Title

Evaluating a Foundation Artificial Intelligence Model for Glaucoma Detection Using Color Fundus Photographs

Journal

Ophthalmology Science, 5(1)

ISSN

2666-9145

Authors

Chuter, Benton

Huynh, Justin

Hallaj, Shahin

et al.

Publication Date

2025

DOI

10.1016/j.xops.2024.100623

Peer reviewed

AMERICAN ACADEMY OF OPHTHALMOLOGY®

Evaluating a Foundation Artificial Intelligence Model for Glaucoma Detection Using Color Fundus Photographs

Benton Chuter, MS,[1] Justin Huynh, MS,[1,2] Shahin Hallaj, MD,[1] Evan Walker, MS,[1] Jeffrey M. Liebmann, MD,[3] Massimo A. Fazio, PhD,[4] Christopher A. Girkin, MD, MSPH,[4] Robert N. Weinreb, MD,[1] Mark Christopher, PhD,[1] Linda M. Zangwill, PhD[1]

Purpose: To evaluate RETFound, a foundation artificial intelligence model, using a diverse clinical research dataset to assess its accuracy in detecting glaucoma using optic disc photographs. The model's accuracy for glaucoma detection was evaluated across race, age, glaucoma severity, and various training cycles (epochs) and dataset sample sizes.

Design: Evaluation of a diagnostic technology.

Participants: The study included 9787 color fundus photographs (CFPs) from 2329 participants of diverse race (White [73.4%], Black [13.6%] and other [13%]), disease severity (21.8% mild glaucoma, 7.2% moderate or advanced glaucoma, 60.3% not glaucoma, and 10.7% unreported), and age (48.8% <60 years, 51.1% >60 years) from the Diagnostic Innovations in Glaucoma Study and the African Descent and Glaucoma Evaluation Study. All fundus photographs were graded as "Glaucomatous" or "Non-glaucomatous."

Methods: The study employed RETFound, a self-supervised learning model, to perform binary glaucoma classification. The diagnostic accuracy of RETFound was iteratively tested across different combinations of dataset sample sizes (50–2000 optic disc photographs), training cycles (5–50), and study subpopulations stratified by severity of glaucoma, age, and race).

Main Outcome Measures: Diagnostic accuracy area under the receiver operating characteristic curve (AUC) for classifying CFP as "Glaucomatous" or "Non-glaucomatous."

Results: Performance increased with larger training datasets and more training cycles, improving from 50 training images and 5 epochs (AUC: 0.52) to 2000 training images and 50 epochs (AUC: 0.86), with reduced gain in performance from approximately 500 and 1000 training images (AUC of 0.82 and 0.83, respectively). Performance was consistent across race and age for all training size and cycle number combinations: Black (AUC = 0.87) vs. other (AUC = 0.86), and >60 years (AUC = 0.84) vs. <60 years (AUC = 0.87). Performance was significantly higher in patients with moderate to severe vs. mild glaucoma (AUC = 0.95 vs. 0.84, respectively).

Conclusions: Good RETFound performance was observed with a relatively small sample size of optic disc photographs used for fine-tuning and across differences in race and age. RETFound's ability to adapt across a range of CFP training conditions and populations suggests it is a promising tool to automate glaucoma detection in a variety of use cases.

Financial Disclosures: Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2025;5:100623 © 2024 Published by Elsevier Inc. on behalf of the American Academy of Ophthalmology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).*

Supplemental material available at www.ophthalmologyscience.org.

Ophthalmology has witnessed remarkable advancements in the application of deep learning artificial intelligence (AI) models.[1] These models have demonstrated considerable success across a spectrum of diseases including glaucoma[2] and diabetic retinopathy,[3] data modalities including retinal color fundus photographs (CFPs) and OCT images,[4,5] and tasks including disease classification[6,7], disease progression,[8] prediction of disease worsening,[9,10] and optic nerve head (ONH) segmentation.[11] A long-standing limitation of the majority of existing algorithms, however, is their dependence on datasets labeled by domain experts—a process that is both labor-intensive and time-consuming [12,13] Such models may also be task-specific and have limited generalizability to different clinical applications.[14]

The emergence of foundation models, a class of large-scale AI models pretrained on extensive datasets and capable of fine-tuning for diverse downstream tasks including detection of a various diseases, holds promise to

https://doi.org/10.1016/j.xops.2024.100623
ISSN 2666-9145/24

address these issues.[15] In medicine, foundation models have demonstrated potential to employ vast datasets through self-supervised learning. In self-supervised learning, models learn useful representations and relevant features from unlabeled data without use of human labor-intensive labels, mitigating this limitation and enabling generalization across various tasks.[16] In addition, 1 foundation model can be used to detect several different diseases making it a flexible option for a wide range of tasks.

RETFound, a recently developed foundational model, poses a possible solution to the challenges of labeled data acquisition and generalizability in developing AI for ophthalmology.[17] RETFound is built upon 1.6 million unlabeled retinal images using self-supervised learning, aiming to provide a generalizable solution that outperforms existing models in diagnosing and predicting sight-threatening eye diseases while requiring fewer labeled data. RETFound employs a transformer-based architecture to handle the complexity and variance inherent in retinal imaging, with pretext tasks such as masked autoencoding to support development of retinal image representations that can be used in subsequent fine-tuned applications. Potential applications include detection of ophthalmic diseases such as diabetic retinopathy, glaucoma, and age-related macular degeneration as well as oculomic challenges including identification of ischemic heart disease, stroke, heart failure, and Parkinson's disease. Operating with both fundus photography and OCT, this foundation model represents a significant advancement in medical AI, offering potential to reduce the annotation workload and improve AI applications in retinal imaging. However, its ability to do so across varying datasets and conditions remains to be established.[17]

Preliminary assessments of RETFound suggest potential application across diseases and imaging modalities.[17] However, the model's subsequent fine-tuning and evaluation have thus far been confined to publicly accessible datasets, where variations in image and label quality are prevalent. In addition, RETFound was developed from a limited United Kingdom cohort. It is crucial to further validate the RETFound model with large geographically and demographically diverse external datasets.

Addressing this gap, our research aims to perform a validation study of RETFound on a large, well-characterized, diverse CFP dataset of eyes with and without glaucoma. Through this study, we seek to rigorously evaluate the performance and applicability of RETFound in an independent context, thereby facilitating its potential integration into diagnostic workflows and enhancing the precision of ophthalmic assessments. Furthermore, it is important to determine how many images and training cycles are necessary to finetune the RETFound model on a new dataset and task. This work explores the ability of RETFound to detect glaucoma using CFPs, with varying amounts of training time and data. We also determine its generalizability to individuals of different races, ages, and disease severity. In this way, we sought to address the question: for a given dataset size and number of training cycles, what kind of performance, with what variability and generalizability, can a user expect from RETFound in assigning binary labels to CFPs?

