The Impact of Anatomic Racial Variations on Artificial Intelligence Analysis of Filipino Retinal Fundus Photographs Using an Image-Based Deep Learning Model

Carlo A. Kasala, MD[1], Kaye Lani Rea B. Locaylocay, MD-MBA[1], Paolo S. Silva, MD[1,2,3]

[1]Eye and Vision Institute, The Medical City, Pasig City, Philippines
[2]Beetham Eye Institue, Joslin Diabetes Center, Boston, Massachusetts, USA
[3]Department of Ophthalmology, Harvard Medical School, Boston, Massachusetts, USA

Correspondence: Carlo A. Kasala, MD
Office Address: Eye and Vision Institute Department Office, 4th Floor, The Medical City, Ortigas Avenue, Pasig City, Philippines 1602
Office Phone Number: +639171692275
Email Address: carlokasalamd@gmail.com

Disclosures: C.A.K. and K.L.R.B.L. report no financial conflict of interest. P.S.S. receive research support and/or honorarium from Optos plc, Eye Health, Kubota Vision that is not related to this paper.

ABSTRACT

Objectives: This study evaluated the accuracy of an artificial intelligence (AI) model in identifying retinal lesions, validated its performance on a Filipino population dataset, and evaluated the impact of dataset diversity on AI analysis accuracy.

Methods: This cross-sectional, analytical, institutional study analyzed standardized macula-centered fundus photos taken with the Zeiss Visucam®. The AI model's output was compared with manual readings by trained retina specialists.

Results: A total of 215 eyes from 109 patients were included in the study. Human graders identified 109 eyes (50.7%) with retinal abnormalities. The AI model demonstrated an overall accuracy of 73.0% (95% CI 66.6% – 78.8%) in detecting abnormal retinas, with a sensitivity of 54.1% (95% CI 44.3% – 63.7%) and specificity of 92.5% (95% CI 85.7% – 96.7%).

Conclusions: The availability and sources of AI training datasets can introduce biases into AI algorithms. In our dataset, racial differences in retinal morphology, such as differences in retinal pigmentation, affected the accuracy of AI image-based analysis. More diverse datasets and external validation on different populations are needed to mitigate these biases.

Keywords: artificial intelligence; deep learning; retinal imaging; dataset diversity; racial variations

Philipp J Ophthalmol 2024;49:130-137

Artificial Intelligence (AI) is a potentially paradigm-changing innovation in diagnosing and treating retinal disease. Deep Learning (DL) is a newer and more sophisticated subtype of AI that is commonly used to process information from text, audio, and photographs. Whereas older AI models require pre-programmed instructions to analyze information, newer DL algorithms can build on previous information it has been fed to essentially "learn" new things and draw conclusions.[1] For example, DL algorithms are first taught what a retinal fundus photo is. It learns how to identify normal landmarks. Once it can properly do this and encounters a retinal lesion, it learns to identify it as an abnormal finding. It can be taught what this lesion is, and the DL algorithm identifies features and relates it to what it was previously trained on. When it encounters it in another image, it should then be able to identify what it is, despite variations in how it looks.

Applications in medicine include the analysis of diagnostic examinations, including retinal fundus photographs. These AI models can identify visually-threatening conditions and guide patient treatment. AI is particularly useful in low-resource settings where access to specialized care is limited. However, the performance of AI algorithms largely depends on the dataset used for training.[2,3]

This study evaluated the performance of a commercially available AI model that was trained primarily on a dataset comprised of East Asian fundus photos.[4] It is an image-based DL algorithm that can identify and outline the following retinal lesions: drusen, hemorrhages, hard exudates, cotton-wool spots, vascular abnormalities, glaucomatous disc changes, membranes, chorioretinal atrophy or scars, and macular holes.[4] External validation was done on open-source datasets available online; however, these external datasets were limited to lesions seen in Diabetic Retinopathy (DR): hemorrhages, hard exudates, and cotton-wool spots.[4] We assessed its diagnostic performance on a local Filipino population dataset and determined if there are racial differences in retinal features that may affect AI performance.

