- 1 Title: A New Foundation Model's Accuracy in Glaucoma Detection using Ocular Coherence Tomography Images
- 2 Authors: Benton Chuter<sup>1</sup>, Justin Huynh<sup>1,2</sup>, Evan Walker<sup>1</sup>, Shahin Hallaj<sup>1</sup>, Jalil Jalili<sup>1</sup>, Jeffrey Liebmann<sup>3</sup>, Massimo
- **3** A Fazio<sup>4</sup>, Christopher A. Girkin<sup>4</sup>, Robert N. Weinreb<sup>1</sup>, Mark Christopher<sup>1</sup>, Linda M. Zangwill<sup>1</sup>
- 4 1: Hamilton Glaucoma Center, Viterbi Family Department of Ophthalmology, University of California, San Diego,
- 5 La Jolla, CA, United States
- 6 2. School of Medicine, University of Illinois Urbana-Champaign, Urbana, IL, United States.
- 7 3. Department of Ophthalmology, Harkness Eye Institute, Bernard and Shirlee Brown Glaucoma Research
- 8 Laboratory, New York, NY, United States.
- 9 4: Department of Ophthalmology and Vision Sciences, University of Alabama at Birmingham, Birmingham, AL,
- 10 United States
- 11
- 12
- 13 Running Head: Glaucoma Classification Using RETFound OCT
- 14 Corresponding Author:
- 15 Linda M. Zangwill, Ph.D.
- 16 Professor
- 17 Richard K. Lansche M.D. and Tatiana A. Lansche Endowed Chair
- 18 Hamilton Glaucoma Center
- 19 Shiley Eye Institute
- 20 Viterbi Family Department of Ophthalmology -0946
- 21 University of California, San Diego
- 22 9500 Gilman Drive
- **23** La Jolla, CA 92093-0946
- 24 Phone: (858) 534-7686
- 25 lzangwill@health.ucsd.edu
- 26

## 27 Address for reprints:

- 28 Linda M. Zangwill, Ph.D.
- 29 Hamilton Glaucoma Center
- 30 9415 Campus Point Drive
- 31 Room 175
- **32** La Jolla, CA 92037
- 33
- 34
- 35

36	Precis
37	The study found high accuracy for glaucoma detection from OCT optic nerve head RNFL scans in a diverse study
38	population by adapting an existing foundation model (RETFound). Performance improved with larger datasets and
39	more training cycles, achieving an AUC of 0.91 with RNFL scans alone. Results suggest RETFound is promising
40	for automated OCT RNFL-based glaucoma detection across demographics and training conditions.
41	
42	Abstract
43	Purpose: To fine tune and evaluate the performance of the retinal foundation model (RETFound) on a diverse
44	longitudinal clinical research dataset in glaucoma detection from optical coherence tomography (OCT) RNFL scans.
45	Subanalyses of the model performance were evaluated across different subgroups, various dataset sample sizes and
46	training cycles (epochs).
47	
48	Design: Evaluation of a diagnostic technology
49	
50	Subjects, Participants, and Controls: 15,216 Spectralis OCT RNFL circle scans of 747 individuals of diverse race
51	(56.9% White, 37.8% Black/African American, and 5.3% Other/Not reported, glaucoma severity (30.8% mild,
52	18.4% moderate-to-severe, and 50.9% no glaucoma), and age (44.8% <60 years, 55.2% >60 years) from the
53	Diagnostic Innovations in Glaucoma Study (DIGS) and the African Descent and Glaucoma Evaluation Study
54	(ADAGES). All OCT scans were labeled as "Non-glaucomatous" or "Glaucomatous."
55	
56	Methods: RETFound was employed to perform binary glaucoma classification. The diagnostic accuracy of
57	RETFound was iteratively tested across different combinations of dataset sample sizes (50 to 2000 OCT RNFL
58	circle scans), epochs (5 to 50), and study subpopulations stratified by severity of glaucoma, age, and race).
59	
60	Main Outcome Measures: Area under receiver operating characteristic curve (AUC) for classifying RNFL scans as
61	"Non-glaucomatous" or "Glaucomatous."
62	

63	Results: Performance metrics improved with larger training datasets and more training cycles, rising from an AUC
64	of 0.61 (50 training images and 5 epochs) to AUC 0.91 (2,000 training images and 50 epochs). Gains in performance
65	were marginal as training size increased beyond 500 scans. Performance was similar across race for all training size
66	and cycle number combinations: African American (AUC=0.90) vs other (AUC=0.93). RNFL scans from older
67	patients (>60 years) led to worse performance (AUC=0.85) compared to younger patients (<60 years, AUC=0.95).
68	Performance was significantly higher for RNFL scans from patients with moderate-to-severe glaucoma vs mild
69	glaucoma (AUC=0.99 vs 0.88, respectively).
70	
71	Conclusions: Good RETFound performance was observed with a relatively small sample size of images used for
72	fine tuning and across differences in race and age. RETFound's ability to adapt across a range of OCT training
73	conditions and populations suggests it is a promising tool to automate glaucoma detection in a variety of use cases.
74	
75	Keywords:
76	Foundation Model, Deep Learning, Glaucoma, Artificial Intelligence, RETFound, Self-Supervised Learning, Ocular
77	Coherence Tomography
78	

### 79 Introduction

Glaucoma, a leading cause of blindness worldwide, is characterized by progressive optic neuropathy that
can lead to irreversible vision loss if not detected and managed early.<sup>1,2</sup> Optical coherence tomography
(OCT) is fundamental in the diagnosis and monitoring of glaucoma, providing high-resolution crosssectional images of the retina essential for detecting structural abnormalities associated with various eye
conditions.<sup>3,4</sup> However, the reliance on expert clinicians to interpret OCT images poses both substantial
clinician time burden as well as issues with inter-physician variability in assessment, demonstrating the
need for reliable automated systems.<sup>5</sup>

87

88 The advent of artificial intelligence (AI) promises to revolutionize ophthalmology by addressing these issues to aid in the diagnosis and monitoring of glaucoma.<sup>6,7</sup> Recent studies have leveraged artificial 89 90 intelligence and deep learning (DL) techniques to enhance glaucoma detection using OCT images, demonstrating promising results.<sup>8-12</sup> For instance, Akter et al. employed VGG16, SqueezeNet, and 91 92 ResNet18 models on a dataset of 780 segmented and 780 raw TSNIT OCT B-scans, resulting in an AUC of 0.93 on test data.<sup>13</sup> Another multi-institutional study developed a DL model to diagnose early-onset 93 94 glaucoma using spectral-domain OCT images. Pre-trained on 4,316 OCT images from 1,371 eves with 95 open-angle glaucoma and 193 normal eyes, and then trained on a dataset from 94 patients with early 96 glaucoma and 84 normal subjects, the model achieved an AUC of 0.937, outperforming random forests and support vector machine models.<sup>14</sup> Moreover, another study evaluated various training strategies for 97 98 DL models and explored the influence of demographic and clinical factors from the study groups on the 99 efficacy of glaucoma detection from optic disc images. It contrasts the effectiveness of deep learning 100 algorithms by two independent investigators in different glaucoma populations, showing optimal 101 performance with AUCs of 0.92 for any glaucoma, 0.91 for mild glaucoma, and 0.98 for moderate-to-102 severe glaucoma across significant datasets like DIGS/ADAGES and the Matsue Red Cross Hospital 103 (MRCH) datasets, involving varied patient demographics from the U.S. and Japan.<sup>12</sup>

