspond to the AUCs on the test sets of POAAGG and REFUGE.

pigmentation or optic disc size can vary among ancestry groups, further highlighting the need to validate in datasets with a wide range of age, ethnicity, and genetic backgrounds. Second, studies show that deep learning models often underperform when applied to images captured on different cameras, posing an obstacle to widespread adoption[1]. We did not consider distribution shift induced by using different cameras since the same device was used throughout this study. However, we next plan to assess our model's performance when using images from a non-mydriatic camera (which does not require retinal dilation and is more acceptable to participants for this reason). We are currently planning a prospective study in Penn Medicine Biobank (PMBB) subjects that involves onsite screenings, including non-mydriatic imaging, to validate this model. To account for test data arising from different cameras than the training data, we can apply standard data augmentation techniques for improving accuracy, which shifts the training dataset images in various ways (i.e., color/lighting, lens distortion, resolution) to mimic changes due to different cameras[33,34]. By training on these perturbed images, the neural network learns to predict accurately for a broad range of camera configurations. Third, our images came from an ophthalmology clinic-based study, which had specific inclusion/exclusion criteria that may not fully reflect the real-world population. We must conduct more research on how co-morbid pathologies can impact our model's performance. Additionally, the performance of the model must be tested in real-world settings to ensure its generalizability and determine the optimal number of selected informative images in non-clinic settings. Fourth, we do not fully meet the sensitivity (>0.85) and specificity (>0.95) criteria proposed by Prevent Blindness America (1995)[35] at this time. However, we emphasize that our model is currently best suited for a pre-screening setting (i.e., referring patients with signs of glaucoma to an ophthalmologist) and that future integration of OCT data could further enhance its performance. Finally, although obtaining good empirical results, the performance of these deep learning models cannot be guaranteed, which calls for more advanced techniques to ensure the trustworthiness of these models.

In conclusion, we developed a novel deep learning pipeline that selects high-quality images for detection of POAG in an African ancestry population. As our next steps, we aim to test our model in other ancestry groups to ensure its generalizability, to adapt and apply our model to non-mydriatic images, and to explore how demographic variables influence screening outcomes. We also plan to explore if our model can be used to identify patients at high risk of rapid POAG progression. As a long-term goal, we envision conducting a clinical trial to test this model's performance compared to diagnosis from glaucoma specialists, and ultimately seeking FDA approval for use in POAG screening. Through implementation in primary care offices, public areas like shopping malls, large community screenings, or rural or international locations without access to eye care professionals, this system has the potential to prevent cases of irreversible blindness before they occur.

Methods

Study dataset

All fundus images utilized in this study were from obtained from subjects enrolled in the POAAGG study between 2010 to the present. POAAGG subjects self-identified as African ancestry (Black, Afro-Caribbean, or

African American) and were aged 35 years and older. Subjects were enrolled during regularly scheduled appointments at the Ophthalmology Department at University of Pennsylvania and other sites, and underwent an onsite interview and examination[36]. At the time of enrollment, each subject was classified as a POAG case, suspect, or control by a glaucoma specialist or ophthalmologist based on previously published criteria[18]. In brief, cases were defined as having an open iridocorneal angle and characteristic optic nerve defects with corresponding visual field loss, while controls were patients seen in regularly scheduled ophthalmology appointments without a glaucoma diagnosis or confounding ocular conditions. Suspects were excluded from this study. All subjects provided written informed consent for their participation in the study. The POAAGG study was approved by the University of Pennsylvania Institutional Review Board under protocol #812036 and followed the tenets of the Declaration of Helsinki.

Input into AI model

Dilated fundus images were taken using the Topcon TRC 50EX retinal camera (Topcon Corp. of America, Paramus, NJ) at enrollment and at subsequent appointments. Multiple images were taken at a single timepoint from each subject eye to account for blinking and other confounding factors. Original dimensions of the images were 2392 pixels wide by 2048 pixels high and images were optic nerve head centered.