Methods

Data Collection

This study used CFPs from the Diagnostic Innovations in Glaucoma Study (DIGS) (clinicaltrials.gov identifier: NCT00221897)[18] and the African Descent and Glaucoma Evaluation Study (ADAGES) (clinicaltrials.gov identifier: NCT00221923).[19] The study's recruitment and methods received approval from the institutional review boards of each involved institution (University of California, San Diego, University of Alabama at Birmingham, and Columbia University), in line with the Declaration of Helsinki and the Health Insurance Portability and Accountability Act, ensuring all participants gave informed consent at recruitment. While the methods of these studies have been detailed in previous publications,[18,19] key relevant details are summarized here.

The DIGS and ADAGES studies represent a collaborative effort involving the University of California, San Diego Hamilton Glaucoma Center and Viterbi Family Department of Ophthalmology, the University of Alabama at Birmingham Department of Ophthalmology, and the Columbia University Medical Center Edward S. Harkness Eye Institute. The study population includes a diverse population of individuals of African, European, and Asian descent. The study protocols include semiannual collection of stereo fundus photographs and visual field (VF) tests as part of their longitudinal framework. For this analysis, 9787 fundus photographs were included. Fundus photographs were captured as simultaneous stereoscopic ONH images between 1986 and 2019. Several different cameras were used over the years, including a Nidek Stereo Camera Model 3-DX (Nidek Inc). Visual field assessments were conducted using the Humphrey Field Analyzer II with a 24-2 standard testing pattern and the Swedish Interactive Thresholding Algorithm, discarding tests with >33% in fixation losses, false-negative, or false-positive errors. The mean deviation (MD) from VF testing, conducted closest to the time of image capture and within a year, was used to approximate the severity of glaucoma damage at the imaging time for all ONH images.

Glaucoma Labels

To assess glaucoma status for DIGS/ADAGES images, stereophotographs were reviewed by 2 independent, masked graders using a stereoscopic viewer such as the Asahi Pentax Stereo Viewer II (Pentax). Criteria for a glaucoma status label comprised evidence of, as demonstrated by excavation, neuroretinal rim thinning or notching, or localized or diffuse retinal nerve fiber layer defect, as assessed by these graders based solely on fundus imaging, with or without VF damage. Disagreements between graders were resolved by a third, experienced grader. Stereo image pairs were then separated into individual images of the ONH. The resulting dataset included 7411 stereo pairs, which were divided into 14 822 individual ONH images, taken from 4363 eyes belonging to 2329 participants. In this study, the data underwent cross-sectional evaluation with images assigned a binary label (Nonglaucomatous or Glaucomatous) at the image level. Images where poor quality precluded confident determination of Glaucomatous Optic Neuropathy status by any of the graders were excluded to yield the images used in this study.

Image Preprocessing

All CFPs selected for this study underwent preprocessing through an automated segmentation tool designed to identify and extract a square crop centered on the ONH, with each side measuring 2.5 times the optic disc's diameter. When applied to images that had

passed the manual quality screen, cropping algorithm fail rate was <1%. This uniform crop was established to ensure a consistent basis for examining the ONH area. A 2.5x disc diameter frame was chosen to accommodate the widest range of visual information accessible from the various cameras and configurations used to acquire the images, despite potential limitations in capturing details far from the disc. Previous studies have validated the effectiveness of this framing size for accurate primary open-angle glaucoma detection with convolutional neural networks (CNNs).[20,21]

Following this standardized cropping, the images were resized to a standard 224 × 224 pixel dimension, and a specialist (M.C.) performed a manual inspection of each one to verify the precise alignment of the ONH. The choice of 224 × 224 pixels was informed by its compatibility with the input requirements of the RETFound and this resolution's prior effectiveness in diagnosing primary open-angle glaucoma in our past experiments.[21] The available images included either simultaneous stereo photographs, sequential stereo images, created by taking 2 successive shots with a monocular fundus camera to simulate a stereo effect. In our analysis, stereo images were separated and analyzed as if they were single-view to include all available data.

Self-Supervised Learning, RETFound

Self-supervised learning attempts to improve data use efficiency by creating supervisory signals without externally provided labels.[22] Models engage in "pretext tasks" that do not require labels. This approach utilizes vast quantities of unlabeled data to develop versatile feature representations suitable for various tasks.

After this initial training, models undergo fine-tuning for specific applications, such as classification or segmentation. Self-supervised learning has demonstrated superiority over supervised learning methods, such as those involving pretraining on ImageNet with categorical labels, across numerous computer vision challenges, achieving higher performance with less data during fine-tuning.[16] Moreover, self-supervised learning models excel over supervised counterparts in tests involving new, domain-divergent data, showcasing their strong generalization and superior fine-tuning capabilities.[23,24] This underscores self-supervised learning's significant promise for medical AI, where data is plentiful, tasks are varied, and labels are limited.[25,26]

RETFound adapts self-supervised learning to ophthalmic imaging modalities, namely CFP and OCT. As described in its original publication, RETFound was created from large-scale unlabeled retinal images through self-supervised learning.[17] Two distinct RETFound models were developed, 1 using CFP and the other using OCT, through application on natural and retinal images from the Moorfields Eye Hospital-MoorfIelds Diabetic imAge dataSet and other publicly available datasets, involving 904170 CFPs and 736442 OCTs. RETFound was then adapted for various challenging detection and prediction tasks by fine-tuning with task-specific labels. In addition to the diagnostic classification of ocular diseases such as diabetic retinopathy and glaucoma, RETFound also identified heart failure, stroke, and Parkinson's disease.