104	However, these DL models often require large, high-quality labeled datasets for effective training, which
105	are not universally available and are resource-intensive to produce. <sup>12,15</sup> This reliance on specialist-
106	annotated data limits their scalability and applicability in diverse clinical settings. <sup>12,15</sup>
107	
108	Self-supervised learning (SSL) helps address this labeling issue. SSL enhances data utilization by
109	extracting features without the need for ground truth labels, creating versatile feature representations for
110	various applications. <sup>16–18,19</sup> Using large pools of unlabeled data, this approach can train robust,
111	generalizable models that can be adapted for a variety of tasks and can outperform supervised learning
112	methods in classification tasks. <sup>20,21</sup> This characteristic makes SSL-based models a promising approach for
113	medical applications with limited labeled data. <sup>22,23</sup>
114	
115	Recently, an SSL-based foundation model that was trained on a large number (>1.6 million) of
116	ophthalmic images, RETFound, was described. <sup>24</sup> A primary intended use of foundation models such as
117	RETFound is to serve as a starting model that can then be adapted and/or fine-tuned to perform a specific
118	task of interest without needing a restrictively large or expensive training dataset. <sup>24</sup> Preliminary
119	evaluations of RETFound show its utility across multiple diseases, tasks, and imaging modalities.
120	However, its development and testing have been primarily confined to publicly available datasets with
121	inconsistent image and label quality and an initial dataset from the UK. To ensure its robustness and
122	applicability in real-world settings, it is crucial to further validate RETFound using larger, more diverse
123	datasets from multiple regions and demographic groups.
124	
125	To address this need, our research focuses on a validation study of RETFound using a comprehensive
126	dataset of OCT images from eyes with and without glaucoma. This study assesses RETFound's
127	performance in detecting glaucoma using OCT optic nerve head (ONH) circle scans from our unique,
128	diverse dataset. We explore the number of images and training iterations required for fine-tuning
129	RETFound to achieve high accuracy in this new context. Our investigation tests RETFound's ability to

400	1 / / 1			• •	.1 .	· · · ·	•		1	1	1 .
130	detect glaucoma	a 11ς1nσ ( )(	1 images	evamining	the im	nact of va	arving	training	durations	and	data
100	ucteet glaucome	i using OC	1 mages	, examming	s une mi		ar ymg	uanning	uurations	anu	uau

- 131 volumes. Additionally, we assess its generalizability across different ethnicities, ages, and stages of
- 132 disease to understand its performance, variability, and applicability in glaucoma detection.
- 133

134 <u>Methods</u>

- 135 Data Collection
- 136 This research used OCTs from the Diagnostic Innovations in Glaucoma Study (DIGS, clinicaltrials.gov

137 ID: NCT00221897)<sup>25</sup> and the African Descent and Glaucoma

- 138 Evaluation Study (ADAGES, clinicaltrials.gov ID: NCT00221923).<sup>26</sup> The recruitment process and
- 139 methodology were approved by the institutional review boards at each participating site, adhering to the
- 140 ethical standards outlined in the Declaration of Helsinki and the Health Insurance Portability and
- 141 Accountability Act. All subjects provided informed consent during recruitment. While comprehensive
- 142 descriptions of these studies have been presented in earlier publications,<sup>25,26</sup> the critical aspects pertinent
- 143 to this work are highlighted below.

144

145 The DIGS and ADAGES studies are a joint initiative between multiple institutions: the University of 146 California San Diego Hamilton Glaucoma Center and Viterbi Family Department of Ophthalmology, the 147 University of Alabama at Birmingham Department of Ophthalmology, and the Columbia University 148 Medical Center Edward S. Harkness Eye Institute. The participants in these studies are a diverse mix of 149 individuals with African, European, and Asian heritage. The studies' protocols involve biannual collection 150 of OCT photographs, stereo fundus images, and visual field (VF) tests as part of their ongoing research. 151 152 This analysis included a total of 15,216 Spectralis (Heidelberg Engineering GmbH, Heidelberg, 153 Germany) OCT images of 747 participants (1231 eyes), taken from 2008 to 2019. It included macula-154 centered posterior pole scans from the Glaucoma Module Premier Edition, which consisted of 61 B-scans,

each with 768 A-scans, covering a  $30^{\circ} \times 25^{\circ}$  region. Quality assessment for the SD-OCT images was

156	conducted by the University of California, San Diego Imaging Data Evaluation and Analysis Reading
157	Center following standardized protocols. Any SD-OCT images with low signal quality or those
158	containing artifacts were discarded.
159	
160	VF assessments were carried out with the Humphrey Field Analyzer II, using the 24-2 test pattern and the
161	Swedish Interactive Thresholding Algorithm standard testing algorithm. Tests with fixation losses, false-
162	negative, or false-positive errors exceeding 33% were discarded. To gauge the severity of glaucoma
163	damage at the time of imaging, the mean deviation (MD) from the VF test taken closest to the time of
164	image capture, and within a year, was used for all ONH images.
165	
166	Glaucoma Labels
167	Ground truth glaucoma status required patients to have both repeatable glaucomatous visual field damage
168	(GVFD) and glaucomatous optic neuropathy (GON). Eyes from patients who had neither GVFD or GON
169	were labeled as "non-glaucomatous." In determining GON, stereophotographs underwent review by two

- 170 independent, blinded graders using a stereoscopic viewer. If the two graders disagreed, a third
- 171 experienced grader adjudicated. GVFD was defined as a VF PSD (P< 0.05) and/or Glaucoma Hemifield
- 172 Test outside normal limits on at least two consecutive tests. Glaucomatous eyes were further categorized
- 173 into two groups based on the severity of glaucoma as indicated by 24-2 VF mean deviation (MD).
- 174 Patients with an MD of -6.0 or worse decibels (dB) were classified as having "moderate-to-severe"
- 175 glaucoma, while those with a VF MD better than -6.0 dB were categorized as having "mild" glaucoma.
- 176

## 177 Image Preprocessing

SD-OCT images were resized to a uniform 224 × 224 pixel dimension to meet the input requirements for
 RETFound. This resolution also has proven sufficient for diagnosing primary open-angle glaucoma
 (POAG) in earlier experiments.<sup>27</sup>

### 182 Evaluation of RETFound using OCTs for glaucoma label assignment

- 183 We assessed the practical use and performance of RETFound in detecting glaucoma from these
- 184 preprocessed OCT images. We conducted comprehensive iterative testing to measure RETFound's
- 185 performance, captured as the area under the receiver operating characteristic (AUC), to predict glaucoma
- 186 status. This involved training RETFound with different datasets of OCT images varying in size (50, 100,
- 187 200, 500, 1,000, 2,000) and for different numbers of training epochs (5, 10, 20, 50). Through this
- 188 approach, we aimed to determine whether RETFound could achieve or exceed the performance of other
- deep learning models in classification tasks with relatively minimal training (fine-tuning) on smaller,
- 190 labeled datasets. Additionally, we evaluated whether RETFound's results were generalizable across
- 191 differences in glaucoma severity, age, and race.
- 192