All images were uploaded to the Scheie Image Reading Center at UPenn using a secure server. Images were taken between 01/13/2004 and 06/25/2019; uploaded to the Reading Center between 01/22/2016 and 04/20/2021; and assessed by non-physician graders at the Reading Center between 06/06/2016 and 05/10/2021[37]. Images from case (Glaucoma) and control (Non-glaucoma) eyes were included in this study, as well as both left and right eyes. Three non-physician graders were trained by two glaucoma specialists to grade these images for a variety of quantitative and qualitative features, as described elsewhere[37]. Before grading, the graders selected the "best" images from the series of images taken at a single time point for each subject eye. The stereo viewer (Screen-Vu stereoscope, Portland, OR) was used to select the pair of images that showed optimal focus, clarity, gain, and contrast; were centered over the disc and surrounding peripapillary atrophy; allowed complete disc visibility; and had greatest magnification. Moving forward, we will call the selected pair of optimal images "***expert***".

The major steps in the development and evaluation of our deep learning model are detailed in Fig. 4.

Step 1: selection of informative images

Fundus images were organized by subject visits, with all images taken at a single timepoint saved in the same folder. Inspired by the process of human graders selecting optimal images for grading (described above), we proposed to develop a model that screens patients using the most informative images from the series taken at a single timepoint. We then investigated if usage of informative images improves model performance, compared to using all images or expert images.

We tested two methods to select the six most informative images for screening. The methods included:

1. Segmentation Binary Classifier ("***Binary***"): We trained a binary classifier on the POAAGG dataset to select the top six images with highest confidence. For the binary classification task, we utilized images from the training set and applied the same preprocessing pipeline used for training the glaucoma classifiers. Each image was labeled based on whether it was selected by the Reading Center annotators: images that were selected were labeled as positive, while the others were labeled as negative. We trained a binary classifier using a ViT-Base model and achieved an AUROC of 0.905.
2. -Based Model ("***Seg***"): We finetuned an image segmentation model, Segformer, to segment the cup and disc from each image. This method measures size as a proxy for informativeness. Segformer was trained on the REFUGE-1 dataset, which has cup/disc segmentation annotations, obtaining a test set accuracy of 85% (calculated by computing the area of the intersection between predicted and annotated area over the total annotated area). Because the POAAGG dataset did not have segmentation annotations, we instead qualitatively evaluated its performance. We found that the model mostly demonstrated reliable segmentation of the optic cup, but under-performed in images that were blurred. For the task of distinguishing the cup/disc region from the background—which is critical for image selection—our segmentation model achieves a mean Intersection-over-Union (IoU) of 76.2%. We then selected the six images with the largest cup-to-disc ratio (CDR) for screening.

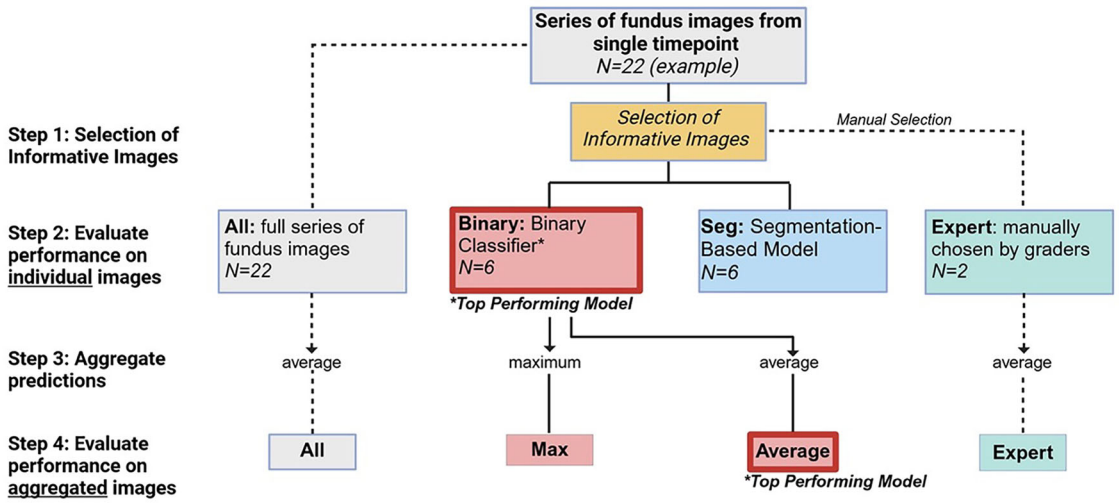

Fig. 4 | Overview of the major steps in developing and evaluating the deep learning model. At each subject appointment, a series of fundus images was taken to account for blinking and other confounding factors. In this figure, we used N = 22 as an example of the number of images collected and outlined the top-performing models in red. *Step 1*: We selected 6 informative images from this series of 22 images using two different methods: a segmentation-based model (Seg) and a Binary Classifier (Binary). Expert images were optimal images manually chosen by non-physician graders at the Scheie Image Reading Center. *Step 2:* We made predictions on individual images (i.e., not yet aggregated) and calculated AUCs, comparing Binary, Seg, Expert, and all images. *Step 3*: Predictions for the Binary model were aggregated by using the average and the maximum predicted probabilities. The average was used for all images and expert images. *Step 4*: We evaluated final model performance on aggregated images.