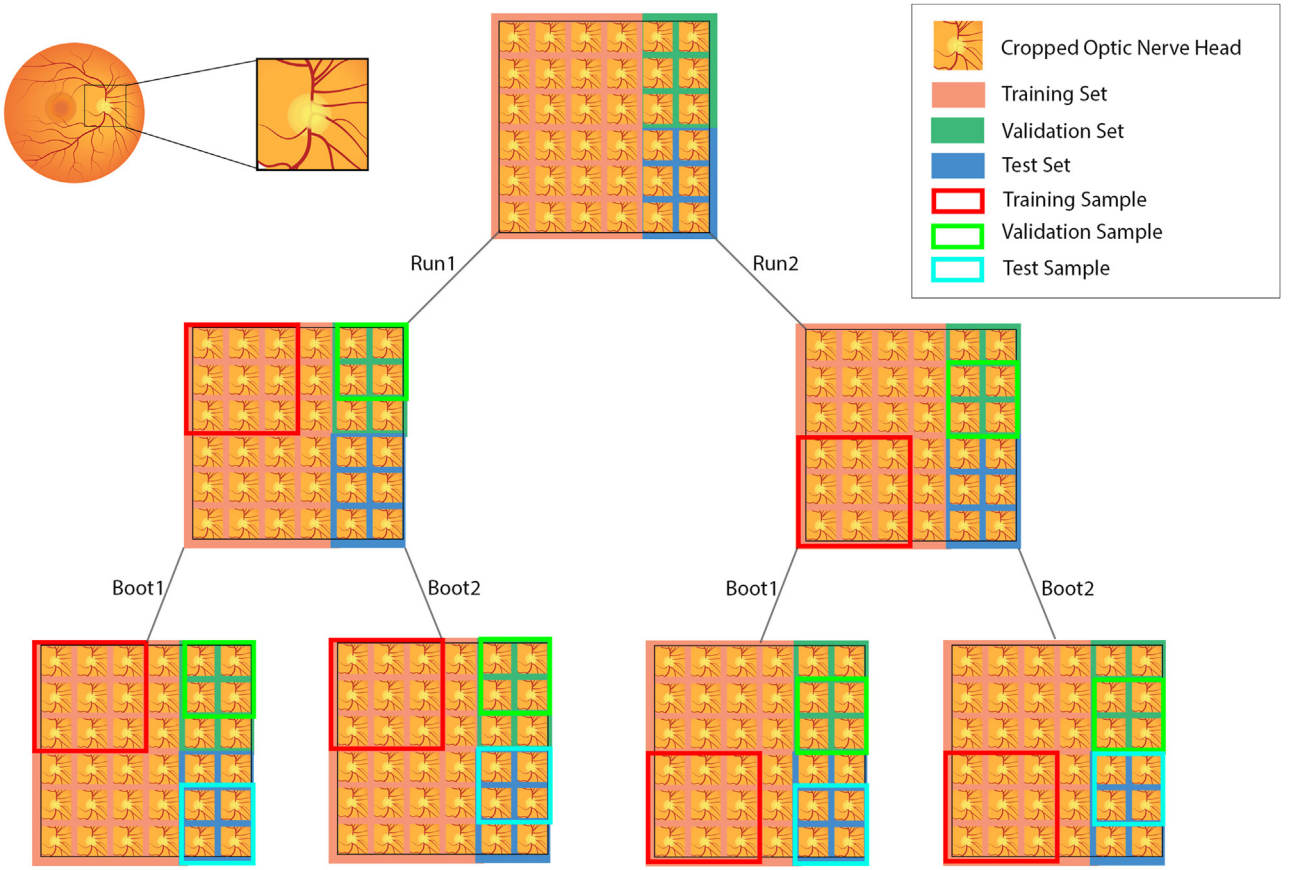

Figure 1. Schematic showing how the data was split per training run, per bootstrap test run. For each run, the train and validation samples sum to the number of images being tested for that specific epoch and sample size combination. Not to scale.

Evaluation of RETFound Using CFPs for Glaucoma Label Assignment

We assessed the practical application and performance of RETFound for glaucoma detection from CFPs. Comprehensive iterative testing of RETFound performance, captured as area under the receiver operating characteristic (AUC) in predicting glaucoma status when trained with datasets of CFPs of varying size (50, 100, 200, 500, 1000, 2000) for varying trainings (epochs: 5, 10, 20, 50), was conducted. In this way we evaluate whether it can perform at or above the level of other deep learning models in categorization tests with relatively little subsequent training (fine-tuning) on relatively small labeled datasets. We also evaluate whether RETFound's results are generalizable to differences in glaucoma severity, age, and race.

Number of Images Variation and Dataset Split

Nine thousand seven hundred eighty-seven images from 2104 patients were randomized into train (6884—1472 patients), validation (976—211 patients), and test (2127—421 patients) pools according to a standard 70-10-20 split by patient (Fig 1). Demographic features of the study population as a whole and for each of these pools can be found in Table 1, further stratified for all images used by glaucoma severity, race, and age in Table S2 (available at www.ophthalmologyscience.org). RETFound was then iteratively tested for varying dataset sizes (50, 100, 200, 500, 1000, 2000) and epochs (5, 10, 20, 50) for a total of 24 size-epoch combinations. These dataset size and epoch ranges were chosen after initial testing suggested these ranges would represent broad model performance from poor to good.

Table 1. Overview of Study Population

Characteristics	All (n = 2104 Subjects; 3973 eyes)	Train (n = 1472 Subjects; 2788 eyes)	Validation (n = 211 Subjects; 395 eyes)	Test (n = 421 Subjects; 790 eyes)
Baseline age	58.4 (57.7, 59.0)	58.3 (57.5, 59.1)	58.2 (56.0, 60.4)	58.6 (57.1, 60.1)
Baseline age classification				
Age <60	1027 (48.8%)	714 (48.5%)	110 (52.1%)	202 (48.0%)
Age >60	1076 (51.2%)	757 (51.4%)	101 (47.9%)	219 (52.0%)
Unknown or not reported	1 (0.0%)	1 (0.1%)	0 (0.0%)	0 (0.0%)
Sex				
Female	1225 (58.2%)	867 (58.9%)	119 (56.4%)	239 (56.8%)
Male	878 (41.7%)	604 (41.0%)	92 (43.6%)	182 (43.2%)
Unknown or not reported	1 (0.0%)	1 (0.1%)	0 (0.0%)	0 (0.0%)
Race				
American Indian/Alaska Native	3 (0.1%)	2 (0.1%)	1 (0.5%)	0 (0.0%)
Asian	128 (6.1%)	95 (6.5%)	10 (4.7%)	23 (5.5%)
Black or African American	287 (13.6%)	189 (12.8%)	31 (14.7%)	67 (15.9%)
Native Hawaiian or Other Pacific Islander	4 (0.2%)	3 (0.2%)	1 (0.5%)	0 (0.0%)
Unknown or not reported	138 (6.6%)	101 (6.9%)	14 (6.6%)	23 (5.5%)
White	1544 (73.4%)	1082 (73.5%)	154 (73.0%)	308 (73.2%)
Ethnicity				
Hispanic	21 (1.0%)	15 (1.0%)	2 (0.9%)	4 (1.0%)
Not Hispanic	1061 (50.4%)	718 (48.8%)	113 (53.6%)	230 (54.6%)
Unknown or not reported	1022 (48.5%)	739 (50.2%)	96 (45.5%)	187 (44.4%)
Diabetes				
No	1968 (93.5%)	1383 (94.0%)	197 (93.4%)	388 (92.2%)
Yes	136 (6.5%)	89 (6.0%)	14 (6.6%)	33 (7.8%)
Hypertension				
No	1646 (78.2%)	1160 (78.8%)	165 (78.2%)	321 (76.2%)
Yes	458 (21.8%)	312 (21.2%)	46 (21.8%)	100 (23.8%)
24-2 VF MD (dB)	−2.60 (−2.81, −2.39)	−2.62 (−2.87, −2.37)	−2.28 (−2.81, −1.76)	−2.72 (−3.21, −2.23)
Baseline disease severity				
Mild glaucoma	867 (21.8%)	597 (21.4%)	83 (21.0%)	187 (23.7%)
Moderate to advanced glaucoma	285 (7.2%)	202 (7.2%)	21 (5.3%)	62 (7.8%)
Nonglaucomatous	2396 (60.3%)	1680 (60.3%)	252 (63.8%)	464 (58.7%)
Unknown or not reported	425 (10.7%)	309 (11.1%)	39 (9.9%)	77 (9.7%)
Axial length (mm)	24.06 (23.99, 24.13)	24.02 (23.93, 24.11)	24.14 (23.90, 24.37)	24.14 (23.98, 24.30)
Spherical equivalent	−0.74 (−0.84, −0.64)	−0.73 (−0.85, −0.62)	−0.70 (−1.06, −0.34)	−0.80 (−1.03, −0.56)
IOP (mmHg)	18.96 (18.71, 19.21)	18.97 (18.67, 19.27)	18.59 (17.80, 19.37)	19.10 (18.51, 19.68)
CCT (μm)	553.35 (551.49, 555.21)	554.13 (551.89, 556.38)	549.30 (543.90, 554.69)	552.63 (548.47, 556.79)
Baseline visit glaucoma classification				
No	2396 (60.3%)	1680 (60.3%)	252 (63.8%)	464 (58.7%)
Yes	1577 (39.7%)	1108 (39.7%)	143 (36.2%)	326 (41.3%)
Last visit glaucoma classification				
No	2336 (58.8%)	1639 (58.8%)	252 (63.8%)	445 (56.3%)
Yes	1637 (41.2%)	1149 (41.2%)	143 (36.2%)	345 (43.7%)