METHODS

This analytical, cross-sectional study retrieved fundus photos from 123 patients (246 eyes) from the database of The Medical City Eye Instrument Center. All photos were taken using the Zeiss Visucam® fundus camera (Oberkochen, Germany). Only fundus photos with clear ocular media from patients of Filipino descent were included. Additionally, all fundus photos included in this dataset were centered on the macula with full view of the optic nerve nasally (i.e. ETDRS standard fundus photo field 2).[5] Fundus photos with media opacities, views other than standard fundus photo field 2 (i.e. peripheral retina), and those with lesions other than those previously mentioned were excluded.

Data Retrieval

The official readings of the colored retinal fundus photos were retrieved from the electronic medical record. Each reading contained the analysis done by 2 board-certified retina specialists (hereafter referred to as "human graders"), which included a text description of all the lesions seen in the fundus photo along with an overall diagnosis or impression. Information on the presence or absence of specific retinal lesions in each fundus photo, as well as descriptive information such as age, sex and race, were tabulated in a customized data collection form. The photos were then submitted to the AI model for analysis.

Data Analysis

Prior to initiation of the study, the investigators agreed not to disclose the identity of the AI developer and AI model used. Examples of fundus photos after AI image analysis are shown in **Figure 1**. Sensitivity for each lesion was set to the manufacturer's default setting of "medium." Findings were similarly tabulated in the same data collection form. Additionally, fundus photos that had any of the previously mentioned retinal lesions were labeled as "abnormal retinas."

This study used a binary system of "1" and "0" to indicate the presence or absence of a retinal lesion or abnormal retina. Topographic locations of the lesions were not taken into account. Fundus photos were assigned an alphanumerical code in our customized data collection form for identification. The values from human graders and AI analysis were

totaled. Values totaling "2" or "0" indicated that the findings were congruent (both were either "1" meaning they both identified the lesion or "0" meaning they both did not identify the lesion). Values totaling "1" indicated that the findings were not congruent (either human or AI had a value of "1" and the other had a value of "0" meaning only one of these groups identified the lesion).

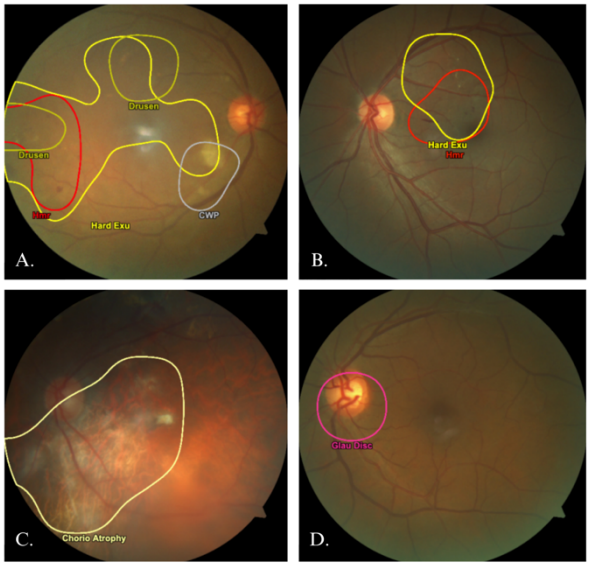

Figure 1. Examples of fundus photographs after AI image analysis. The AI algorithm highlights clusters of lesions and labels them accordingly. Pictured here are drusen, hemorrhages, hard exudates, cotton wool patches (A); hard exudates and hemorrhages (B); chorioretinal atrophy (C); glaucomatous disc change (D).

This study was approved by The Medical City Institutional Review Board.

Statistical Analysis

A minimum sample size of 175 fundus photos was computed for this study based on a 5% level of significance and a probability of disagreement of 0.13. Descriptive statistics summarized the general and clinical characteristics of the fundus photos, including their age and sex. Kappa agreement analysis determined the level of agreement between the analysis of the AI model versus human graders. Sensitivity (Sn), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (LR+), negative likelihood ratio (LR-), and diagnostic accuracy (Acc) were reported with their 95% confidence intervals to assess the diagnostic performance of Fundus AI in detecting retinal lesions. Macular holes were excluded from analysis as no fundus photos containing this lesion were found in this dataset.