### **193** Number of Images Variation & Dataset Split

A total of 15216 images from 747 patients were randomized into training (10708 images from 512

patients), validation (1497 images from 99 patients), and test (3011 images from 212 patients) pools using

a standard 70-10-20 patient-based split (as illustrated in Figure 1). Demographic information for the entire

- dataset and for each of these pools is available in Table 1. RETFound was then evaluated across various
- dataset sizes (50, 100, 200, 500, 1,000, 2,000) and epochs (5, 10, 20, 50), totaling 24 size-epoch
- 199 combinations. These specific ranges were selected after initial testing indicated they provided a
- 200 comprehensive spectrum of model performance, from subpar to optimal.
- 201

202 For each of these size-epoch combinations, models were trained, validated, and tested on subsets

randomly sampled from the predetermined training, validation, and test pools in a 70-10-20 ratio. The

total number of images reported represented the current size for the specific size-epoch combination being

205 evaluated (number of training samples + number of validation samples = current size), as shown in Figure

- 206 1. To account for and assess variability across different training runs, each size-epoch combination
- 207 underwent 10 separate training runs, with the sampling process repeated for each. For each of these

training runs, an additional 100 bootstrap runs were performed, resulting in a total of 240 training runs
and 24,000 bootstrap runs. Further implementation details are provided in Supplementary Table 2.

- 210
- 211 Analysis

212 Given the balanced distribution of glaucoma cases within the DIGS/ADAGES dataset (as outlined in

Table 1), each run's performance was measured using the area under the receiver operating characteristic

214 (AUC). 95% confidence intervals (CIs) for these were calculated using a cumulative density function,

sorting the bootstrapped estimates and selecting the values corresponding to the 2.5th and 97.5th

216 percentiles to determine the CI range. This approach effectively captures variability between runs,

especially when the image count is high. The generalizability of the models was also assessed by

stratifying the results by race (Black vs. non-Black), age (<60 years vs. ≥60 years), and severity of

219 glaucoma (MD >-6.0 dB vs. MD  $\leq$ -6.0 dB)).

220

### 221 <u>Results</u>

222 This study included 15216 images from 747 subjects and 1232 eyes, divided into subsets for testing (3011 223 images from 212 subjects and 356 eyes), training (10708 images from 512 subjects and 812 eyes), and 224 validation (1497 images from 99 subjects and 167 eyes), as shown in Table 1. The average age of 225 participants in the study is 60.3 years, with 44.8% (n=335) under 60 years of age and 55.2% (n=412) over 226 60 years. Females (n=438, 58.6%) outnumber males (n=309, 41.4%). The majority of the study 227 population is White (56.9%, n=425), followed by Black (37.8%, n=282) and Asian (3.9%, n=29) 228 individuals. The racial status for 6(0.8%) participants was unspecified or unrecorded. Eve-level 229 characteristics such as VF MD, axial length, spherical equivalent, intraocular pressure (IOP), and central 230 corneal thickness (CCT) are documented for the training, validation, and test sets in Table 1. 30.8% 231 (n=379) of patients' eyes indicated mild glaucoma (VF MD >-6 dB) compared to 18.4% (n=226) with 232 moderate-to-severe glaucoma (VF MD  $\leq$ -6 dB), while 50.9% (n=626) had no glaucoma. The datasets are

evenly distributed between the two outcomes of interest at the latest visit: Glaucoma (n=605, 49.1%) and
Not Glaucoma (n=626, 50.9%).

235

236 Table 2, Figure 2, and Supplementary Figures S1 and S2 demonstrate the model's performance across 237 various epoch and OCT image sample size combinations. Increases in either epoch count or sample size 238 were associated with increased performance, however, sample size had a larger impact than epoch count. 239 Across all epochs, as the image count grew from 50 to 2000, the AUC also increased; at a constant of 50 240 epochs, this increase ranges from 0.64 at 50 images to 0.85 at 200 images and 0.91 at 2000 images. At a 241 constant sample size of 2,000 images, AUC increased from 0.86 at 5 epochs to 0.91 at 50 epochs. The rate 242 of improvement diminished as the sample size increased, with large gains in performance when moving 243 from 50 to 500 images and relatively small gains when adding additional images after 500. 244 245 The 95% confidence intervals for AUC narrow with the increase in the number of images, indicative of 246 higher confidence in the AUC values with larger datasets; at a constant 50 epochs, the 95% CI range is 247 0.37 at 50 epochs, 50 images and decreases to 0.12 at 50 epochs, 2000 images. This narrowing of 95% 248 CIs with the increased sample size and epoch count indicates better model stability and performance 249 consistency, as more data is available for training over greater numbers of epochs. An increase in the 250 number of images appears to have a larger effect than a proportional increase in epochs. For 2000 training 251 and validation images, the CI range only decreases from 0.14 at 5 epochs to 0.12 at 50 epochs. Model 252 performance at 2000 images, over 50 epochs, consistently achieves excellent performance with a mean 253 AUC of 0.91.

254

Model performance was also assessed in relation to demographic factors (Table 2, Figure 3 and Supplemental Figure 3) across sample sizes and epoch numbers. With respect to age groups (<60 years vs.  $\geq$ 60 years), AUCs were comparable at small sample sizes, but AUC was consistently higher in the <60 years group at larger sample sizes (0.95 (0.83 – 0.99) vs. 0.85 (0.73 – 0.93) at 2,000 images and 50

259 epochs). These differences were not statistically significant. Similarly in comparing racial groups (Black / 260 African American vs. White), results were comparable at small sample sizes, but AUC slightly in the 261 White participants at larger sample sizes  $(0.93 \ (0.84 - 0.97) \ vs. \ 0.90 \ (0.76 - 0.98)$  at 2,000 images and 50 262 epochs). Again, these differences were not statistically significant. Finally, with respect to disease 263 severity, AUC was consistently higher for the moderate-to-severe group than the early glaucoma group 264 (0.95 (0.83 - 0.99) vs. 0.85 (0.73 - 0.93) at 2,000 images and 50 epochs). These results were statistically 265 significant at the 500, 1,000, and 2,000 sample size cases across all epoch numbers (0.93 (0.84 - 0.97) vs.)266 0.90(0.76 - 0.98) at 2,000 images and 50 epochs).

267

## 268 Discussion

269 RETFound demonstrates strong diagnostic performance for OCT images, benefiting from larger datasets 270 and longer training times. Its efficiency and accuracy make it a promising tool for clinical applications. 271 The performance of RETFound varied with the number of images and epochs during training and was 272 generalizable across differences in age and race. As the number of images and epochs increases, there is a 273 general trend of improvement in diagnostic accuracy, with limited improvement in diagnostic accuracy 274 after increasing the sample size from 500 (average AUC 0.87, 25.0 patients, 40.1 eyes) to 1000 (average 275 AUC 0.90, 50.0 patients, 80.2 eyes) images. This suggests that although RETFound benefits from a larger 276 volume of data and extended training, good diagnostic performance is possible with relatively small 277 sample sizes. Likely because this foundation model is pre-trained on a large dataset using a self-278 supervised approach, it is able to acquire strong prior knowledge of informative retinal image features, 279 allowing for efficient fine-tuning using smaller sample sizes and fewer epochs to achieve strong 280 performance.