CCT = central corneal thickness; dB = decibels; IOP = intraocular pressure; MD = mean deviation; VF = visual field.

For each of these size-epoch combinations, models were trained, validated, and tested on train, validation, and test sets generated by random sampling from the predetermined train, validation, and test pools of 6684, 976, and 2127 images, respectively (70-10-20 split), where the total number of images reported represents the current size for the current size-epoch combination being tested (#train_sample + #validation_sample = current size) (Fig 1). To both accommodate and assess variability inherent between training runs, each size-epoch combination involved 10 separate training runs for which this train-validation-test sampling process was repeated. For each of these training runs, a further 100 bootstrap runs were conducted, for a total of 240 training runs and 24 000 bootstrap runs. Details regarding implementation details can be found in Table S3 (available at www.ophthalmologyscience.org).

Analysis

Given the relatively balanced nature of the DIGS/ADAGES dataset with respect to the presence of glaucoma (Table 1), performance for each run was assessed via AUC; 95% confidence intervals (CIs) for these AUCs were calculated by 2 methods: via assumption of a normal distribution, and by a cumulative density function (CDF). For CIs calculated using the CDF method, this involves sorting the bootstrapped estimates and selecting values that correspond to the 2.5th and 97.5th percentiles of the bootstrapped distribution. These different CIs serve distinct purposes: CDF CI better captures the degree of variability between runs, especially at low numbers of images. Normal distribution CI, with the large sample size, suggests the significance of differences in performance between epoch and image number combinations. In addition, the generalizability of the models was evaluated by stratifying by race, (African descent vs. not of African descent), age (above and below the median 60 years), and severity of glaucoma (mild VF MD >-6 decibels [dB] vs. moderate to severe VF MD <-6 dB).

Results

This study included 2104 subjects and 3973 eyes, with subsets for testing (421 subjects and 790 eyes), training (1472, 2788), and validation (211, 395), as displayed in Table 1. The average age of participants in the study is 58 years, with 48.8% (n = 1027) of participants <60 years of age and 51.1% (n = 1076) >60 years of age. Females (n = 1225, 58.2%) slightly outnumbered males (n = 878, 41.7%). A majority of the study population are White (1544, 73.5%), followed by people of Black/African descent (287, 13.6%) and Asian (128, 6.1%). Racial status was unknown or unrecorded for 138 (6.6%) of study participants. Eye-level characteristics including 24-2 VF MD (dB), axial length (mm), spherical equivalent, intraocular pressure (mmHg), and central corneal thickness (μm), are presented for the training, validation, and test set in Table 1. Twenty-one point eight percent (n = 867) of participants' eyes had mild glaucoma (dB >-6 on 24-2 VF MD) compared with 7.2% (n = 285) with moderate or severe glaucoma (VF MD dB <-6), while 60.3% (n = 2396) were not glaucomatous and 10.7% (n = 425) did not have a recorded dB measurement. The datasets are relatively well-balanced between the 2 categories of interest at most recent visit: glaucoma (n = 2336, 58.8%) or not glaucoma (n = 1637, 41.2%).

Figures 2, S3 and S4 (available at www.ophthalmologyscience.org) illustrate the model performance for different epoch and image sample size combinations. Ninety-five percent CIs calculated with the standard normal distribution method are much narrower than those using CDF and show statistically significant ($P < 0.001$) differences between each epoch and image number combination. Cumulative density function CIs in contrast are broader and show a relatively high degree of variability between training and testing runs at small sample sizes.

When holding epoch or sample size constant to observe changes in AUC in response to increasing image number or epoch, respectively, there is a positive trend of increasing AUC with a larger number of images, suggesting that more data contribute to better model performance. For all epochs, as the number of images increases from 50 to 2000, the AUC also increases, indicating improved model performance; for 50 epochs this represents an increase from 0.574 at 50 images to 0.859 at 2000 images. However, the rate of improvement diminishes with more images and epochs, as

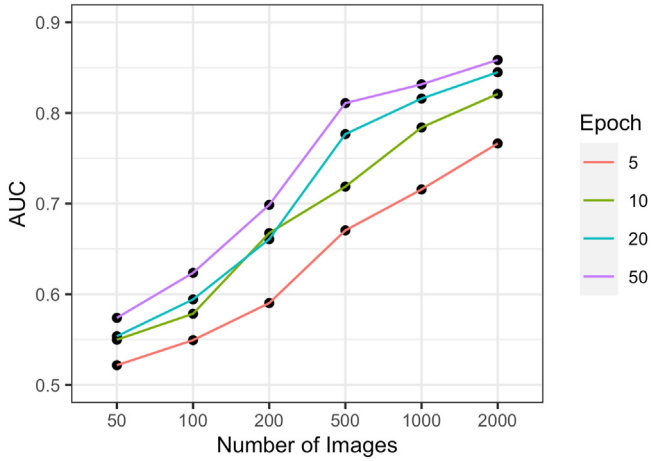

Figure 2. Plots demonstrating the relationship between number of images (x-axis) and diagnostic performance, as measured by area under the receiver operating characteristic curve (AUC, y-axis), at each tested epoch number.