RESULTS

A total of 215 eyes from 109 patients were included in the study (**Table 1**). The mean age of patients was 50.94 (±13.54) years. Approximately half (54.13%) of this dataset were of fundus photos from male patients. There were 109 fundus photos (50.70%) labeled as abnormal retinas based on the previous analysis of human graders. The most common lesions identified by human graders were hemorrhages in 51 fundus photos (23.72%), drusen in 49 fundus photos (22.79%), and glaucomatous disc changes in 21 fundus photos (9.77%). No fundus photos with macular holes were identified in this dataset. In contrast, the AI model identified fewer lesions, with only 67 fundus photos (31.16%) identified as an abnormal retina. Drusen were identified in 28 fundus photos (13.02%), hemorrhages in 19 fundus photos (8.88%), and glaucomatous disc changes in 17 fundus photos (7.91%) (**Table 2**).

Table 1. Demographic profile of patients

Demographic Profile	**N=109 patients**
Mean age ± SD, years	50.94±13.54
Sex, n(%)	
Male	59 (54.13)
Female	50 (45.87)

Table 2. Frequency of fundus lesions in our Filipino dataset (n=215)

Lesion Type	**Detected by AI, n(%)**	**Detected by Human Graders, n(%)**
Drusen	28 (13.02)	49 (22.79)
Hemorrhage	19 (8.88)	51 (23.72)
Hard Exudate	8 (3.70)	15 (6.98)
Cotton Wool Spot	8 (3.72)	11 (5.12)
Vascular Abnormality	0 (0.00)	11 (5.12)
Glaucomatous Disc Change	17 (7.91)	21 (9.77)
Membrane	0 (0.00)	4 (1.86)
Chorioretinal Atrophy/Scar	3 (1.40)	3 (1.40)
Abnormal Retina	67 (31.16)	109 (50.70)

Agreement rates between human graders and the AI model were near perfect for cotton wool spots ($\varkappa$ = 0.84, 95% CI 0.65 - 1.00, p<0.001) and moderate for drusen ($\varkappa$ = 0.46, 95% CI 0.31 - 0.60, p<0.001),

hemorrhages ($\varkappa$ = 0.44, 95% CI 0.30 - 0.59, p<0.001), and glaucomatous disc changes ($\varkappa$ = 0.48, 95% CI 0.28 - 0.69, p<0.001) (**Table 3**). AI analysis had the highest accuracy at 98.6% (95% CI 96.0% – 99.7%) for identifying cotton wool spots, with a specificity of 100% (95% CI 98.2% – 100%) and sensitivity of 72.7% (95% CI 39.0% – 94.0%). It had the least accuracy for identifying drusen (83.7% [95% CI 78.1% – 88.4%]) and hemorrhages (84.1% [95% CI 78.5% – 88.7%]) **(Table 4)**.

Table 3. Agreement rates between the AI model and human grader analysis (n=215)

Lesion Type	Concordant Findings (%)	Discordant Findings (%)	Kappa (95% CI)	*p-value*
Drusen	83.72	17.28	0.46 (0.31 – 0.60)	<0.001
Hemorrhage	71.52	28.48	0.44 (0.30 – 0.59)	<0.001
Hard Exudate	95.82	4.19	0.32 (0.12 – 0.53)	<0.001
Cotton Wool Spot	91.54	8.46	0.84 (0.65 – 1.00)	<0.001
Vascular Abnormality	94.88	5.12	0.00 (0.00 – 0.00)	<0.001
Glaucomatous Disc Change	83.87	16.13	0.48 (0.28 – 0.69)	<0.001
Membrane	98.14	1.86	0.00 (0.00 – 0.00)	<0.001
Chorioretinal Atrophy/Scar	98.14	1.86	0.32 (-0.17 – 0.82)	<0.001

DISCUSSION

This study serves as an external validation of the AI model on a Filipino population. In our dataset, the AI had the best performance at identifying cotton-wool spots but least in identifying drusen and hemorrhages. It performed better in more severe disease with many clusters of lesions. Less common findings, such as chorioretinal atrophies/scars, and retinal membranes, were underrepresented in this dataset. There were no fundus photos with macular holes present in this dataset.

