

EyeScreen

Development and Potential of a Novel Machine Learning Application to Detect Leukocoria

Alec Bernard, BS, MS,¹ Shang Zhou Xia, BS,² Sahal Saleh, MD,¹ Tochukwu Ndukwe, MD,¹ Joshua Meyer, BS,² Elliot Soloway, PhD,² Mandefro Sintayehu, MD,³ Blen Teshome Ramet, MD,⁴ Bezawit Tadegegne, MD,³ Christine Nelson, MD, FACS,⁵ Hakan Demirci, MD⁵

Purpose: Early diagnosis and treatment of retinoblastoma are of paramount importance for a positive clinical outcome. The most common sign of retinoblastoma is leukocoria, or white pupil. Effective, easy-to-perform, community-based screening is needed to improve outcomes in lower-income regions. The EyeScreen (developed by Joshua Meyer from the University of Michigan) Android (Google LLC) smartphone application is an important step toward addressing this need. The purpose of this study was to examine the potential of the novel use of low-cost technologies—a cell phone application and machine learning—to identify leukocoria.

Design: A cell phone application was developed and refined with the feedback from on-site, single-population use in Ethiopia. Application performance was evaluated in this technology validation study.

Participants: One thousand four hundred fifty-seven participants were recruited from ophthalmology and pediatric clinics in Addis Ababa, Ethiopia.

Methods: Photographs obtained with inexpensive Android smartphones running the EyeScreen Application were used to train an ImageNet (ResNet) machine learning model and to measure the performance of the app. Eighty percent of the images were used in training the model, and 20% were reserved for testing.

Main Outcome Measures: Performance of the model was measured in terms of sensitivity, specificity, receiver operating characteristic (ROC) curve, and precision-recall curve.

Results: Analyses of the participant images resulted in the following at the participant level: sensitivity, 87%; specificity, 73%; area under the ROC curve, 0.93; and area under the precision-recall curve, 0.77.

Conclusions: EyeScreen has the potential to serve as an effective screening tool in the areas of the world most affected by delayed retinoblastoma diagnosis. The relatively high initial performance of the machine learning model with small training datasets in this early-phase study can serve as a proof of concept for future use of machine learning and artificial intelligence in ophthalmic applications. *Ophthalmology Science* 2022;2:100158 © 2022 by 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/>).

Retinoblastoma is the most common intraocular cancer in children, with an annual incidence of approximately 9000 new cases globally.¹ Mortality rates are highly disparate in different parts of the globe, and areas with the highest prevalence also show the highest mortality rates. In Africa and Asia, 40% to 70% of children with retinoblastoma die, compared with 3% to 5% of children in Europe and North America.^{2,3} Early detection and treatment are essential for effective treatment of retinoblastoma before the tumor spreads beyond the eye and the chances of survival decrease.⁴ Leukocoria, or white pupil, is the most common presenting sign of retinoblastoma, and this is the sign initially detected by family or friends in 80% of patients.⁵ The interval from family finding of leukocoria to the pediatrician examination ranged from 1 to 42 months, with a mean of 2 months, and the interval from pediatrician examination to ocular oncology examination was from 1 to 29 months, with a mean of 2 months.⁶ This

delay is longer in Africa and Asia and contributes to much of the disparity in survival in these regions.⁷ Poor survival rates are correlated directly with delay in diagnosis and treatment abandonment. Although treatment abandonment is associated with numerous socioeconomic, cultural, and educational factors, evidence exists that diagnostic delay can be mitigated with effective screening programs.^{3,8,9}

Current screening tools have significant drawbacks. Red reflex testing is the current standard; it relies on the skill of the screening personnel and has variable sensitivity, ranging from 85% to <5%, depending on the location of pathologic features and training of the providers.^{10,11} Newer, technology-based tools have a need for ophthalmologists or trained technicians and are limited in their widespread application.⁸ These tools include a miniature direct ophthalmoscope, a portable fundus camera, and a clip-on fundus camera. In addition to the need for trained

providers, these tools suffer from lack of wide accessibility and cost factors.⁸ Currently, 2 applications exist that are designed to screen for leukocoria, but they rely on subjective assessments or have low sensitivity and specificity and additionally are designed for use by parents and not for real-time screening.¹²