281

RETFound matches or exceeds the performance of previously developed convolutional neural networks
(CNNs) based DL methods while requiring significantly fewer samples for training.<sup>28</sup> When trained for 50
epochs on 200 images, it begins to approach the performance of ResNet-50, which was trained on much

285	larger DIGS/ADAGES OCT datasets comprising tens of thousands of images while using a similar
286	definition of glaucoma (GVFD and GON) as well as comparable distribution of severity of disease. <sup>28</sup>
287	Specifically, RETFound achieves an AUC of 0.87 (95% CI: 0.77–0.93) with just 500 images, matching
288	the performance of the prior CNN (AUC = $0.86$ , 95% CI: $0.84-0.87$ , n= $25,751$ ). <sup>28</sup> Additionally,
289	RETFound surpasses previous CNNs when trained and validated on more than 500 images, achieving an
290	AUC of 0.90 (95% CI: 0.83–0.95) for 1000 images. Even with only 20 epochs of training, RETFound
291	exceeds the performance of prior CNNs with an AUC of 0.88 (95% CI: 0.80–0.93) when trained on 1000
292	OCT images. RETFound OCT also shows comparable if not superior performance to previous CNNs
293	using ONH fundus images. Specifically, RETFound OCT achieved an AUC of 0.91 (95% CI: 0.83–0.95)
294	when trained and validated on only 2000 OCT images, similar to the AUC of 0.91 (95% CI: 0.89-0.92)
295	reported for CNNs trained on 20,828 ONH fundus images. <sup>28</sup>
296	
297	RETFound's performance also compares favorably to previous transformer models applied to the
298	DIGS/ADAGES dataset. While a transformer model achieved an AUC of 0.92 on 22,464 OCT images
299	from the DIGS/ADAGES dataset, <sup>29</sup> RETFound achieved similar results with an AUC of 0.91 (95% CI:
300	0.83-0.95) using just one-tenth of the data. This outcome demonstrates RETFound's efficiency, as it
301	requires fewer labeled training samples to match or surpass the performance of prior approaches,
302	benefiting from both increased training time and sample size.

303

In the original study,<sup>24</sup> RETFound's application to a publicly available dataset for glaucoma classification yielded comparable or mildly inferior outcomes compared to its performance following fine-tuning on the clinical DIGS/ADAGES dataset. This includes glaucoma detection on the PAPILA dataset, for which they reported a mean AUC of 0.86 (0.84, 0.87).<sup>24</sup> However, it should be noted that the original study did not directly test RETFound's performance with OCT for glaucoma prediction, only fundus images for Glaucoma and "Multi-class disease", such that a direct comparison cannot be cleanly made.

311 Further, discrepancies in performance may result from various factors, including variations in ground 312 truth definition of glaucoma disease severity, study population, image quality, or other elements.<sup>30</sup> 313 Numerous studies have reported high accuracy in glaucoma detection; however, direct comparisons 314 across studies can be difficult due to data source differences, including variations in disease severity, 315 which is often not reported despite its significant impact on accuracy.<sup>7,31</sup> In particular, accuracy for 316 identifying mild glaucoma is often substantially lower than identifying moderate-to-severe disease,<sup>12,32</sup> as 317 shown in this work, where mean AUC for detecting moderate-or-severe glaucomatous disease rose to 318 0.99 (0.95, 1.00) at 2000 images and 50 epochs, compared to 0.88 (0.79, 0.94) for detecting mild disease. 319 320 When comparing RETFound model performance when stratified for detecting mild versus moderate-to-321 severe glaucoma, significant differences are observed. At 50 epochs, the model's performance on 1000 322 images from patients with mild glaucoma shows a mean AUC of 0.87 (95% CI: 0.787 to 0.934), whereas 323 for moderate to severe glaucoma, the mean AUC significantly improves to 0.986 (95% CI: 0.950 to 324 0.999). Similarly, with 2000 images, the AUC for mild glaucoma is 0.881 (95% CI: 0.791 to 0.937), 325 compared to an AUC of 0.991 (95% CI: 0.954 to 0.999) for moderate to severe cases. Notably, even with 326 just 200 images, the model achieved an impressive AUC of 0.965 (95% CI: 0.887 to 0.992) for moderate 327 to severe glaucoma. These findings highlight the model's enhanced sensitivity in detecting more advanced 328 stages of glaucoma. This may be attributed to the more pronounced structural changes in the optic nerve 329 head and retinal nerve fiber layer in moderate-to-severe glaucoma, which are more easily recognized by 330 the model. The subtle changes in mild glaucoma may present a greater challenge for detection, requiring 331 higher image resolution or additional clinical features to improve model performance. 332 333 This study also finds that the RETFound model maintains strong performance across various 334 demographic groups, showing no statistically significant differences in AUC when stratified by race. This 335 outcome implies that the model effectively generalizes across diverse populations, a crucial trait for its 336 clinical use in different real-world scenarios. However, while not reaching statistical significance, our

337 results suggest that the RETFound model's performance may vary between age groups. Specifically, when 338 trained for 50 epochs on 2000 images the model achieved an AUC of 0.95 (95% CI: 0.83 to 0.99) for 339 subjects below 60 years of age, whereas for those above 60, the AUC was lower at 0.85 (95% CI: 0.73 to 340 (0.93). This may be due in part to differences in the severity of disease among these populations. This 341 suggests that the model is more effective at detecting glaucoma in younger patients, potentially due to 342 more pronounced retinal changes in younger individuals or differences in disease pathology. The ROC 343 curve in Figure 3 further illustrates these differences, with the curve for subjects under 60 showing higher 344 sensitivity and specificity across most thresholds compared to those over 60. This discrepancy could be 345 due to age-related changes in retinal structures that are harder to detect in older individuals or reflect 346 differences in the underlying disease pathology. These findings underscore the importance of considering 347 demographic factors such as age in the development and evaluation of AI models for medical diagnostics. 348 While the RETFound model shows promise, its varying performance across age groups highlights the 349 need for further refinement and possibly the development of age-specific models or adjustments <sup>11</sup>.