The United States Food and Drug Administration requires a minimum performance threshold sensitivity of 85.0% and specificity of 82.5% for an AI model to be used for screening for more than mild diabetic retinopathy (DR).[6] High sensitivity values are crucial for detecting early manifestations of disease, while specificity determines the burden of patient load sent for further evaluation. Studies show that AI models can have acceptable performance on particular datasets, with sensitivity performance varying in between datasets.[7] Wang *et al.* used DL algorithms to detect referable DR in retinal fundus photographs with a sensitivity of 97% and specificity of 87.9%.[7] Jeong *et al.* reviewed the performances of different AI models in the diagnosis and screening of DR, age-related macular degeneration (AMD), and glaucomatous optic neuropathy.[8] These AI models were trained on a binary system to detect referable vs non-referable disease, as well as a separate, stage-based system. Sensitivity performance of these AI models ranged from 82.1% to 98.9% and specificity from 94.1% to 97.3%.[8]

Table 4. Diagnostic performance of the AI model using human grader analysis as reference standard (n=215)

Lesion Type	Sn (95% CI)	Sp (95% CI)	LR+ (95% CI)	LR- (95% CI)	PPV (95% CI)	NPV (95% CI)	Acc (95% CI)
Drusen	42.9 (28.8 – 57.8)	95.8 (91.5 – 98.3)	10.2 (4.6 – 22.5)	0.6 (0.5 – 0.8)	75 (55.1 – 89.3)	85 (79.1 – 89.8)	83.7 (78.1 – 88.4)
Hemorrhage	35.3 (22.4 – 49.9)	99.4 (96.6 – 100)	57.5 (7.9 – 420)	0.7 (0.5 – 0.8)	94.7 (74.0 – 99.9)	83.1 (77.1 – 88.1)	84.1 (78.5 – 88.7)
Hard Exudate	20.0 (2.52 – 55.61)	99.52 (97.34 – 99.99)	41.40 (22.4 – 419)	0.8 (0.59 – 1.10)	66.67 (16.49 –95.29)	96.26 (94.97 –97.23)	95.85 (92.27 –98.09)
Cotton Wool Spot	72.7 (39 – 94)	100 (98.2 – 100)	-	0.3 (0.1 – 0.7)	100 (63.1 – 100)	98.6 (95.8 – 99.7)	98.6 (96.0 – 99.7)
Vascular Abnormality	0 (0 – 28.5)	100 (98.2 – 100)	-	1.00	-	94.9 (94.9 – 94.9)	94.9 (91.0 – 97.4)
Glaucomatous Disc Change	47.6 (25.7 – 70.2)	96.4 (92.7 – 98.5)	13.2 (5.6 – 31.0)	0.5 (0.4 – 0.8)	58.8 (32.9 – 81.6)	94.4 (90.3 – 97.2)	91.6 (87.1 – 95.0)
Membrane	0 (0 – 60.2)	100 (98.3 – 100)	-	1.00 (1.00 – 1.00)	-	98.1 (98.1 – 98.1)	98.1 (95.3 – 99.5)
Chorioretinal Atrophy / Scar	33.3 (0.8 – 90.6)	99.1 (96.6 – 99.9)	35.3 (4.3 – 292)	0.7 (0.3 – 1.5)	33.3 (0.8 – 90.6)	99.1 (96.6 – 99.9)	98.1 (95.3 – 99.5)
Abnormal Retina	54.1 (44.3 – 63.7)	92.5 (85.7 – 96.7)	7.2 (3.6 – 14.3)	0.5 (0.4 – 0.6)	88.1 (77.8 – 94.7)	66.2 (58.0 – 73.8)	73.0 (66.6 – 78.8)