A need exists for simple, inexpensive, and widely available screening technologies to improve outcomes in this needlessly deadly disease. Smartphone applications are a logical choice for this screening method because of their widespread availability and ease of use. Few data have been published regarding the reliability of identifying leukocoria with smartphone cameras, and no data are available regarding the use of an application designed to detect leukocoria by combining data from multiple directions of gaze. The literature also shows that evaluation of the red reflex in multiple gazes, in addition to primary gaze, increases the detection of leukocoria.¹³ Building on existing tools for screening for retinoblastoma and the anecdotal evidence suggesting that flash photography has enabled early detection, we developed a novel smartphone application for use in screening populations for retinoblastoma in multiple directions of gaze.¹⁴ This application is called EyeScreen (developed by Joshua Meyer from the University of Michigan) and was tested in Addis Ababa, Ethiopia, for accuracy and feasibility. The objective of this study was to determine the performance of the model in identifying leukocoria, measured in terms of sensitivity, specificity, receiver operating characteristic (ROC) curve, and precision-recall curve.

Methods

This study was approved by the University of Michigan Institutional Review Board and by the institutional review board of the St. Paul's Hospital Millennium Medical College in Addis Ababa, Ethiopia (identifier: HUM00090656). Informed consent was obtained from all participants, and all research adhered to the tenets of the Declaration of Helsinki.

The EyeScreen software is a smartphone application designed for use with Android devices (Google LLC). The EyeScreen application was installed in Google Pixel 3a phones provided to the researchers for the duration of this study. Participants were recruited from St. Paul's Hospital Millennium Medical College, a crowded and busy hospital in Addis Ababa, Ethiopia. Researchers were located in the general ophthalmology clinic, pediatric ophthalmology clinic, neonatal intensive care unit, emergency department, and general pediatric clinic. These locations varied in lighting, crowding, privacy, and other factors. All patients were eligible to participate if they, or their parent if younger than 18 years, were able to give informed consent.

Consent was obtained from English-speaking participants and their parents for underage participants. A local translator was used for patients who spoke Amharic. Patients were seated, and lights were dimmed as much as possible (sometimes requiring sheets over the windows). Four different directions of gaze were captured for each participant: up gaze, left gaze, right gaze, and center gaze. Small stuffed animals (beanie babies) were used to assist in directing gaze for younger participants. The initial study size target was 1200 patients because of recruitment considerations, but the nature of the machine learning model is such that the more data

run through the training program, the more sensitive the test is likely to be.

The development of EyeScreen was highly iterative and technically challenging, responding to issues arising in the clinical setting. For example, current smartphone cameras use a preemptive flash technique in which bursts of light are sent out before the photograph is taken. This action constricts pupils to remove the common red-eye effect; however, in our situation, the preemptive flash technique is counterproductive because a solid red reflex has a definite impact on subsequent analysis. To capture the eye in a maximum naturally dilated (without chemical dilation) state with the red reflex present, we needed first to find the preflash setting that was buried deep in the system software. We did find a way to disable the preflash light, thereby avoiding the issue of pupillary constriction. Additionally, having the patient in a darker environment assisted in obtaining a clearer view of the red reflex. But, limitations in on-site lighting conditions required still further adjustments to the light-processing balance in the application software. Still further, a lack of reliable wireless or cellular connection required the application to be updated to store photographs temporarily until an upload opportunity became available.

To aid the picture taker—and to standardize the distance the cell phone was from a participant—EyeScreen displayed boxes around the participant's eyes when they were detected to be the proper distance. The boxes were a sign to the picture taker to take the picture. The image of the participant's eyes was then displayed on the screen. EyeScreen provided the picture taker with the option to accept the pictures and move to the next direction of gaze or to retake the image (Fig 1). This process allows for multiple attempts at each gaze if the participant is moving or the image is blurry. The user interface for the Android application displayed helpful tips for optimal photography, and the system required no specialized training (Fig 2).

Participants' age, race or ethnicity, sex, and presence of ocular conditions were recorded and attached to their images. In the ophthalmology clinics, these ophthalmic conditions were obtained from the patient's chart. All collected patient information was deidentified.