350

351 One limitation is that this study did not explicitly explore the impact of batch size on training dynamics, 352 as theoretically batch size could affect gradient smoothness and convergence stability. However, initial 353 experiments showed no significant effect on results, and relevant hyperparameters are included in 354 Supplemental Table 2. The methodology of this study is also limited by its binary classification approach, 355 distinguishing only between glaucoma and non-glaucoma. This simplification may not adequately capture 356 the nuanced spectrum of ocular diseases and the variability in normal human optic nerve head structure. 357 Enhanced diagnostic accuracy in stratifying disease severity suggests that a model with a broader range of 358 categories that includes glaucoma suspects, or glaucomatous optic nerve damage without visual field damage, could be more beneficial for glaucoma detection.<sup>24</sup> However, a binary classification system is 359 360 essential for generating referral suggestions and plays a key role in telehealth, screening, primary care, 361 and clinical decision-making tools. Relying solely on OCT RNFL imaging also has its limitations, and 362 incorporating additional imaging techniques or diagnostic information could enhance the model's

performance. Additionally, using a relatively uniform dataset in this study may limit the applicability of
the results to diverse populations. These limitations are not unique to glaucoma detection and reflect
broader challenges often faced when implementing AI-driven methods in ophthalmology.

366

367 Overall, the findings underscore RETFound's adaptability and efficiency across various training 368 configurations, demonstrating significant performance gains even with limited training samples or 369 computational resources-common constraints in real-world clinical settings. Many healthcare facilities 370 face challenges in acquiring large volumes of expertly labeled data and the necessary computational 371 infrastructure for extensive model training. Models developed externally on separate data often may 372 suffer worse performance when applied to an independent local dataset. As such, RETFound's reduced 373 dependence on extensive labeled datasets and its ability to maintain high performance across diverse 374 training conditions make it a viable and innovative tool for integrating AI into ophthalmological practices. 375 Fine-tuning enables models to be adapted to specific clinical settings, patient demographics, and disease 376 presentations, thereby optimizing their diagnostic accuracy and utility. Crucially, models can reach high 377 performance with small dataset sizes; this study suggests fine-tuning with only 500 or even 200 images 378 may be sufficient for accurate glaucoma detection. This study highlights the potential of foundational 379 models trained on large, unlabeled datasets to overcome barriers to AI adoption, enhancing glaucoma 380 detection in telehealth, primary care, community, and clinical environments.

381

Future efforts will focus on integrating fundus images as well as OCT data in models to enable a multimodal approach. By combining fundus and OCT imaging for multimodal assessment of RETFound, its diagnostic potential can be further evaluated and performance may be improved. Expanding the validation study to cover a broader spectrum of eye conditions, beyond just glaucoma, to include multiple disease categories, could also greatly enhance our understanding of RETFound's flexibility and significance. Foundational AI models have the potential to significantly advance the field of ophthalmology, but they require extensive validation before they can be implemented in clinical practice.

- 389 Furthermore, including diverse and representative datasets in these studies will be essential for evaluating
- 390 the models' real-world performance and reliability.
- 391 Bibliography
- **392** 1. Weinreb RN, Aung T, Medeiros FA. The pathophysiology and treatment of glaucoma: a review.
- **393** JAMA 2014;311:1901–1911.
- 2. Tham Y-C, Li X, Wong TY, et al. Global prevalence of glaucoma and projections of glaucoma burden
  through 2040: a systematic review and meta-analysis. Ophthalmology 2014;121:2081–2090.
- 396 3. Fujimoto JG, Drexler W, Schuman JS, Hitzenberger CK. Optical Coherence Tomography (OCT) in
   397 ophthalmology: introduction. Opt Express 2009;17:3978–3979.
- 4. Schuman JS, Hee MR, Puliafito CA, et al. Quantification of nerve fiber layer thickness in normal and
  glaucomatous eyes using optical coherence tomography. Arch Ophthalmol 1995;113:586–596.
- 5. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, et al. Artificial intelligence in retina. Prog Retin Eye
  Res 2018;67:1–29.
- 402 6. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and
  403 referral in retinal disease. Nat Med 2018;24:1342–1350.
- 404 7. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. Br J
  405 Ophthalmol 2019;103:167–175.
- 8. Akter N, Fletcher J, Perry S, et al. Glaucoma diagnosis using multi-feature analysis and a deep learning
  technique. Sci Rep 2022;12:8064.
- 9. Wu C-W, Shen H-L, Lu C-J, et al. Comparison of different machine learning classifiers for glaucoma
  diagnosis based on spectralis OCT. Diagnostics (Basel) 2021;11.
- 410 10. Ran AR, Tham CC, Chan PP, et al. Deep learning in glaucoma with optical coherence tomography: a
  411 review. Eye 2021;35:188–201.
- 412 11. Noury E, Mannil SS, Chang RT, et al. Deep Learning for Glaucoma Detection and Identification of
- 413 Novel Diagnostic Areas in Diverse Real-World Datasets. Transl Vis Sci Technol 2022;11:11.
- 414 12. Christopher M, Nakahara K, Bowd C, et al. Effects of study population, labeling and training on
  415 glaucoma detection using deep learning algorithms. Transl Vis Sci Technol 2020;9:27.
- 416 13. Akter N, Perry S, Fletcher J, et al. Glaucoma Detection and Feature Visualization from OCT Images417 Using Deep Learning. medRxiv 2023.
- 418 14. Asaoka R, Murata H, Hirasawa K, et al. Using Deep Learning and Transfer Learning to Accurately
- 419Diagnose Early-Onset Glaucoma From Macular Optical Coherence Tomography Images. Am J
- 420 Ophthalmol 2019;198:136–145.
- 421 15. Ashtari-Majlan M, Dehshibi MM, Masip D. Deep Learning and Computer Vision for Glaucoma
  422 Detection: A Review. arXiv preprint arXiv:230716528 2023.
- 423 16. Rani V, Nabi ST, Kumar M, et al. Self-supervised Learning: A Succinct Review. Arch Comput
- 424 Methods Eng 2023;30:2761–2775.

- 425 17. He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning. In:
- 426 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2020:9726–
  427 9735.
- 428 18. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual
  429 representations. International conference on machine learning 2020:1597.
- 430 19. Caron M, Touvron H, Misra I, et al. Emerging Properties in Self-Supervised Vision Transformers. In:
- 431 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE; 2021:9630–9640.
- 432 20. Ye Z. SSL-DG: Rethinking and Fusing Semi-supervised Learning and Domain Generalization in
  433 Medical Image Segmentation. arXiv preprint arXiv:231102583 2023.
- 434 21. Tayebi Arasteh S, Misera L, Kather JN, et al. Enhancing diagnostic deep learning via self-supervised
  435 pretraining on large-scale, unlabeled non-medical images. Eur Radiol Exp 2024;8:10.
- 436 22. Denner S, Zimmerer D, Bounias D, et al. Leveraging foundation models for content-based medical
  437 image retrieval in radiology. arXiv preprint arXiv:240306567 2024.
- 438 23. Pai S, Bontempi D, Hadzic I, et al. Foundation model for cancer imaging biomarkers. Nat Mach Intell
  439 2024;6:354–367.
- 24. Zhou Y, Chia MA, Wagner SK, et al. A foundation model for generalizable disease detection from
  retinal images. Nature 2023;622:156–163.
- 442 25. 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 Sci2006;47:3381–3389.
- 26. 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.
- 447 27. Fan R, Bowd C, Christopher M, et al. Detecting glaucoma in the ocular hypertension study using deep
  448 learning. JAMA Ophthalmol 2022;140:383–391.
- 449 28. Christopher M, Bowd C, Walker E, et al. Comparison of Deep Learning Glaucoma Detection Using
- 450 Optic Nerve Head Fundus Photos and Optical Coherence Tomography. Investigative Ophthalmology &451 Visual Science 2022;63.
- 452 29. Huynh J, Gonzalez R, Walker E, et al. Multimodal Transformer Model to Detect Glaucoma from
- 453 OCT and Retinal Nerve Fiber Layer (RNFL) Thickness. Investigative Ophthalmology & Visual Science
  454 2023;64:362–362.
- 455 30. Christopher M, Belghith A, Weinreb RN, et al. Retinal nerve fiber layer features identified by
- unsupervised machine learning on optical coherence tomography scans predict glaucoma progression.
  Invest Ophthalmol Vis Sci 2018;59:2748–2756.
- 458 31. Liu H, Li L, Wormstone IM, et al. Development and validation of a deep learning system to detect 459 glaucomatous optic neuropathy using fundus photographs. JAMA Ophthalmol 2019;137:1353–1360.
- 460 32. Christopher M, Bowd C, Proudfoot JA, et al. Performance of Deep Learning Models to Detect
- 461 Glaucoma Using Unsegmented Radial and Circle OCT Scans of the Optic Nerve Head. Investigative
- 462 Ophthalmology & Visual Science 2021;62:1014–1014.