Sn - sensitivity; Sp - specificity, LR+ - positive likelihood ratio, LR- - negative livelihood ratio, PPV - positive predicative value, NPV - negative predicative value, Acc – accuracy, CI - confidence interval

Most image-based AI models are trained to diagnose disease; however, this AI model identifies specific retinal fundus findings without giving an overall diagnosis.[4] The training dataset included 103,262 gradable fundus photos from eyes of East Asian descent.[4] Sensitivity ranged from 88.2% to 99.1%, and specificity from 90.5% to 97.9%. **Table 5** shows a comparison of the sensitivity and specificity of the internal training and external validation datasets used by the developer and our

Filipino dataset. Only hemorrhages, hard exudates, and cotton-wool spots were included in this table as these were the only lesions represented in all three datasets. Compared to the internal training dataset, the AI model had decreased performance in the external validation dataset, and even more significantly decreased performance in our Filipino dataset.

Table 5. Comparison of the AI model's performance for the analysis of hemorrhages, hard exudates and cotton wool spots across the internal East Asian training dataset,[3] external validation datasets[3] and this study's Filipino dataset.

Lesion Type	East Asian Dataset[3]		External Dataset[3]		Filipino Dataset	
	Sn, %	*Sp, %*	*Sn, %*	*Sp, %*	*Sn, %*	*Sp, %*
Hemorrhage	97.2	96.8	88.9	96.6	35.3	99.4
Hard Exudate	99.1	97.2	92.6	100.0	20.0	99.52
Cotton Wool Spot	98.4	95.4	92.3	94.0	72.7	100.0

Sn: Sensitivity; Sp: Specificity

Gudis *et al.* discussed how the availability and sources of datasets create unintentional biases in AI models.[9] As a hypothetical example, an AI model trained to detect cancer survival rates may erroneously associate racial and demographic disparities to poor survival rates.[9] Applied to ophthalmology, AI models may be more likely to associate lesions with certain populations. For example, since AMD is more common in the Caucasian population, it may erroneously associate drusen to be a finding seen in Caucasian retinas only.

Racial differences in morphology can affect AI performance. For instance, retinal differences among Chinese, Malay, and Indian descent include vessel caliber, tortuosity, bifurcation, and fractal dimension and pigmentation.[10-12] These play an important role in the way the AI analyzes the fundus photos. In lighter-skinned individuals, there is better demarcation of the retinal vasculature from the underlying retina because there is more contrast between the dark red color of retinal vessels and the lighter orange retinal pigmentation.[12] This decrease in contrast affects the AI model algorithm's performance, as one of the filters used in DL algorithms is the edge detection or outlining of structures.[13]

We generated color histograms from representative Filipino, East Asian and Caucasian fundus photos from The Medical City Eye Instrument Center's database using Affinity Photo image editing software (Nottingham, United Kingdom) (**Figure 2**). Filipino retinas displayed low contrast especially between retinal vessels and the surrounding retinal tissue. Color histogram analysis showed red, green and blue (RGB) color channels clustered towards the left of the histogram (darker colors) with significant overlap of the color channels. Interestingly, the green channels appeared more pronounced than the red channels in the Filipino fundus photo. The East Asian fundus photo had a more defined separation of the RGB color channels, and a more pronounced red channel compared to the Filipino fundus photo. The Caucasian fundus photo appeared more vibrant with better contrast qualitatively. There was greater separation of the RGB color channels across the spectrum with less overlap of the channels. There was also better distribution across the spectrum towards the right (brighter colors) and the left (darker colors) of the histogram.