The images underwent automatic image processing within the application to bound the eyes in boxes to preserve participant privacy. The images of the eyes alone then were uploaded to a secure server at the University of Michigan. The images then were reviewed by an ocular oncologist (H.D.) and classified into simple categories (normal vs. abnormal or leukocoria; Fig 3). Eyes with leukocoria were assigned the label "abnormal." During the labeling process, very poor quality and ungradable images were removed (i.e., images in which the pupil was not visible). These images then were used to fine-tune a pretrained machine learning model. Each image underwent preprocessing, including resizing to 224×224 and normalizing to the dataset mean. Then, each image was augmented by rotating 90° , 180° , and 270° ; adjusting saturation, sharpness, brightness, contrast, and gamma; and adjusting hue by 0.5 and -0.5 (polar opposites of color hue). Additionally, images were partitioned by patient using unique identifiers recorded at the time of image capture. All testing was carried out at the level of the participant, with all gaze directions tested together for each participant. This is more clinically relevant than testing by individual photograph. Our application uses an ImageNet model, specifically, ResNet, an open-source deep learning network developed for use in image processing applications.¹⁵ This model was selected because of its performance in image processing and to assess the potential of open-source, free models. Our model uses only unstructured data. We used a pretrained ImageNet model and retrained the model using our dataset. The ImageNet deep learning model was trained with 80% of the images and tested on

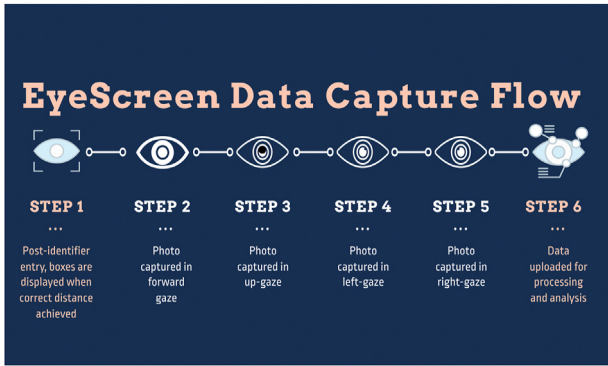


Figure 1. Diagram showing basic screening steps in use of the EyeScreen application.

the remaining 20%. Because of the imbalance in our dataset toward normal eyes, we selected a threshold to maximize sensitivity at the expense of specificity. These images were randomized, and training and testing sets also were separated at the patient level. The batch size was 100, the learning rate was 3×10^{-6} , and we used an Adam optimizer. After 10 epochs, we saw that the

accuracy, sensitivity, and specificity did not change by a significant amount. We continued to train for 100 epochs, and the model did not diverge. Cross-validation was not able to be performed with the small amount of positive cases.

Results

More than 4000 images of eyes were obtained, approximately 4 per participant. [Table 1](#) shows demographic information of participants and the distribution of normal red reflex and abnormal red reflex eyes, separated by training and testing datasets at the level of the participant and including image counts.

Eighty percent of the participant images were used in training the model with multiple ResNet training processes completed. One hundred fifty iterations were completed before the model accuracy converged to a stable value. The remaining 20% of the images were used to test the accuracy of the model. Sample images used in testing are shown in [Figure 4](#).

The model using the testing set of images from 291 participants showed sensitivity of 87% and specificity of 73%. The ROC curve and area under the ROC curve for