Table 4. Summary of the Model Performance on the Local Datasets as Captured by AUC, with 95% Confidence Intervals as Calculated by Cumulative Density Function

			Disease Severity		Age		Race	
Epoch	Number of Images	Overall (n = 421 Subjects; 790 Eyes; 2127 Images)	*Mild Glaucoma (n = 162 Subjects; 270 Eyes; 871 Images)*	*Moderate to Advanced Glaucoma (n = 58 Subjects; 75 Eyes; 355 Images)*	*Age Below 60 (n = 201 Subjects; 383 Eyes; 865 Images)*	*Age Above 60 (n = 251 Subjects; 468 Eyes; 1262 Images)*	*Black or African American (n = 67 Subjects; 128 Eyes; 370 Images)*	*Other Races (n = 354 Subjects; 662 Eyes; 1757 Images)*
5	50	0.522 (0.420, 0.627)	0.514 (0.424, 0.606)	0.555 (0.415, 0.740)	0.525 (0.384, 0.655)	0.527 (0.453, 0.633)	0.529 (0.394, 0.688)	0.526 (0.449, 0.631)
5	100	0.549 (0.456, 0.631)	0.547 (0.467, 0.620)	0.576 (0.405, 0.741)	0.526 (0.413, 0.628)	0.532 (0.416, 0.615)	0.541 (0.382, 0.712)	0.549 (0.478, 0.625)
5	200	0.590 (0.447, 0.657)	0.558 (0.417, 0.615)	0.688 (0.518, 0.802)	0.580 (0.471, 0.675)	0.606 (0.515, 0.685)	0.633 (0.450, 0.761)	0.584 (0.469, 0.659)
5	500	0.670 (0.598, 0.716)	0.634 (0.567, 0.684)	0.769 (0.673, 0.843)	0.649 (0.510, 0.717)	0.651 (0.579, 0.707)	0.700 (0.579, 0.808)	0.655 (0.594, 0.704)
5	1000	0.716 (0.668, 0.765)	0.675 (0.619, 0.734)	0.838 (0.774, 0.886)	0.698 (0.608, 0.783)	0.692 (0.637, 0.740)	0.762 (0.660, 0.857)	0.696 (0.651, 0.745)
5	2000	0.766 (0.726, 0.801)	0.726 (0.678, 0.768)	0.890 (0.842, 0.922)	0.768 (0.683, 0.830)	0.733 (0.681, 0.775)	0.808 (0.697, 0.887)	0.749 (0.707, 0.787)
10	50	0.550 (0.471, 0.677)	0.537 (0.478, 0.645)	0.575 (0.436, 0.770)	0.557 (0.385, 0.664)	0.542 (0.450, 0.667)	0.569 (0.422, 0.763)	0.541 (0.472, 0.649)
10	100	0.578 (0.415, 0.678)	0.564 (0.414, 0.648)	0.652 (0.411, 0.809)	0.554 (0.350, 0.674)	0.583 (0.429, 0.677)	0.584 (0.370, 0.767)	0.585 (0.455, 0.664)
10	200	0.667 (0.527, 0.717)	0.634 (0.505, 0.686)	0.769 (0.643, 0.840)	0.642 (0.447, 0.721)	0.653 (0.571, 0.720)	0.692 (0.528, 0.793)	0.656 (0.562, 0.708)
10	500	0.719 (0.668, 0.763)	0.685 (0.632, 0.733)	0.825 (0.706, 0.881)	0.712 (0.622, 0.783)	0.683 (0.613, 0.738)	0.746 (0.638, 0.843)	0.701 (0.636, 0.750)
10	1000	0.784 (0.747, 0.830)	0.749 (0.702, 0.802)	0.904 (0.863, 0.935)	0.796 (0.723, 0.858)	0.752 (0.708, 0.801)	0.818 (0.705, 0.897)	0.772 (0.729, 0.818)
10	2000	0.821 (0.787, 0.854)	0.792 (0.751, 0.826)	0.932 (0.889, 0.957)	0.832 (0.774, 0.887)	0.796 (0.749, 0.837)	0.850 (0.760, 0.921)	0.819 (0.783, 0.853)
20	50	0.554 (0.433, 0.680)	0.541 (0.425, 0.656)	0.582 (0.351, 0.770)	0.554 (0.430, 0.703)	0.548 (0.412, 0.652)	0.568 (0.401, 0.795)	0.547 (0.429, 0.651)
20	100	0.594 (0.388, 0.657)	0.574 (0.402, 0.634)	0.650 (0.428, 0.785)	0.596 (0.366, 0.696)	0.572 (0.449, 0.666)	0.612 (0.374, 0.754)	0.584 (0.446, 0.651)
20	200	0.661 (0.581, 0.722)	0.625 (0.545, 0.693)	0.762 (0.570, 0.838)	0.655 (0.572, 0.750)	0.641 (0.466, 0.709)	0.710 (0.596, 0.845)	0.644 (0.547, 0.703)
20	500	0.777 (0.700, 0.815)	0.740 (0.653, 0.785)	0.904 (0.812, 0.938)	0.786 (0.659, 0.845)	0.744 (0.675, 0.790)	0.814 (0.691, 0.895)	0.765 (0.673, 0.806)
20	1000	0.816 (0.774, 0.847)	0.786 (0.741, 0.821)	0.931 (0.892, 0.957)	0.823 (0.756, 0.876)	0.788 (0.744, 0.829)	0.838 (0.747, 0.921)	0.812 (0.773, 0.844)
20	2000	0.845 (0.812, 0.873)	0.819 (0.784, 0.851)	0.944 (0.904, 0.968)	0.857 (0.806, 0.902)	0.821 (0.779, 0.856)	0.866 (0.770, 0.933)	0.841 (0.810, 0.869)
50	50	0.574 (0.455, 0.680)	0.559 (0.450, 0.668)	0.613 (0.425, 0.818)	0.585 (0.433, 0.666)	0.558 (0.469, 0.687)	0.590 (0.420, 0.754)	0.560 (0.463, 0.670)
50	100	0.624 (0.408, 0.691)	0.594 (0.421, 0.660)	0.719 (0.426, 0.832)	0.605 (0.391, 0.702)	0.628 (0.467, 0.696)	0.661 (0.352, 0.793)	0.622 (0.443, 0.685)
50	200	0.699 (0.621, 0.794)	0.661 (0.590, 0.762)	0.800 (0.665, 0.921)	0.694 (0.583, 0.810)	0.667 (0.587, 0.771)	0.741 (0.629, 0.882)	0.681 (0.603, 0.780)
50	500	0.811 (0.764, 0.846)	0.779 (0.715, 0.820)	0.927 (0.879, 0.955)	0.825 (0.748, 0.888)	0.776 (0.722, 0.824)	0.843 (0.735, 0.926)	0.804 (0.756, 0.850)
50	1000	0.832 (0.796, 0.865)	0.803 (0.760, 0.844)	0.936 (0.891, 0.963)	0.844 (0.777, 0.890)	0.805 (0.759, 0.850)	0.846 (0.734, 0.923)	0.829 (0.784, 0.866)
50	2000	0.859 (0.829, 0.884)	0.837 (0.802, 0.865)	0.946 (0.906, 0.971)	0.869 (0.824, 0.909)	0.838 (0.800, 0.871)	0.869 (0.775, 0.932)	0.856 (0.822, 0.883)

AUC = area under the receiver operating characteristic curve.
The number of images represents the sum of images used for training and validation stratified by age, disease severity, and race.