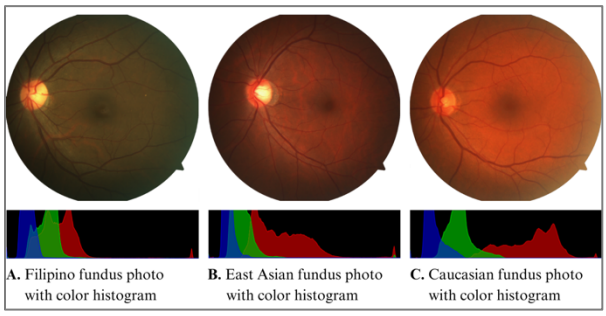

Figure 2. Comparison of a Filipino (A), East Asian (B) and Caucasian (C) fundus photo with the corresponding color histogram demonstrating differences in retinal pigmentation.

We also performed individual pixel color analysis on a representative fundus photo in our dataset in **Figure 3** using the same image editing software. Hex codes are a set of alphanumeric values that determine the combination of colors in an RGB format. It allows us to objectively compare the differences in color wavelengths and brightness of a specific color on a pixel. The first two values represent the levels of red; the second two values represent the levels of green; the last two values represent the levels of blue. Higher values (letters or numbers) represent more vivid representations of a specific color. **Figure 3** shows a fundus photo in our dataset with a faint flame-shaped hemorrhage against a deeply pigmented retina. We sampled the color from a pixel within the lesion and compared it to that of a pixel from the surrounding background retina and found that the color difference was very minimal.

4 out of the 6 total values were equal between the two points. The 2 dissimilar values were also only a few orders of magnitude different — for example, the hex code for a pixel in the hemorrhage in **Figure 3** is #54381E, and the surrounding retina is #50381B. In this case, the 2nd value "4" is only 4 steps away from "0," and the 6th value "E" is only 3 steps away from "B." What this shows is that this pixel within the hemorrhage is only slightly more red and slightly more blue, with equal levels of green, compared to a representative pixel in the surrounding retina. Even qualitatively, the edges of this lesion were very difficult to delineate from the surrounding pigmented retina. The AI failed to identify this finding in our dataset. In contrast, cotton-wool spots, which are brighter yellow and appear more distinct than hemorrhages, were more readily identified by the AI model, as seen in **Figure 4**. Hex code analysis of representative pixels showed a greater color difference, with only 1 out of 6 values equal between the two points. This is an example of a fundus photo that the AI model was able to properly analyze.

An ideal dataset should include an equal number of normal and abnormal datapoints from a diverse population. Some sources even note that datasets should also include low-quality images to mimic real-world scenarios.[14,15] Imbalances in datasets create biases in DL algorithms, as the AI models generally do not have an understanding of disease prevalence in the real world, unlike human graders.

A study by Burlina *et al.* used generative AI techniques to create synthetic retinal images out of known abnormal retinas of light-skinned individuals and normal retinas of dark-skinned individuals. They essentially created new fundus photos of dark-skinned individuals with retinal lesions that they could use to train their AI model, as these were lacking in their dataset. After re-training their AI model with these new images, they noted improved performance at analyzing fundus photos of dark-skinned individuals.[16] This shows that the paucity of these images in the dataset does affect the AI model's performance, as the DL algorithm learns how to discern the nuances in this group compared to other groups.

This disparity in AI datasets is fueled by the fact that most AI models are developed in resource-rich countries, where their models are trained first on an internal dataset comprised of the local population. In a review by Celi *et al.* in 2022, they looked at the distribution of nationalities used in datasets for AI in medicine. They found that most models were initially trained with American and Chinese data.[17] They postulate that this is probably due to the domination of these countries in the technology industry, utilizing their resources for cloud storage and computer processing speed to analyze massive amounts of data.[17] However, this means that resource-poor countries are underrepresented in their data. Reports say that up to 45% of the global population may not have proper representation in ophthalmic datasets used to train AI models.[18] This highlights the importance of utilizing more diverse datasets in training and validating AI models.