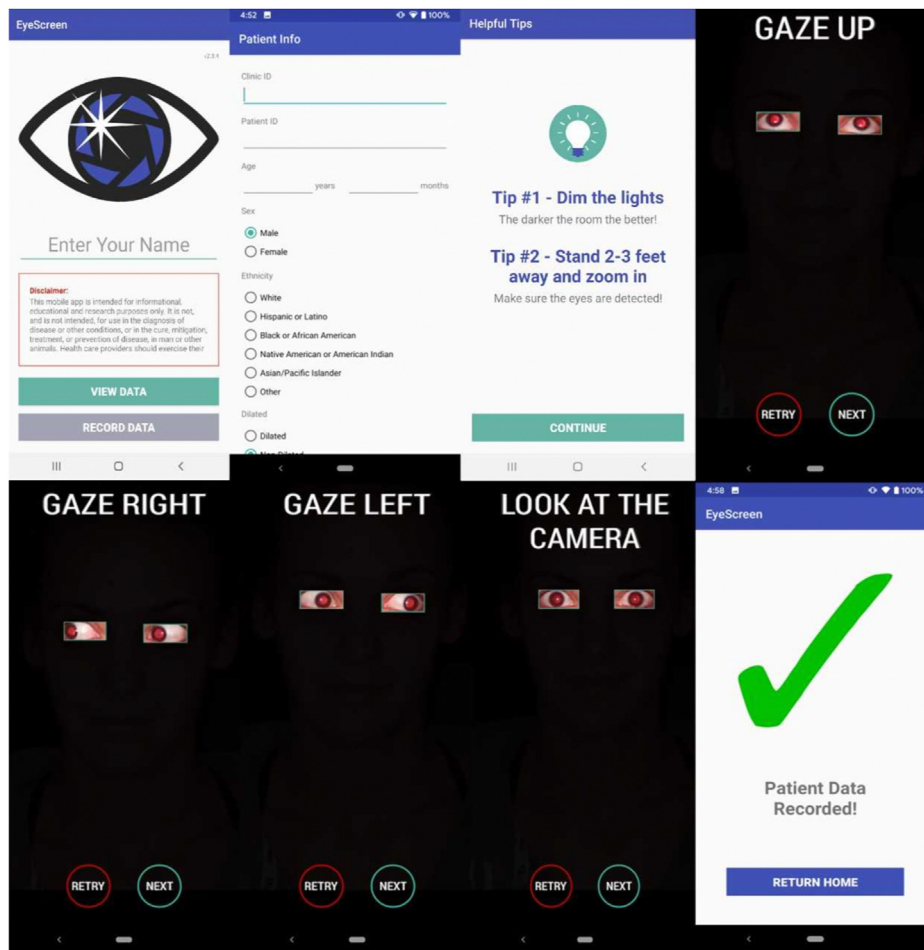


Figure 2. Screenshots obtained by the authors showing the logo and user interface for the Android EyeScreen app (developed by Joshua Meyer from University of Michigan), before taking photographs.

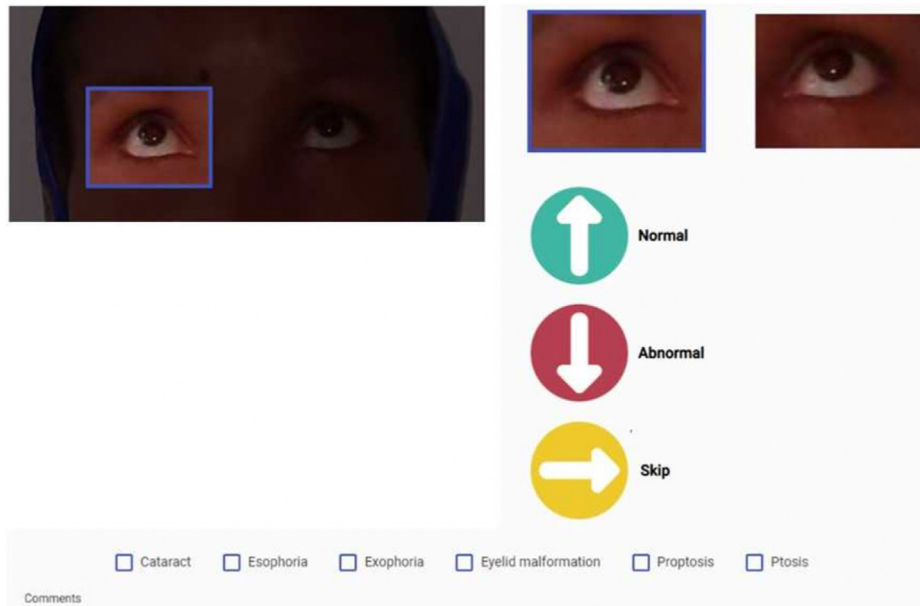


Figure 3. Ophthalmology-facing user interface used to label images for network training.

this dataset are shown in Figure 5. Figure 6 shows the precision-recall curve and area under the precision-recall curve value. The bulk of the testing improvements occurred within 10 epochs. Table 2 shows the confusion matrix for test data at the level of the participant. Figure 7 shows eye photographs that the model classified incorrectly to demonstrate characteristics of difficult images. The area under the ROC curve and area under the precision-recall curve were selected metrics to allow for evaluation of algorithm performance. Sensitivity and specificity are reported to allow for direct assessment of clinical relevance. Positive and negative predictive values are not reported because the study population may not be representative of the global population in need of screening for retinoblastoma.