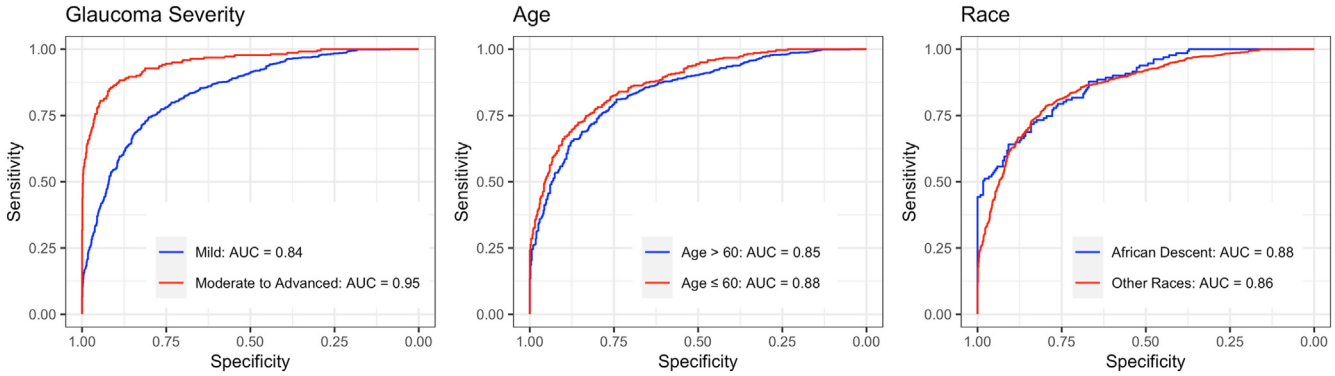

Figure 5. Generalizability of the best performing RETFound model across age, race, and severity of glaucoma. Area under the receiver operating characteristic curves are stratified by glaucoma severity (left: mild, moderate/severe), age (middle: >60 years, <60 years), and race (right: African descent, other). The best performing model was defined as the model with the highest combined (AUC), fine-tuned from a single training run. AUC = area under the receiver operating characteristic curve.

indicated by the plateauing curves from approximately 500 to 1000 images with both 20 and 50 epochs in the training set (Fig 2, S3 and S4, available at www.ophthalmologyscience.org). The diminishing returns in AUC with increasing epochs and images suggest a point of diminishing returns where additional epochs and number of images do not yield significant performance gains.

Table 4 further illustrates these trends. For 20 epochs, this constitutes a rise in AUC from 0.554 with 50 training and validation images, to 0.845 with 2000 images. The 95% CIs for AUC narrow with the increase in the number of images, indicative of higher confidence in the AUC values with larger datasets; 95% CI range is 0.214 at 50 epochs, 50 images and decreases to 0.075 at 50 epochs, 2000 images. This narrowing of 95% CIs with the rise in images and training cycles indicates better model stability and performance consistency, as more data are made available for training over greater numbers of epochs. An increase in the number of images seems to have a larger effect than a proportional increase in epochs. For 2000 training and validation images, the CI range only decreases from 0.075 at 5 epochs to 0.054 at 50 epochs. Model performance at 2000 images, over 50 epochs, is consistently excellent, with a mean AUC of 0.86.

The results also demonstrate no significant differences in performance when stratifying by race and age (Table 4, Figs 5 and 6). Comparing the 95% CI for AUCs derived from images from patients with mild glaucoma vs. moderate or advanced glaucoma, we found a significant difference at 50 epochs and 1000 or 2000 training and validation images, where the model performed significantly better on images from patients with moderate to advanced glaucoma compared with mild glaucoma (Figs 5 and 6). At 50 epochs, for 1000 images from patients with mild glaucoma mean AUC was 0.81 (0.76, 0.84), while for moderate to severe glaucoma it was 0.94 (0.86, 0.96). Similarly, at 50 epochs for 2000 images, for mild glaucoma mean AUC was 0.86 (0.80, 0.87), while for moderate to severe glaucoma mean AUC was 0.95 (0.91, 0.97).

Discussion

Although RETFound benefits from a larger volume of data and extended training, good diagnostic performance for detection of glaucoma from fundus photographs is possible with relatively small sample sizes. Specifically, as the number of images and epochs increases, there is a general trend of

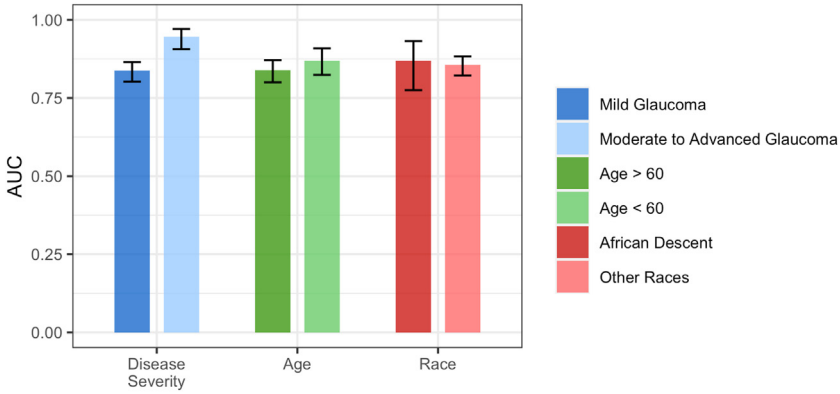

Figure 6. Bar plots of mean AUC values for 50 epochs, 2000 training/validation images, stratified by age (>60 years, <60 years), glaucoma severity (mild, moderate/severe), and race (African descent, other), with confidence intervals. AUC = area under the receiver operating characteristic curve.

improvement in diagnostic accuracy, with limited improvement after increasing the sample size past 500 images (on average 107.5 patients, 203.0 eyes) for higher numbers of epochs (20, 50) or past 1000 images (on average 215 patients, 406.0 eyes) for lower numbers of epochs (5, 10). In addition, the good diagnostic accuracy was generalizable across differences in age and race. Because this foundation model was pretrained on a large dataset using a self-supervised approach, it was able to acquire strong prior knowledge of informative retinal image features, allowing for efficient fine-tuning using smaller sample sizes and fewer epochs to achieve strong performance for our specific task of detecting glaucoma from fundus photographs.