Step 2: prediction of POAG from each image

To use pre-trained models, we normalized the input pixel values and resized the input. We followed standard data augmentation techniques for training, specifically using the following optimizations: random cropping with a size of 224*224 (in pixels), random horizontal flips, and color jittering.

Next, for training, we made predictions of POAG on each image using two state-of-the-art machine learning models: ResNet and Vision-Transformer (ViT). ResNet models are based on convolutional layers and include ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152. The numbers indicate the counts of convolution layers in model architectures. ViTs are based on transformer layers. We included ViT-Base and ViT-Large, which contain 12 and 24 transformer layers, respectively. We trained ResNet from scratch and used the pre-trained weights on ImageNet for ViT.

The model was trained on a single Nvidia A6000 GPU, using PyTorch as the deep learning framework. In our training, we used stochastic gradient descent as the optimizer, with a step size of 1e-4 and momentum of 0.9. For our main results, we trained the models for 15 epochs. We also included results with more epochs (400) on ResNet-101. To account for the class imbalance on the POAAGG dataset, whose positive labels (Glaucoma) are approximately twice of the number of negative labels (Non-Glaucoma), we used weighted cross-entropy, with penalty on positive labels as 1.0 and on negative labels 2.0. As long as a sufficient number of images from the positive class are available, these techniques help to restore balance to an imbalanced dataset, allowing us to achieve similar accuracy in real-world populations where the imbalance ratio can be closer to 1:20.

We then compared the AUC using ResNet versus ViT. Based on these results, we used ViT in our final POAG screening pipeline.

Step 3: combination of POAG predictions

Next, we aimed to ensemble POAG predictions from selected informative images from the same timepoint. We used two approaches to combine predictions from informative images: *average* predicted confidence and *maximum* predicted confidence across selected informative images. We then classified the patient for that visit as Glaucoma if the averaged or maximum confidence was above 0.5, and Non-Glaucoma otherwise.

Evaluation of model performance

To account for the class imbalance, namely {Glaucoma vs. Non-Glaucoma}, we measured the model performance using AUC. To account for randomness in the training process, we conducted multiple trainings using different random seeds (resulting in different train/test splits, as well as different optimization runs) and reported the means and SDs of AUCs.

First, we assessed if informative image selection improved model prediction. After selecting informative images from the binary classifier, segmentation-based model, and experts (Binary, Seg, Expert), we made predictions on individual images (i.e., not yet aggregated) and calculated AUCs by comparing individual predicted confidences to their true labels. After finding that the binary classifier led to the highest performance (see Results), we used this method of informative image selection in our pipeline moving forward.

We next aggregated predictions from Binary-selected images using the *average* and *maximum* predicted probabilities. We found that the binary classifier using average predicted probability performed on par with expert-selected images (see Results), and finalized our screening pipeline (Fig. 1).

Transferability across datasets

To study how the trained model performed on the POAAGG dataset and transferred across different ancestry groups, we evaluated on both the POAAGG and REFUGE-1 datasets. The REFUGE dataset (Retinal Fundus Glaucoma Challenge) is collected from a Chinese population[38]. The POAAGG dataset contains 64,129 images while the REFUGE dataset contains 1200 images which are equally split into training, validation, and test sets.

We trained resnet101 on the POAAGG and REFUGE datasets. Also, to ablate the effect of dataset size, we also trained ResNet-101 on a subset of POAAGG with the same size as REFUGE. We named this subset $POAAGG_{small}$.