Discussion

Compared with high-income countries, children with retinoblastoma from low-income countries seek treatment an older age with more advanced disease and a higher rate of metastasis.³ Patients from high-income countries are diagnosed at a median age of 14.1 months, with <1% of patients having extraocular retinoblastoma and <1% having metastasis. In contrast, patients from low-income countries received a diagnosis at a median age of 30.5 months, with 49.1% having extraocular retinoblastoma and 18.9% having metastasis. Clearly, a significant delay in the recognition of retinoblastoma occurs in low-income countries when compared with recognition of retinoblastoma in

Table 1. Demographics of Participants by Testing and Training Dataset and Total Image Count

	Participants Classified as Showing Normal Red Reflex		Participants Classified as Showing Abnormal Red Reflex		Total Images Captured (Eye Pairs)
	Training	Testing	Training	Testing	
Total no.	944	236	222	55	4356
Age (mos)					
Mean	12.2	11.5	13.1	9.9	
0–2 (count)	135	44	30	7	1300
3–12 (count)	460	103	119	30	2300
13–19 (count)	74	28	24	10	416
20–29 (count)	26	5	7	1	244
30–89 (count)	123	28	29	3	640
Sex, no. (%)					
Female	332 (35)	79 (33)	72 (32)	25 (45)	
Male	612 (65)	157 (67)	150 (68)	30 (55)	

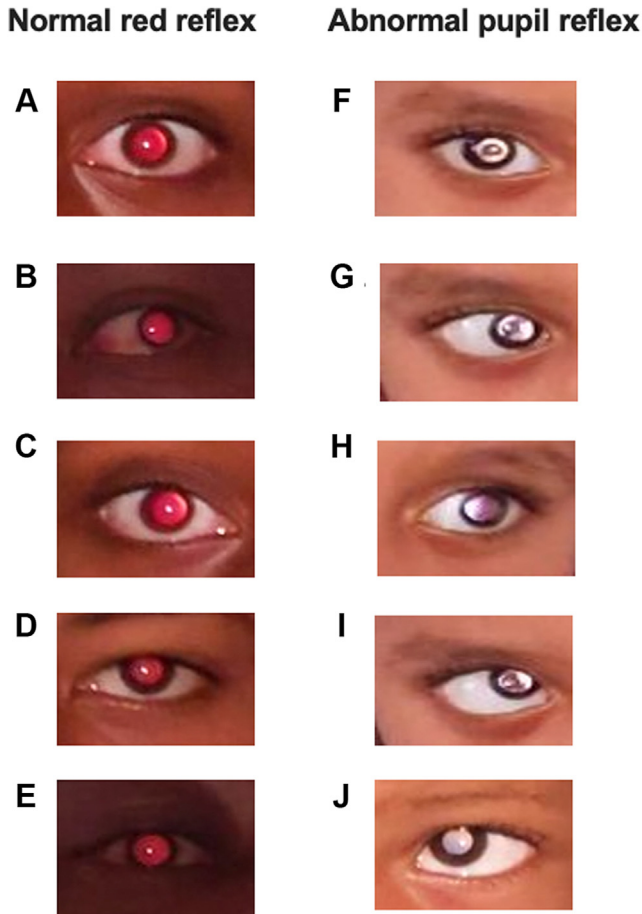


Figure 4. Sample images of (A–E) normal red reflex and (F–J) abnormal pupil reflex in model training and testing.

high-income countries. A review of presentations of retinoblastoma in Ethiopia found that the most common presenting sign was proptosis, signifying late presentation and late referral patterns for these patients.¹⁶ Considering 85% of patients with global retinoblastoma are from the low-income countries, an unmet need exists for effective, easy, and low-cost screening tools for retinoblastoma.

Leukocoria, or white pupil, is the most common presenting symptom in 63% of patients globally. In a meta-analysis, Subhi et al¹⁷ reported that estimated sensitivity of abnormal red reflex testing for detecting ocular pathologic features was 7.5% and specificity was 97.5%. The positive predictive value was 53%, and the negative predictive value was 74%. Leukocoria, or white pupil, can be detected via flash photography and is the most common presenting sign for retinoblastoma. Two smartphone-based screening applications evaluate the pupil color; both have significant drawbacks. MDEyeCare is available only on iOS devices (Apple, Inc). Android phones—typically available at much lower price points—make up 85% of the cellular phone sector, whereas iOS phones make up 12% as of December 2020.¹⁸ A need exists for an Android phone-based application such as EyeScreen. Given the widespread use of Android smartphones in