RETFound performance equals or surpasses previously developed CNNs while being trained with significantly fewer samples.[21,27,28] When trained for 5 epochs on 2000 images, it begins to approach the performance of a previously developed CNN, ResNet-50, that had been trained on similar DIGS/ADAGES, but used significantly larger datasets, on the order of 10 000 images.[21,27,28] When trained for 10 epochs, it requires 1000 images to match prior CNN performance (CNN AUC = 0.74 [0.69, 0.79], n = 9473),[21,27,28] and surpasses prior CNNs when trained and validated on 2000 images (RETFound AUC = 0.821 [0.787, 0.857], n = 2000). When trained for 20 epochs, it surpasses prior CNNs when trained on only 500 images (RETFound AUC = 0.777 [0.705, 0.819], n = 2000).

RETFound performance also compares favorably to our previous results based on a transformer model applied to DIGS/ADAGES data.[21,27,28] On a similar DIGS/ADAGES dataset, the transformer model achieved a mean AUC (95% CI) of 0.77 (0.71, 0.82), which was comparable or worse than the results achieved by RETFound in the current study (Table 4). RETFound benefits from both increased training time and training samples but overall requires fewer labeled training samples to match or surpass the performance of prior approaches.

The findings underscore RETFound's adaptability and efficiency across a spectrum of training configurations. It demonstrates that substantial performance gains are achievable even with constrained training samples or limited computational power, situations often encountered in clinical settings. Many health care facilities grapple with the challenges of obtaining large volumes of expertly labeled data and the requisite computational infrastructure for extensive model training. RETFound's reduced reliance on extensive labeled datasets and its ability to deliver high performance across a variety of training conditions position it as a viable and innovative tool for integrating medical AI more broadly into ophthalmic practices. This study highlights the potential of using foundation models trained on large unlabeled datasets to address existing barriers to the adoption of AI technologies in a variety of settings. It offers an avenue for enhancing glaucoma detection in telehealth, primary care, community, and clinical settings.

In the original study,[17] RETFound's application to a publicly available dataset for glaucoma classification yielded equivalent or mildly superior outcomes compared with its performance following fine-tuning on the clinical DIGS/ADAGES dataset. This includes glaucoma detection on the PAPILA dataset,[29] for which they reported a mean AUC 0.86 (0.84, 0.87) and a "Glaucoma Fundus" dataset, with mean AUC 0.94 (0.94, 0.95).[17] Such a discrepancy in performance may be a result of many factors, including variations in disease severity, study population, image quality, or other factors. Numerous studies have reported high accuracy in glaucoma detection; however, direct comparisons across studies can be difficult because the disease severity is often not reported despite its apparent large impact on accuracy.[1,30] In particular, accuracy for identifying mild glaucoma is often substantially lower than identifying moderate or severe disease.[14,31] As shown in this work, where mean AUC for detecting moderate or severe glaucomatous disease rose to 0.95 (0.91, 0.97) at 2000 images and 50 training cycles, compared with 0.84 (0.80, 0.87) for detecting mild disease.

One limitation of this study is its reliance on binary classification (i.e., glaucoma vs. not glaucoma), which simplifies the complex spectrum of eye conditions. As stratification demonstrates increased performance when distinguishing severity of disease, this may suggest that a broader categorical model allowing >2 labels, as performed in the original RETFound study,[17] may aid glaucoma detection. However, a binary classification that can be used to make referral recommendations is important in implementing telehealth, screening, primary care, and clinical decision support tools. Another limitation is that the model relies solely on fundus photography; including other imaging or diagnostic data will likely enhance performance. Overall, these limitations are not necessarily specific to glaucoma detection and are indicative of broader challenges commonly faced when adopting AI-based techniques in ophthalmology.

Future work will focus on integrating OCT data into the models. Evaluating RETFound in multimodal approaches that include both fundus and OCT imaging could provide additional validation of diagnostic performance. Furthermore, expanding the validation study to encompass a broader spectrum of eye conditions, moving beyond binary glaucoma classification to include diseases with categorical labels, would significantly enhance our understanding of RETFound's versatility and relevance. Foundation AI models potentially represent an important advancement in applying AI within ophthalmology, but strong validation is required before integration into ophthalmic care.

Footnotes and Disclosures

Originally received: May 18, 2024.
Final revision: September 5, 2024.
Accepted: September 6, 2024.
Available online: September 14, 2024.Manuscript no. XOPS-D-24-00150.

[1] Hamilton Glaucoma Center, Viterbi Family Department of Ophthalmology, University of California, San Diego, La Jolla, California.

[2] School of Medicine, University of Illinois Urbana-Champaign, Urbana, Illinois.

[3] Bernard and Shirlee Brown Glaucoma Research Laboratory, Department of Ophthalmology, Harkness Eye Institute, New York, New York.

[4] Department of Ophthalmology and Vision Sciences, University of Alabama at Birmingham, Birmingham, Alabama.

Disclosures:

All authors have completed and submitted the ICMJE disclosures form.

The authors have made the following disclosures:

B.C.: Grants − UCSD MEDGAP 2023 National Institutes of Health ULTR001442

J.H.: Grants − T35: Short-Term National Research Service Award (NRSA).

J.L.: Grants − Research to Prevent Blindness, NY, NY; Consultant − Thea, Inc., Galimedix, Inc., Alcon, Inc., Carl Zeiss Meditech, Inc., ONL, Inc.; Leadership or fiduciary role in other board, society, committee or advocacy group, paid or unpaid − The Glaucoma Foundation, NY, NY.

M.A.F.: Financial support − National Eye Institute, EyeSight Foundation of Alabama, Research to Prevent Blindness, Heidelberg Engineering, GmbH, Topcon and Wolfram Research.

C.A.G.: Financial support − National Eye Institute, Topcon, EyeSight Foundation of Alabama, Research to Prevent Blindness, Heidelberg Engineering; Grants − EY018926.

R.N.W.: Financial support − Topcon Medical; Consultant − Topcon; Research equipment − Carl Zeiss Meditec, Centervue, Heidelberg Engineering, Optovue, Topcon.

M.C.: Financial support − NEI, The Glaucoma Foundation; Co-founder − AISight Health Inc.; Inventor − AISight Health Inc.; Board member − AISight Health Inc.; Shares − AISight Health Inc.

L.M.Z.: Grants − The Glaucoma Foundation, National Institutes of Health, National Eye Institute, Heidelberg Engineering; Consultant − AbbVie, Topcon Medical Systems; Patents planned, issued or pending − AISight Health; Shares − AISight Health Inc.; Others − Carl Zeiss Meditec, Icare, Optovue, Optomed, Topcon.

Supported by the National Institutes of Health, UL1TR001442 (BC), National Institutes of Health Bridge2AI common fund, grant number: OT2OD032644 (S.H.), National Eye Institute Grants R01EY027510, R01EY034146, R0111008, R01EY19869, P30EY022589, (L.M.Z.); R01EY028284 and R01EY026574 (M.A.F.); K99EY030942 and R00EY030942 (M.C.); National Institutes of Health grants (R01EY023704, R01EY029058, K12EY024225, and R01MD014850 [R.N.W.]), and a Research to Prevent Blindness unrestricted grant (R.N.W.).