- 463
- 464 **Figures**
- 465 Figure 1: Diagram showing model workflow, involving OCT preprocessing, fine-tuning runs, &
- 466 bootstraps.



467

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



472

- 473 Figure 3: Best performing RETFound models for age, race, and severity of glaucoma. Area under the receiver
- 474 operating characteristic curve (AUC) curves are stratified by Glaucoma Severity (left: Mild, Moderate-to-severe),
- 475 Age (middle: >60y, <60y) and Race (right: African Descent, Other). The best performing model was defined as the
- 476 model with the highest combined (AUC), fine-tuned from a single training run with 2000 images.





#### 479 Tables

#### 480 Table 1: Overview of all Cohorts

	Overall ( $n = 747$ subjects; 1231 eyes; 15216 images)	Training $(n = 512)$ subjects; 812 eyes; 10708 images)	Validation (n = 99 subjects; 167 eyes; 1497 images)	Testing $(n = 212)$ subjects; 356 eyes; 3011 images)
Age	60.3 (59.2, 61.4)	61.7 (60.4, 63.0)	54.8 (51.4, 58.2)	58.1 (56.1, 60.1)
Age Classification				
Age < 60	335 (44.8%)	205 (40.0%)	60 (60.6%)	114 (53.8%)
Age > 60	412 (55.2%)	307 (60.0%)	39 (39.4%)	98 (46.2%)
Sex				
Female	438 (58.6%)	282 (55.1%)	65 (65.7%)	134 (63.2%)
Male	309 (41.4%)	230 (44.9%)	34 (34.3%)	78 (36.8%)
Race				
American Indian/ Alaska Native	2 (0.3%)	1 (0.2%)	0 (0.0%)	1 (0.5%)
Asian	29 (3.9%)	23 (4.5%)	3 (3.0%)	3 (1.4%)
Black or African American	282 (37.8%)	185 (36.1%)	31 (31.3%)	104 (49.1%)
Native Hawaiian or Other Pacific Islander	3 (0.4%)	3 (0.6%)	0 (0.0%)	0 (0.0%)
Unknown or Not Reported	6 (0.8%)	4 (0.8%)	2 (2.0%)	0 (0.0%)
White	425 (56.9%)	296 (57.8%)	63 (63.6%)	104 (49.1%)
Ethnicity				
Hispanic	18 (2.4%)	13 (2.5%)	5 (5.1%)	3 (1.4%)
Not Hispanic	647 (86.6%)	446 (87.1%)	82 (82.8%)	181 (85.4%)
Unknown or Not Reported	82 (11.0%)	53 (10.4%)	12 (12.1%)	28 (13.2%)
Diabetes				
No	652 (87.3%)	437 (85.4%)	93 (93.9%)	189 (89.2%)
Yes	95 (12.7%)	75 (14.6%)	6 (6.1%)	23 (10.8%)
Hypertension				
No	463 (62.0%)	299 (58.4%)	68 (68.7%)	142 (67.0%)
Yes	284 (38.0%)	213 (41.6%)	31 (31.3%)	70 (33.0%)
24-2 VF MD (dB)	-3.31 (-3.71, -2.92)	-4.10 (-4.62, -3.59)	-1.61 (-2.42, -0.80)	-1.08 (-1.54, -0.63)
Baseline Disease Severity				
Mild Glaucoma	379 (30.8%)	313 (37.9%)	25 (15.0%)	41 (11.5%)
Moderate-to-severe Glaucoma	226 (18.4%)	195 (23.6%)	13 (7.8%)	18 (5.1%)
Non-Glaucomatous	626 (50.9%)	317 (38.4%)	129 (77.2%)	297 (83.4%)
Axial Length (mm)	23.9 (23.9, 24.0)	24.0 (23.9, 24.1)	23.7 (23.5, 24.0)	23.8 (23.7, 24.0)
Spherical Equivalent	-0.51 (-0.64, -0.38)	-0.58 (-0.74, -0.42)	-0.58 (-0.90, -0.26)	-0.31 (-0.55, -0.07)
IOP (mmHg)	14.7 (14.4, 14.9)	14.7 (14.4, 15.1)	14.6 (14.1, 15.2)	14.6 (14.2, 15.0)
CCT (µm)	539 (536, 542)	539 (535, 542)	544 (536, 552)	538 (533, 543)
Baseline Visit Glaucoma Classification				
GVFD & GON	605 (49.1%)	508 (61.6%)	38 (22.8%)	59 (16.6%)
Non-glaucomatous	626 (50.9%)	317 (38.4%)	129 (77.2%)	297 (83.4%)
Last Visit Glaucoma Classification				
GVFD & GON	605 (49.1%)	508 (61.6%)	38 (22.8%)	59 (16.6%)
Non-glaucomatous	626 (50.9%)	317 (38.4%)	129 (77.2%)	297 (83.4%)

IOP: intraocular pressure, CCT: central corneal thickness, VF: visual field, MD: mean deviation

481 482 \*Number of patients when stratified by characteristic may not sum to total (2104 for "All"); this remainder were unreported for the characteristic 483 of interest.

484 \*\*Age of 5 subjects progressed past 60y during the study; these are here reported at baseline age <60y.