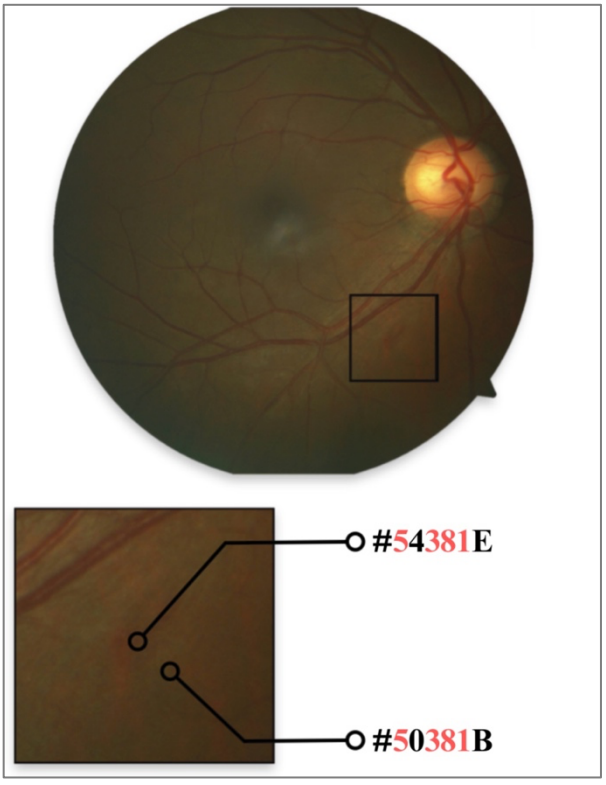

Figure 3. Hex code analysis of discrete pixels within a faint flame-shaped hemorrhage compared to the surrounding retina with deep pigmentation. The AI failed to identify this lesion in our dataset.

This study was limited by the small dataset of fundus photos available. Apart from dataset diversity, dataset volume is also needed to properly train and validate AI models. Massive amounts of fundus photos need to be analyzed for AI models to recognize patterns. A significant number of these fundus photos should contain abnormal findings,

otherwise specificity values will be elevated simply by the AI identifying normal fundus photos.

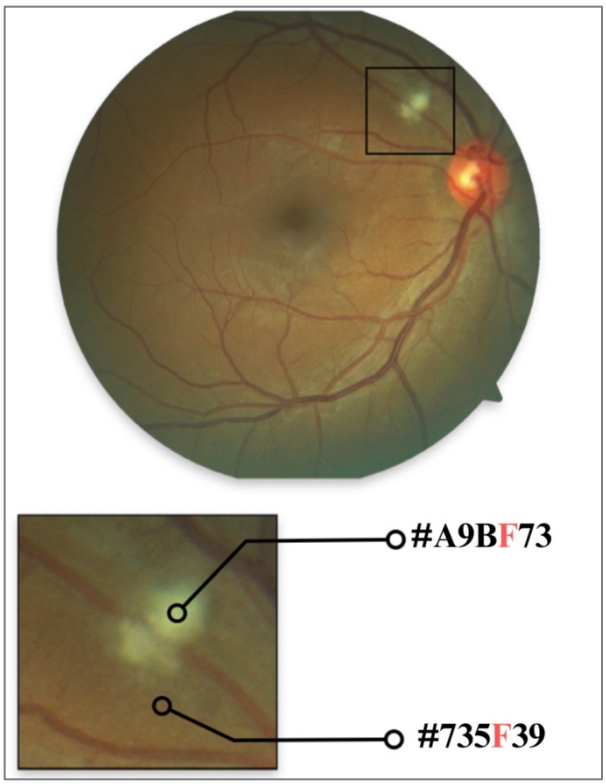

Figure 4. Hex code analysis of discrete pixels within a cotton-wool spot compared to the surrounding retina.

In this dataset, only 50.7% of the fundus photos were labeled as "abnormal retinas." On manual evaluation by human graders, most of these fundus photos displayed only early or mild retinal disease, i.e. few faint hemorrhages or single, small druse. Based on how the AI model processes data, it may have more difficulty identifying these faint lesions on "medium" sensitivity settings. Running the images on a "high" setting could increase its sensitivity, but at the cost of specificity, i.e. increasing the number of false positives.