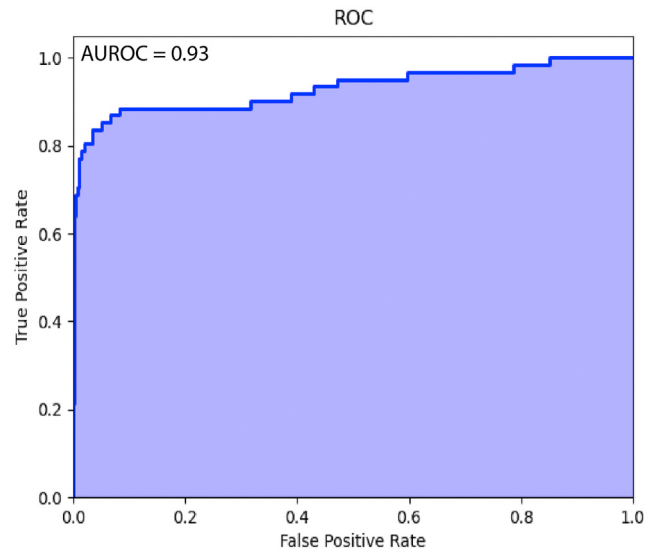


Figure 5. Receiver operating characteristic (ROC) curve with area under the ROC curve (AUROC) for EyeScreen testing dataset.

Sub-Saharan Africa and other resource-limited settings, a leukocoria-detecting application would be accessible for users in these areas.

CRADLE, an Android-based and iOS application, examines existing photographs on the user's device and as such would not be effective as a community-based screening tool.¹⁹ Another iOS-based application, MDEyeCare, was assessed in a small study of 28 patients; it requires relatively standard conditions for photographs, which may be difficult to achieve in a clinical setting. EyeScreen is an improvement because it addresses the limitations of the other 2 Android applications.

In this study, in a resource-limited setting and under varied clinical conditions, we demonstrated the potential of the EyeScreen application. The performance of the application also likely will continue to improve as additional photographs and populations are added into the training dataset. In addition to the use of low-cost smartphones, the

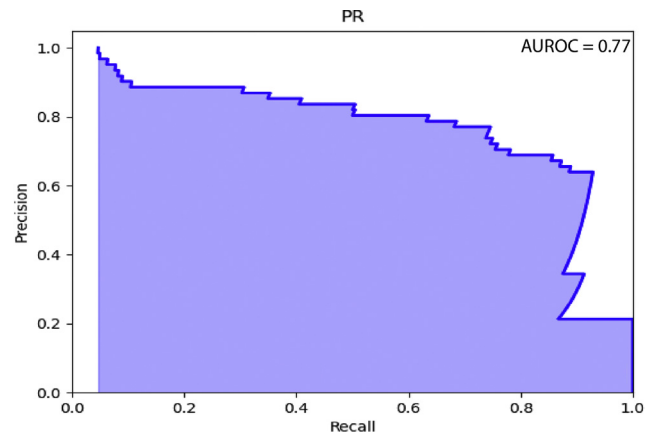


Figure 6. Precision-recall (PR) curve with area under the PR curve (AUPRC) for EyeScreen testing dataset. AUROC = receiver operating characteristic.

Table 2. Confusion Matrix for Test Dataset

Actual	Predicted		Total
	Negative	Positive	
Negative	TN = 173	FP = 63	236
Positive	FN = 7	TP = 48	55
Total	180	110	

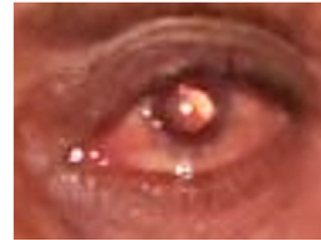
FN = false-negative; FP = false-positive; TN = true-negative; TP = true-positive.

use of low-resource-demand machine learning technology allowed EyeScreen to serve as a proof of concept that may develop into an effective tool as soon as sensitivity is increased further with additional data and the application is tested in multiple settings and populations. Currently, because of the rarity of retinoblastoma, false-positive findings would exceed true-positives greatly. The sensitivity of EyeScreen remains too low in its current form to be deployed for use in screening for a disease in which false-negative findings have such dire consequences. We have taken care to develop EyeScreen in a population subject to typical increased difficulty in screening. Children with darker fundus pigmentation may have abnormal reflex testing results because an increased amount of melanin pigment can give a duller red reflex. Factors that change the appearance of normal eyes can affect model performance, and we wanted to ensure that our input data were from this patient base from the outset to account for these variations. One limitation of our study thus far is that the patient photographs came from a single population in Addis Ababa, Ethiopia, and further development may benefit from the inclusion of additional populations. Another important limitation to note is that this study represents an early-stage proof of concept, and further validation and examination of the model is necessary before implementation in clinical care. Because of the small number of patients with positive results, cross-validation and confidence estimates were not completed; however, this study demonstrated that the EyeScreen application, with its low-cost, efficient technologies, shows promise in developing into an effective tool in a resource-limited setting with further study.