- 486 Table 2: Summary of the model performance on the local datasets as captured by AUC, with 95% confidence
- 487 intervals. The number of images represents the sum of images used for training and validation. Stratified by Age,

### 488 Disease Severity, and Race.

Epoch	Number	Overall	Age		Disease Severity		Race		
	of	(n = 212	Age Below 60 Age Above 60		Mild Glaucoma	Moderate-to-severe	Black or African	Other Races	
	Images	subjects; 356	(n = 114 subjects;	(n = 101 subjects;	(n = 29 subjects;	Glaucoma	American	(n = 108 subjects;	
		eyes; 3011	196 eyes; 1421	164 eyes; 1590	43 eyes; 666	(n = 19 subjects; 27	(n = 104 subjects; 172	184 eyes; 1586	
		images)	images)	images)	images)	eyes; 221 images)	eyes; 1425 images)	images)	
5	50	0.61 (0.44, 0.70)	0.59 (0.37, 0.69)	0.61 (0.50, 0.71)	0.60 (0.44, 0.69)	0.64 (0.41, 0.76)	0.61 (0.40, 0.75)	0.61 (0.43, 0.71)	
5	100	0.65 (0.56, 0.73)	0.60 (0.42, 0.71)	0.65 (0.57, 0.74)	0.64 (0.55, 0.72)	0.70 (0.55, 0.79)	0.63 (0.52, 0.76)	0.66 (0.57, 0.76)	
5	200	0.71 (0.64, 0.79)	0.67 (0.57, 0.79)	0.68 (0.59, 0.77)	0.69 (0.61, 0.76)	0.80 (0.71, 0.88)	0.70 (0.59, 0.83)	0.72 (0.63, 0.79)	
5	500	0.80 (0.71, 0.87)	0.78 (0.67, 0.88)	0.75 (0.62, 0.85)	0.76 (0.67, 0.83)	0.93 (0.85, 0.97)	0.80 (0.66, 0.92)	0.80 (0.70, 0.87)	
5	1000	0.84 (0.76, 0.91)	0.81 (0.64, 0.93)	0.79 (0.67, 0.89)	0.80 (0.72, 0.88)	0.96 (0.91, 0.99)	0.82 (0.67, 0.94)	0.85 (0.77, 0.91)	
5	2000	0.86 (0.78, 0.92)	0.84 (0.69, 0.96)	0.80 (0.68, 0.89)	0.82 (0.74, 0.90)	0.97 (0.93, 0.99)	0.84 (0.69, 0.95)	0.87 (0.79, 0.93)	
10	50	0.62 (0.44, 0.72)	0.60 (0.37, 0.70)	0.62 (0.51, 0.73)	0.61 (0.44, 0.71)	0.66 (0.41, 0.79)	0.61 (0.41, 0.75)	0.62 (0.44, 0.74)	
10	100	0.72 (0.63, 0.81)	0.68 (0.53, 0.83)	0.68 (0.57, 0.77)	0.69 (0.60, 0.77)	0.83 (0.67, 0.93)	0.71 (0.58, 0.87)	0.73 (0.63, 0.81)	
10	200	0.80 (0.71, 0.87)	0.78 (0.65, 0.90)	0.74 (0.59, 0.85)	0.75 (0.66, 0.84)	0.93 (0.83, 0.97)	0.79 (0.64, 0.91)	0.80 (0.68, 0.88)	
10	500	0.85 (0.78, 0.91)	0.82 (0.64, 0.93)	0.79 (0.67, 0.88)	0.81 (0.73, 0.88)	0.96 (0.91, 0.99)	0.83 (0.69, 0.93)	0.87 (0.79, 0.92)	
10	1000	0.86 (0.79, 0.92)	0.85 (0.70, 0.96)	0.80 (0.69, 0.90)	0.83 (0.74, 0.90)	0.97 (0.92, 0.99)	0.84 (0.71, 0.96)	0.89 (0.80, 0.94)	
10	2000	0.89 (0.81, 0.94)	0.89 (0.71, 0.97)	0.82 (0.71, 0.92)	0.85 (0.78, 0.92)	0.98 (0.93, 1.00)	0.87 (0.74, 0.97)	0.91 (0.82, 0.96)	
20	50	0.69 (0.46, 0.79)	0.64 (0.41, 0.76)	0.67 (0.51, 0.77)	0.67 (0.45, 0.76)	0.77 (0.43, 0.89)	0.68 (0.45, 0.83)	0.71 (0.44, 0.80)	
20	100	0.76 (0.64, 0.85)	0.72 (0.57, 0.86)	0.72 (0.59, 0.82)	0.72 (0.62, 0.81)	0.88 (0.68, 0.96)	0.75 (0.61, 0.89)	0.77 (0.63, 0.87)	
20	200	0.82 (0.73, 0.90)	0.82 (0.67, 0.92)	0.76 (0.64, 0.88)	0.78 (0.68, 0.87)	0.96 (0.89, 0.99)	0.81 (0.66, 0.93)	0.83 (0.72, 0.91)	
20	500	0.85 (0.78, 0.91)	0.83 (0.67, 0.94)	0.80 (0.68, 0.89)	0.82 (0.73, 0.89)	0.97 (0.92, 0.99)	0.84 (0.69, 0.94)	0.87 (0.79, 0.93)	
20	1000	0.88 (0.80, 0.93)	0.88 (0.72, 0.97)	0.81 (0.69, 0.90)	0.84 (0.76, 0.91)	0.98 (0.92, 01.00)	0.85 (0.70, 0.95)	0.90 (0.83, 0.96)	
20	2000	0.90 (0.82, 0.94)	0.91 (0.76, 0.98)	0.84 (0.73, 0.92)	0.87 (0.78, 0.93)	0.99 (0.94, 1.00)	0.88 (0.73, 0.97)	0.93 (0.85, 0.97)	
50	50	0.64 (0.43, 0.80)	0.63 (0.40, 0.81)	0.64 (0.48, 0.80)	0.62 (0.43, 0.76)	0.71 (0.42, 0.94)	0.64 (0.41, 0.85)	0.64 (0.41, 0.81)	
50	100	0.77 (0.48, 0.84)	0.74 (0.38, 0.88)	0.72 (0.57, 0.84)	0.73 (0.49, 0.81)	0.88 (0.41, 0.95)	0.75 (0.46, 0.89)	0.78 (0.48, 0.86)	
50	200	0.85 (0.77, 0.91)	0.85 (0.71, 0.96)	0.79 (0.66, 0.89)	0.81 (0.72, 0.88)	0.97 (0.89, 0.99)	0.83 (0.70, 0.94)	0.87 (0.78, 0.93)	
50	500	0.87 (0.77, 0.93)	0.87 (0.72, 0.98)	0.81 (0.67, 0.90)	0.83 (0.72, 0.91)	0.98 (0.93, 1.00)	0.85 (0.69, 0.95)	0.89 (0.76, 0.95)	
50	1000	0.90 (0.83, 0.95)	0.92 (0.80, 0.99)	0.84 (0.72, 0.92)	0.87 (0.79, 0.93)	0.99 (0.95, 1.00)	0.89 (0.75, 0.97)	0.92 (0.83, 0.97)	
50	2000	0.91 (0.83, 0.95)	0.95 (0.83, 0.99)	0.85 (0.73, 0.93)	0.88 (0.79, 0.94)	0.99 (0.95, 1.00)	0.90 (0.76, 0.98)	0.93 (0.84, 0.97)	

# 490 Supplemental Figures

- 491 Supplemental Figure 1: Plots demonstrating the relationship between number of images (x) vs and performance, as
- 492 measured by AUC (y), at each tested epoch number. Dashed lines correspond to the range of 95% Cis.