Data availability

The datasets generated and/or analyzed during the current study are not publicly available due to restrictions on sharing patient images for the POAAGG study, but may be made available from the corresponding author to qualified researchers upon reasonable request and with permission of the University of Pennsylvania Institutional Review Board (IRB). The informed consent form signed by POAAGG subjects requires that all enrolled subjects have an opt in/out period of at least 30 days for participation in additional studies. Furthermore, the requesting investigator must be added to the IRB of the POAAGG study to allow access to patient data.Code Availability The underlying code for this study [and training/validation datasets] is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author and following all IRB procedures, described in the above section. We have regularly shared data over the last 15 years and are in full compliance with NIH data sharing policies.

Code availability

The underlying code for this study [and training/validation datasets] is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author and following all IRB procedures, described in the above section. We have regularly shared data over the last 15 years and are in full compliance with NIH data sharing policies.

Received: 5 September 2024; Accepted: 23 December 2025;
Published online: 06 February 2026

References

1. Thompson, A. C., Jammal, A. A. & Medeiros, F. A. A review of deep learning for screening, diagnosis, and detection of glaucoma progression. *Transl. Vis. Sci. Technol.* **9**, 42 (2020).
2. Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N. & Folk, J. C. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit. Med.* **1**, 39–6 (2018). eCollection 2018.
3. Poplin, R. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
4. Ittoop, S. M., Jaccard, N., Lanouette, G. & Kahook, M. Y. The role of artificial intelligence in the diagnosis and management of glaucoma. *J. Glaucoma* **31**, 137–146 (2022).
5. Weinreb, R. N. et al. Primary open-angle glaucoma. *Nat. Rev. Dis. Prim.* **2**, 1–19 (2016).
6. Quigley, H. A. & Broman, A. T. The number of people with glaucoma worldwide in 2010 and 2020. *Br. J. Ophthalmol.* **90**, 262–267 (2006).
7. Mitchell, P., Smith, W., Attebo, K. & Healey, P. R. Prevalence of open-angle glaucoma in Australia. The Blue Mountains Eye Study. *Ophthalmology* **103**, 1661–1669 (1996).
8. Delgado, M. F. et al. Management of glaucoma in developing countries: challenges and opportunities for improvement. *Clinicoecon Outcomes Res* **11**, 591–604 (2019).
9. Myers, J. S., Fudemberg, S. J. & Lee, D. Evolution of optic nerve photography for glaucoma screening: a review. *Clin. Exp. Ophthalmol.* **46**, 169–176 (2018).
10. Miller, S. E. et al. Glaucoma screening in Nepal: cup-to-disc estimate with standard mydriatic fundus camera compared to portable nonmydriatic camera. *Am. J. Ophthalmol.* **182**, 99–106 (2017).