HUMAN SUBJECTS: Human subjects were included in this study. The study was approved by the institutional review boards of each involved institution (University of California, San Diego, University of Alabama at Birmingham, and Columbia University). All research adhered to the tenets of the Declaration of Helsinki. All participants provided informed consent.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Chuter, Christopher, Zangwill

Data collection: Chuter, Liebmann, Fazio, Girkin, Weinreb, Zangwill

Analysis and interpretation: Chuter, Huynh, Hallaj, Walker, Liebmann, Fazio, Girkin, Weinreb, Christopher, Zangwill

Obtained funding: Zangwill, Fazio, Christopher, Chuter

Overall responsibility: Chuter, Huynh, Hallaj, Walker, Liebmann, Fazio, Girkin, Weinreb, Christopher, Zangwill

Abbreviations and Acronyms:
ADAGES = African Descent and Glaucoma Evaluation Study; **AI** = artificial intelligence; **AUC** = area under the receiver operating characteristic curve; **CDF** = cumulative density function; **CFP** = color fundus photograph; **CI** = confidence interval; **CNN** = convolutional neural network; **dB** = decibels; **DIGS** = Diagnostic Innovations in Glaucoma Study; **MD** = mean deviation; **ONH** = optic nerve head; **VF** = visual field.

Keywords:
Artificial intelligence, Fundus photographs, Glaucoma, RETFound, Self-supervised learning.

Correspondence:
Linda M. Zangwill, PhD, Professor, Richard K. Lansche MD and Tatiana A. Lansche Endowed Chair, Hamilton Glaucoma Center, Shiley Eye Institute, Viterbi Family Department of Ophthalmology -0946, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0946. E-mail: lzangwill@health.ucsd.edu.

References

1. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol*. 2019;103:167−175.
2. Christopher M, Hallaj S, Jiravarnsirikul A, et al. Novel technologies in artificial intelligence and telemedicine for glaucoma screening. *J Glaucoma*. 2024;33:S26−S32.
3. Grzybowski A, Brona P, Lim G, et al. Artificial intelligence for diabetic retinopathy screening: a review. *Eye*. 2020;34:451−460.
4. Mehta P, Petersen C, Wen JC, et al. Automated detection of glaucoma with interpretable machine learning using clinical data and multi-modal retinal images. *BioRxiv*. 2020;27:154−169.
5. Qureshi I, Ma J, Shaheed K. A hybrid proposed fundus image enhancement framework for diabetic retinopathy. *Algorithms*. 2019;12:14.
6. Zedan MJM, Zulkifley MA, Ibrahim AA, et al. Automated glaucoma screening and diagnosis based on retinal fundus images using deep learning approaches: a comprehensive review. *Diagnostics*. 2023;13:2180.
7. Abdullah F, Imtiaz R, Madni HA, et al. A review on glaucoma disease detection using computerized techniques. *IEEE Access*. 2021;9:37311−37333.
8. Thompson AC, Jammal AA, Medeiros FA. A review of deep learning for screening, diagnosis, and detection of glaucoma progression. *Transl Vis Sci Technol*. 2020;9:42.
9. Christopher M, Gonzalez R, Huynh J, et al. Proactive decision support for glaucoma treatment: predicting surgical interventions with clinically available data. *Bioengineering (Basel)*. 2024;11:140.
10. Girard MJA, Schmetterer L. Artificial intelligence and deep learning in glaucoma: current state and future prospects. *Prog Brain Res*. 2020;257:37−64.
11. Benet D, Pellicer-Valero OJ. Artificial intelligence: the unstoppable revolution in ophthalmology. *Surv Ophthalmol*. 2022;67:252−270.
12. Willemink MJ, Koszek WA, Hardell C, et al. Preparing medical imaging data for machine learning. *Radiology*. 2020;295:4−15.
13. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28:31−38.
14. Christopher M, Nakahara K, Bowd C, et al. Effects of study population, labeling and training on glaucoma detection using deep learning algorithms. *Transl Vis Sci Technol*. 2020;9:27.

15. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616:259–265.
16. He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE; 2020:9726–9735.
17. Zhou Y, Chia MA, Wagner SK, et al. A foundation model for generalizable disease detection from retinal images. *Nature*. 2023;622:156–163.
18. Sample PA, Medeiros FA, Racette L, et al. Identifying glaucomatous vision loss with visual-function-specific perimetry in the diagnostic innovations in glaucoma study. *Invest Ophthalmol Vis Sci*. 2006;47:3381–3389.
19. Sample PA, Girkin CA, Zangwill LM, et al. The african descent and glaucoma evaluation study (ADAGES): design and baseline data. *Arch Ophthalmol*. 2009;127:1136–1145.
20. Chuter B, Huynh J, Bowd C, et al. Deep learning identifies high-quality fundus photographs and increases accuracy in automated primary open angle glaucoma detection. *Transl Vis Sci Technol*. 2024;13:23.
21. Fan R, Bowd C, Christopher M, et al. Detecting glaucoma in the ocular hypertension study using deep learning. *JAMA Ophthalmol*. 2022;140:383–391.
22. Rani V, Nabi ST, Kumar M, et al. Self-supervised learning: a succinct review. *Arch Comput Methods Eng*. 2023;30:2761–2775.
23. Chen T, Kornblith S, Norouzi M, Hinton G. *A Simple Framework for Contrastive Learning of Visual Representations*. Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research, Journal of Machine Learning Research 119:1597-1607. 2020.
24. Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE; 2021:9630–9640.
25. Ye Z. SSL-DG: rethinking and fusing semi-supervised learning and domain generalization in medical image segmentation. *arXiv*. 2023. https://doi.org/10.48550/arXiv.2311.02583.
26. Tayebi Arasteh S, Misera L, Kather JN, et al. Enhancing diagnostic deep learning via self-supervised pretraining on large-scale, unlabeled non-medical images. *Eur Radiol Exp*. 2024;8:10.
27. Ajitha S, Akkara JD, Judy MV. Identification of glaucoma from fundus images using deep learning techniques. *Indian J Ophthalmol*. 2021;69:2702–2709.
28. Fan R, Alipour K, Bowd C, et al. Detecting glaucoma from fundus photographs using deep learning without convolutions: transformer for improved generalization. *Ophthalmol Sci*. 2023;3:100233.
29. Kovalyk O, Morales-Sánchez J, Verdú-Monedero R, et al. PAPILA: dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. *Sci Data*. 2022;9:291.
30. 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*. 2019;137:1353–1360.
31. Christopher M, Bowd C, Proudfoot JA, et al. Performance of deep learning models to detect glaucoma using unsegmented radial and circle OCT scans of the optic nerve head. *Invest Ophthalmol Vis Sci*. 2021;62:1014.