- 495 Supplemental Figure 2: Plots of the relationship between number of epochs (x) and diagnostic performance, as
- 496 measured by area under the receiver operating characteristic curve (AUC, y), at each tested dataset sample size.
- 497 Dashed lines correspond to the range of 95% Cis.



498

- 500 Supplemental Figure 3: Bar Plots of mean area under the receiver operating characteristic curve (AUC)
- values for 50 epochs, 2000 training/validation images Stratified by Age (<60 years, ≥60 years), glaucoma
- 502 severity (Mild, Moderate-to-severe) and race (African Descent, Other), with confidence intervals.



504

# 506 Supplemental Tables

## 507 Supplemental Table 1

	Disease Severity		А	ge	Race		
	Mild Glaucoma	Moderate to	Age < 60 (n =	Age > 60 (n =	AD (n = 282	Other Race (n =	
	(n = 258	Advanced	335 subjects; 576	412 subjects; 655	subjects; 454	465 subjects; 777	
	subjects; 379 eyes)	Glaucoma (n = 140 subjects; 226 eves)	eyes)	eyes)	eyes)	eyes)	
Age	67.7 (66.2, 69.1)	67.0 (64.9, 69.1)	46.3 (45.1, 47.4)	71.7 (71.0, 72.4)	59.8 (58.2, 61.4)	60.6 (59.0, 62.1)	
Age Classification							
Age < 60	60 (23.3%)	36 (25.7%)	335 (100.0%)	0 (0.0%)	142 (50.4%)	193 (41.5%)	
Age > 60	198 (76.7%)	104 (74.3%)	0 (0.0%)	412 (100.0%)	140 (49.6%)	272 (58.5%)	
Sex							
Female	146 (56.6%)	67 (47.9%)	198 (59.1%)	240 (58.3%)	179 (63.5%)	259 (55.7%)	
Male	112 (43.4%)	73 (52.1%)	137 (40.9%)	172 (41.7%)	103 (36.5%)	206 (44.3%)	
Race							
American Indian/ Alaska Native	0 (0.0%)	1 (0.7%)	2 (0.6%)	0 (0.0%)	0 (0.0%)	2 (0.4%)	
Asian	10 (3.9%)	12 (8.6%)	17 (5.1%)	12 (2.9%)	0 (0.0%)	29 (6.2%)	
Black or African American	89 (34.5%)	48 (34.3%)	142 (42.4%)	140 (34.0%)	282 (100.0%)	0 (0.0%)	
Native Hawaiian or Other Pacific Islander	2 (0.8%)	0 (0.0%)	1 (0.3%)	2 (0.5%)	0 (0.0%)	3 (0.6%)	
Unknown or Not Reported	2 (0.8%)	0 (0.0%)	5 (1.5%)	1 (0.2%)	0 (0.0%)	6 (1.3%)	
White	155 (60.1%)	79 (56.4%)	168 (50.1%)	257 (62.4%)	0 (0.0%)	425 (91.4%)	
Ethnicity							
Hispanic	5 (1.9%)	3 (2.1%)	12 (3.6%)	6 (1.5%)	4 (1.4%)	14 (3.0%)	
Not Hispanic	237 (91.9%)	123 (87.9%)	267 (79.7%)	380 (92.2%)	267 (94.7%)	380 (81.7%)	
Unknown or	16 (6.2%)	14 (10.0%)	56 (16.7%)	26 (6.3%)	11 (3.9%)	71 (15.3%)	
Not Reported							
No	207 (80.2%)	120 (85.7%)	306 (91.3%)	346 (84.0%)	218 (77 3%)	434 (93 3%)	
Yes	51 (19.8%)	20 (14 3%)	29 (8 7%)	66 (16 0%)	64 (22.7%)	31 (6 7%)	
Hypertension	51 (1)10/0)	20 (111570)	25 (01770)	00 (10:070)	01 (221770)	51 (01770)	
No	124 (48 1%)	68 (48 6%)	260 (77.6%)	203 (49 3%)	151 (53.5%)	312 (67 1%)	
Yes	134 (51.9%)	72 (51.4%)	75 (22.4%)	209 (50.7%)	131 (46 5%)	153 (32.9%)	
24-2 VF MD (dB)	-2.40 (-2.83, -	-13.52 (-14.08, -	-1.84 (-2.41, -	-4.53 (-5.05, -	-3.34 (-3.99, -	-3.30 (-3.80, -	
Baseline Disease	1.96)	12.90)	1.28)	4.01)	2.09)	2.80)	
Severity Mild Glaucoma	379 (100.0%)	0 (0.0%)	89 (15.5%)	290 (44.3%)	131 (28.9%)	248 (31.9%)	
Moderate to	0 (0.0%)	226 (100.0%)	55 (9.5%)	171 (26.1%)	71 (15.6%)	155 (19.9%)	
Advanced Glaucoma	0 (01070)	220 (100.070)	00 (0.070)	1,1 (2011/0)	(1010/0)	100 (191970)	
Non-	0 (0.0%)	0 (0.0%)	432 (75.0%)	194 (29.6%)	252 (55.5%)	374 (48.1%)	
Glaucomatous Axial Length (mm)	24.1 (23.9,	24.1 (24.0, 24.3)	24.0 (23.8, 24.1)	23.9 (23.8, 24.0)	23.8 (23.7, 23.9)	24.0 (23.9, 24.1)	
Spherical Fouivalent	-0.4 (-0.6, -0.2)	-0.8 (-1.0, -0.5)	-1.01 (-1.20, -	-0.10 (-0.27, 0.07)	-0.22 (-0.43, -	-0.69 (-0.85, -	
IOP (mmHg)	15.3 (14.8,	13.3 (12.7, 13.9)	14.9 (14.5, 15.3)	14.5 (14.1, 14.8)	15.2 (14.8, 15.6)	14.4 (14.0, 14.7)	
CCT (µm)	536 (532, 541)	532 (527, 537)	544 (539, 548)	536 (532, 540)	531 (526, 536)	544 (540, 548)	
Baseline Visit Glaucoma Classification							
GVFD & GON	379 (100.0%)	226 (100.0%)	144 (25.0%)	461 (70.4%)	202 (44.5%)	403 (51.9%)	
Non- glaucomatous	0 (0.0%)	0 (0.0%)	432 (75.0%)	194 (29.6%)	252 (55.5%)	374 (48.1%)	
Last Visit Glaucoma Classification							
GVFD & GON	379 (100.0%)	226 (100.0%)	144 (25.0%)	461 (70.4%)	202 (44.5%)	403 (51.9%)	
Non- glaucomatous	0 (0.0%)	0 (0.0%)	432 (75.0%)	194 (29.6%)	252 (55.5%)	374 (48.1%)	

# 509 Supplemental Table 2

Parameter	Train Command Value
Processing Unit	A40
Mode	Training
Optimizer	AdamW
nproc_per_node	1
batch_size	16
world_size	1
model	vit_large_patch16
epochs	EPOCH
blr	5e-3
layer_decay	0.65
weight_decay	0.05
drop_path	0.2
nb_classes	2
input_size	244