11. Ting, D. S. W. et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* **318**, 2211–2223 (2017).
12. Li, Z. et al. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology* **125**, 1199–1206 (2018).
13. Rogers, T. W. et al. Evaluation of an AI system for the automated detection of glaucoma from stereoscopic optic disc photographs: the European Optic Disc Assessment Study. *Eye* **33**, 1791–1797 (2019).
14. Tielsch, J. M. et al. Racial variations in the prevalence of primary open-angle glaucoma. The Baltimore Eye Survey. *JAMA* **266**, 369–374 (1991).
15. Broman, A. T. et al. Estimating the rate of progressive visual field damage in those with open-angle glaucoma, from cross-sectional data. *Invest. Ophthalmol. Vis. Sci.* **49**, 66–76 (2008).
16. Munoz, B. et al. Causes of blindness and visual impairment in a population of older Americans: The Salisbury Eye Evaluation Study. *Arch. Ophthalmol.* **118**, 819–825 (2000).
17. Abràmoff, M. D. et al. Considerations for addressing bias in artificial intelligence for health equity. *NPJ Digit. Med.* **6**, 170–179 (2023).
18. Charlson, E. S. et al. The Primary Open-Angle African American Glaucoma Genetics Study: Baseline Demographics. *Ophthalmology* **122**, 711–720 (2015).
19. Verma, S. S. et al. A multi-cohort genome-wide association study in African ancestry individuals reveals risk loci for primary open-angle glaucoma. *Cell* **187**, 464–480.e10 (2024).
20. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int. J. computer Vis.* **128**, 336–359 (2020).
21. Susanna, R., De Moraes, C. G., Cioffi, G. A. & Ritch, R. Why do people (still) go blind from glaucoma?. *Transl. Vis. Sci. Technol.* **4**, 1 (2015).
22. Thomas, S. et al. The effectiveness of teleglaucoma versus in-patient examination for glaucoma screening: a systematic review and meta-analysis. *PLoS One* **9**, e113779 (2014).
23. Rajesh, A. E., Davidson, O. Q., Lee, C. S. & Lee, A. Y. Artificial Intelligence and Diabetic Retinopathy: AI Framework, Prospective Studies, Head-to-head Validation, and Cost-effectiveness. *Diab. Care* **46**, 1728–1739 (2023).
24. Al-Aswad, L. A. et al. Evaluation of a Deep Learning System For Identifying Glaucomatous Optic Neuropathy Based on Color Fundus Photographs. *J. Glaucoma* **28**, 1029–1034 (2019).
25. Ahn, J. M. et al. A deep learning model for the detection of both advanced and early glaucoma using fundus photography. *PLoS One* **13**, e0207982 (2018).
26. Liu, H. et al. Development and Validation of a Deep Learning System to Detect Glaucomatous Optic Neuropathy Using Fundus Photographs. *JAMA Ophthalmol.* **137**, 1353–1360 (2019).
27. Hemelings, R. et al. Deep learning on fundus images detects glaucoma beyond the optic disc. *Sci. Rep.* **11**, 20313–1 (2021).
28. Thakur, A., Goldbaum, M. & Yousefi, S. Predicting glaucoma before onset using deep learning. *Ophthalmol. Glaucoma* **3**, 262–268 (2020).
29. AlRyalat, S. A., Singh, P., Kalpathy-Cramer, J. & Kahook, M. Y. Artificial Intelligence and Glaucoma: Going Back to Basics. *Clin. Ophthalmol.* **17**, 1525–1530 (2023).
30. Bragança, C. P., Torres, J. M., Macedo, L. O. & Soares, C. P.dA. Advancements in glaucoma diagnosis: the role of ai in medical imaging. *Diagnostics (Basel)* **14**, 530 (2024).
31. Chen, D. et al. Applications of artificial intelligence and deep learning in glaucoma. *Asia-Pac. J. Ophthalmol.* **12**, 80–93 (2023).
32. Christopher, M. et al. Effects of study population, labeling and training on glaucoma detection using deep learning algorithms. *Transl. Vis. Sci. Technol.* **9**, 27 (2020).
33. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
34. Taori, R. et al. *Measuring robustness to natural distribution shifts in image classification* (Proceedings of the 34th International Conference on Neural Information Processing Systems, Curran Associates Inc, Red Hook, NY, USA, 2020).
35. Stamper, R. L. Glaucoma screening. *J. Glaucoma* **7**, 149–150 (1998).
36. Salowe, R. J. et al. Recruitment strategies and lessons learned from a large genetic study of African Americans. *PLOS Glob. Public. Health* **2**, e0000416 (2022). Epub 2022 Aug 5.
37. Addis, V. et al. Non-physician grader reliability in measuring morphological features of the optic nerve head in stereo digital images. *Eye (Lond)*, (2019).
38. Orlando, J. I. et al. REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Med. Image Anal.* **59**, 101570 (2020).

Acknowledgements

This research was supported by the Collaborative Research in Trustworthy AI for Medicine grant from the University of Pennsylvania. The Primary Open-Angle African American Glaucoma Genetics (POAAGG) study was supported by the National Eye Institute, Bethesda, Maryland (grant #1R01EY023557) and Vision Research Core Grant (P30 EY001583). Funds also come from the F.M. Kirby Foundation, Research to Prevent Blindness, The UPenn Hospital Board of Women Visitors, and The Paul and Evanina Bell Mackall Foundation Trust. Support also came from the Ophthalmology Department at the Perelman School of Medicine and the VA Hospital in Philadelphia, PA. The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Author contributions

S.L., R.S., J.M.O., and O.B. conceptualized and designed the study. R.S., R.L., and J.M.O. collected and curated the data and images. S.L., I.L., and O.B. developed the algorithm. G.S.Y. led all statistical analysis. S.L. and R.S. contributed to visualization and wrote the original draft of the manuscript. O.B. and J.M.O. acquired funding and supervised the study's implementation. All authors contributed to the writing review and editing of the manuscript. S.L. and R.S. contributed equally to this study and shared co-first authorship.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-02318-2.

Correspondence and requests for materials should be addressed to Joan O'Brien.