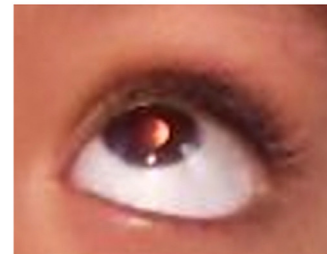
Current community-based screening protocols have poor sensitivity or require training and do not reach many of the most vulnerable populations.²⁰ Further development of effective, free, and simple-to-use applications like EyeScreen could allow increased screening of the populations in community-based settings, allowing earlier referral and treatment. Pediatric screening programs vary widely among

Incorrectly Identified Photos

1.



2.



3.

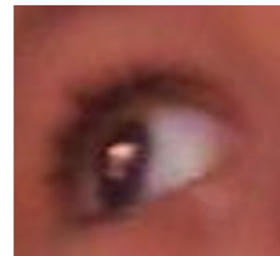


Figure 7. Examples of incorrectly identified eyes in the EyeScreen test dataset.

countries.²¹ This application has its ideal use in settings that already perform community-based screening and intervention, such as childhood vaccination campaigns, and if targeted toward the minimally trained community health worker.

In this study, we explored the feasibility of a low-cost, low-input, end-to-end system using transfer learning algorithms to provide diagnosis of a key presentation of retinoblastoma. Work is ongoing to allow detection of additional ocular pathologic features within the application and to increase its performance for further studies.

Footnotes and Disclosures

Originally received: August 17, 2021.

Final revision: April 7, 2022.

Accepted: April 7, 2022.

Available online: April 15, 2022. Manuscript no. XOPS-D-21-00153.

¹ School of Medicine, University of Michigan, Ann Arbor, Michigan.

² Department of Electrical Engineering & Computer Science, College of Engineering, University of Michigan, Ann Arbor, Michigan.

³ Department of Ophthalmology, St. Paul's Hospital, Addis Ababa, Ethiopia.

⁴ Department of Pediatrics, St. Paul's Hospital, Addis Ababa, Ethiopia.

⁵ Kellogg Eye Center, Department of Ophthalmology and Visual Sciences, University of Michigan Ann Arbor, Michigan.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s): C.N.: Board of directors – World Association of Eye Hospitals; Board trustee – American Society of Ophthalmic Plastic and Reconstructive Surgery Foundation

H.D.: Advisory Board – Castle Bioscience; Financial support – Aura Bioscience

Supported by the University of Michigan Medical School, Ann Arbor, Michigan; the Ravitz Foundation, Kalamazoo, Michigan; and the Richard N. and Marilyn K. Witham Professorship, Ann Arbor, Michigan; and an anonymous foundation (C.N.). The sponsors or funding organizations had no role in the design or conduct of this research.

HUMAN SUBJECTS: Human subjects were included in this study. This study was approved by the University of Michigan Institutional Review Board and by the Institutional Review Board of the St Paul's Hospital Millennium Medical College in Addis Ababa, Ethiopia (HUM00090656). All research adhered to the tenets of the Declaration of Helsinki. All participants or their parents, if younger than 18 years, provided informed consent.

No animal subjects were included in this study.

Author Contributions:

Conception and design: Bernard, Xia, Saleh, Ndukwe, Meyer, Soloway, Sintayehu, Teshome, Tadegegne, Nelson, Demirci

Analysis and interpretation: Bernard, Xia, Meyer, Soloway, Sintayehu, Teshome, Tadegegne, Nelson, Demirci

Data collection: Bernard, Xia, Saleh, Ndukwe, Meyer, Soloway, Sintayehu, Teshome, Tadegegne, Nelson, Demirci

Obtained funding: N/A; Study was performed as part of regular employment/academic duties at University of Michigan Medical School and Kellogg Eye Center.

Overall responsibility: Bernard, Xia, Saleh, Ndukwe, Meyer, Soloway, Nelson, Demirci

Abbreviations and Acronyms:

ROC = receiver operating characteristic.