Based on our testing of this particular AI model on our Filipino dataset, it will need further training and validation on a Filipino population to increase sensitivity before it can be used for screening purposes. The low sensitivity performance means the AI model misses a significant number of fundus photos with retinal lesions. Although it can be argued that most of these fundus photos show non-referable disease to begin with, its performance will still need to be improved prior to clinical use. A larger and more diverse dataset with more abnormal fundus photos will also be needed to better assess performance, especially for chorioretinal atrophies and scars, macular holes, and retinal membranes.

AI analysis of fundus photos shows promise as a tool for screening retinal disease; however, it must be optimized for the target population by training it with diverse and representative datasets. Addressing AI model optimization and dataset bias by incorporating data from various sources is crucial. Further training and validation on a Filipino population are needed to improve sensitivity before clinical use. A larger dataset with more abnormal fundus photos is also necessary for better assessment, especially for less common lesions. Generation of synthetic retinal images using generative AI techniques shows promise as an alternative way to mitigate these biases, especially in populations that have a paucity of available data.

REFERENCES

1. Sarker I. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science* 2021;2:420.
2. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, *et al.* Artificial intelligence in retina. *Prog Retin Eye Res* 2018;67:1-29.
3. Ting DS, Pasquale LR, Peng L, *et al.* Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol* 2019;103(2):167-175.
4. Son J, Shin JY, Kim HD, *et al.* Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. *Ophthalmology* 2020; 127(1):85-94.
5. Early Treatment Diabetic Retinopathy Study Research Group. Grading Diabetic Retinopathy from Stereoscopic Color Fundus Photographs — An Extension of the Modified Airlie House Classification. ETDRS report number 10. *Ophthalmology* 1991;98:786–806.
6. United States Food and Drug Administration. De novo classification request for IDx-DR. January 12, 2018: https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180001.pdf (accessed February 11, 2023).
7. Wang Y, Shi D, Tan Z, *et al.* Screening referable diabetic retinopathy using a semi-automated deep learning algorithm assisted approach. *Front Med* 2021;8.
8. Jeong Y, Hong YJ, Han JH. Review of machine learning applications using retinal fundus images. *Diagnostics* 2022;12(1):134.

9. Gudis DA, McCoul ED, Marino MJ, Patel ZM. Avoiding bias in artificial intelligence. *Int Forum Allergy Amp* Rhinol 2023;13(3):193-195.
10. Li X, Wong WL, Cheung CY, *et al.* Racial differences in retinal vessel geometric characteristics: a multiethnic study in healthy asians. *Invest Opthalmol Vis Sci* 2013; 54(5):3650-3656.
11. Bourne RR. Ethnicity and ocular imaging. *Eye* 2011; 25(3):297-300.
12. Rochtchina E, Wang JJ, Taylor B, *et al.* Ethnic variability in retinal vessel caliber: a potential source of measurement error from ocular pigmentation? — the Sydney Childhood Eye Study. *Invest Opthalmol Vis Sci* 2008;49(4):1362-1363.
13. Carin L, Pencina MJ. On deep learning for medical image analysis *JAMA* 2018;320(11):1192-1193.
14. Lee C, Yanagihara R, Lee A. Using Deep Learning Models to Characterize Major Retinal Features on Color Fundus Photographs. *Ophthalmology* 2020;127(1):95-96.
15. Lee KG, Song SJ, Lee S, *et al.* A Deep Learning-Based Framework for Retinal Fundus Image Enhancement. *PLoS ONE* 2023;18(3):e0282416.
16. Burlina P, Joshi N, Paul W, *et al.* Addressing Artificial Intelligence Bias in Retinal Diagnostics. *Transl Vis Sci Technol* 2021;10(2):13.
17. Celi LA, Cellini J, Charpignon ML, *et al.* Sources of bias in artificial intelligence that perpetuate healthcare disparities — A global review. *PLOS Digit Health* 2022; 1.
18. Jacoba CM, Celi LA, Lorch AC, *et al.* Bias and non-diversity of big data in artificial intelligence: focus on retinal diseases. *Semin Ophthalmol* 2023;38:433-441.