Keywords:

Application, Machine learning, Retinoblastoma, Screening.

Correspondence:

Hakan Demirci, MD, Kellogg Eye Center, University of Michigan, 1000 Wall Street, Ann Arbor, MI 48169. E-mail: hdemirci@med.umich.edu.

References

- Kivela T. The epidemiological challenge of the most frequent eye cancer: retinoblastoma, an issue of birth and death. *Br J Ophthalmol*. 2009;93(9):1129–1131.
- Dimaras H, Kimani K, Dimba EA, et al. Retinoblastoma. *Lancet*. 2012;379(9824):1436–1446.
- Fabian ID, Abdallah E, Abdullahi SU, et al. Global retinoblastoma presentation and analysis by national income level. *JAMA Oncol*. 2020;6(5):685–695.
- Leander C, Fu LC, Peña A, et al. Impact of an education program on late diagnosis of retinoblastoma in Honduras. *Pediatr Blood Cancer*. 2007;49(6):817–819.
- Abramson DH, Beaverson K, Sangani P, et al. Screening for retinoblastoma: presenting signs as prognosticators of patient and ocular survival. *Pediatrics*. 2003;112(6):1248–1255.
- Shields CL, Gorry T, Shields JA. Outcome of eyes with unilateral sporadic retinoblastoma based on the initial external findings by the family and the pediatrician. *J Pediatr Ophthalmol Strabismus*. 2004;41(3):143–149.
- Canturk S, Qaddoumi I, Khetan V, et al. Survival of retinoblastoma in less-developed countries impact of socioeconomic and health-related indicators. *Br J Ophthalmol*. 2010;94(11):1432–1436.
- Vempuluru VS, Kaliki S. Screening for retinoblastoma: a systematic review of current strategies. *Asia Pac J Ophthalmol*. 2021;10(2):192–199.
- Naseripour M. “Retinoblastoma survival disparity”: the expanding horizon in developing countries. *Saudi J Ophthalmol Off J Saudi Ophthalmol Soc*. 2012;26(2):157–161.
- Mussavi M, Asadollahi K, Janbaz F, et al. The evaluation of red reflex sensitivity and specificity test among neonates in different conditions. *Iran J Pediatr*. 2014;24(6):697–702.
- Sun M, Ma A, Li F, et al. Sensitivity and specificity of red reflex test in newborn eye screening. *J Pediatr*. 2016;179:192–196.e4.
- Panwar N, Huang P, Lee J, et al. Fundus photography in the 21st century—a review of recent technological advances and their implications for worldwide healthcare. *Telemed J E-Health Off J Am Telemed Assoc*. 2016;22(3):198–208.
- Li J, Coats DK, Fung D, et al. The detection of simulated retinoblastoma by using red-reflex testing. *Pediatrics*. 2010;126(1):e202–e207.
- Nyamori JM, Kimani K, Njuguna MW, Dimaras H. The incidence and distribution of retinoblastoma in Kenya. *Br J Ophthalmol*. 2012;96(1):141–143.
- Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: understanding transfer learning for medical imaging. *Adv Neural Inf Process Syst*. 2019;32:3347–3357.
- Shifa JZ, Gezmu AM. Presenting signs of retinoblastoma at a tertiary level teaching hospital in Ethiopia. *Pan Afr Med J*. 2017;28.
- Subhi Y, Schmidt DC, Al-Bakri M, et al. Diagnostic test accuracy of the red reflex test for ocular pathology in infants: a meta-analysis. *JAMA Ophthalmol*. 2021;139(1):33–40.
- StatCounter. Mobile operating system market share in Africa—December 2020. Available at: <https://gs.statcounter.com/os-market-share/mobile/africa>; Published online January 10, 2021; Accessed January 10, 2021.
- Khedekar A, Devarajan B, Ramasamy K, et al. Smartphone-based application improves the detection of retinoblastoma. *Eye*. 2019;33(6):896–901.
- Mndeme FG, Mmbaga BT, Kim MJ, et al. Red reflex examination in reproductive and child health clinics for early detection of paediatric cataract and ocular media disorders: cross-sectional diagnostic accuracy and feasibility studies from Kilimanjaro, Tanzania. *Eye*. 2021;35(5):1347–1353.
- Chen AH, Abu Bakar NF, Arthur P. Comparison of the pediatric vision screening program in 18 countries across five continents. *J Curr Ophthalmol*. 2019;31(4):357–365.