Automated Machine Learning for Referable Diabetic Retinopathy Image Classification from Ultrawide Field Images

Leandro Victor L. Arcena, MD-MBA[1], Paolo S. Silva, MD[1,2,3]

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

Correspondence: Leandro Victor L. Arcena, MD-MBA
Office Address: Asian Eye Institute, PHINMA Plaza, Rockwell Center, Makati City
Office Phone Number: +63 9177921763
Email Address: lean.arcena@gmail.com

Disclosure: L.V.L.A reports no financial relationship. P.S.S. receives research support and/or honorarium from Optos plc, AEye Health, Kubota Vision not related to this paper.

ABSTRACT

Objective: To develop and evaluate the diagnostic performance of an automated machine learning (AutoML) model for the detection of referable diabetic retinopathy (refDR) in ultrawide field (UWF) retinal images from local Philippine retinal image datasets.

Methods: A Google AutoML Vision model was trained using 2000 UWF images with a 50/50 ratio of refDR/non-refDR. Images were labeled according to the Early Treatment Diabetic Retinopathy Study (ETDRS) severity grading. RefDR was defined as moderate nonproliferative DR or worse. The dataset was split with 80% for training, 10% for validation, and 10% for testing. Two sets of published UWF image sets were used for external validation. Sensitivity and specificity were calculated in accordance with United States Food and Drug Administration (US FDA) performance requirements of 0.85 and 0.825, respectively.

Results: The area under the precision-recall curve was 0.998. External validation against two datasets showed a sensitivity/specificity of 0.88/0.83 (95% CI 0.80-0.94/0.74-0.89) and 0.83/0.80 (95% CI 0.74-0.89/0.72-0.86), respectively. Positive and negative predictive values were 0.81/0.89 (95% CI 0.73-0.89/0.82-0.94) and 0.75/0.86 (95% CI 0.66-0.83/0.79-0.91), respectively.

Conclusions: The pilot performance of the custom AutoML model constructed using local Philippine data approaches US FDA requirements for the diagnosis of referable DR. The ease of use and intuitiveness of the platform, combined with its performance, support the potential of no-code AI in the detection of refDR.

Keywords: artificial intelligence, machine learning, referable diabetic retinopathy, ultrawide-field, teleophthalmology

Philipp J Ophthalmol 2024;49:138-143

Diabetic Retinopathy (DR) is a prevalent and potentially blinding ocular pathology. Regular screening is crucial due to its often-asymptomatic progression. Advancements in retinal imaging, such as standard 45° retinal photography and ultrawide field (UWF) imaging, have significantly improved the detection and management of DR. The integration of artificial intelligence (AI) in ophthalmology, specifically through deep learning systems for DR detection, shows promising results. No-code AI platforms like Google AutoML (Google, CA, USA) are designed to be accessible to users without programming expertise, making it easier for clinicians to develop and implement AI-driven diagnostic tools. This study explored the application of Google AutoML with UWF retinal images from a local Philippine tertiary hospital image dataset to create and evaluate a machine learning model for detecting referable diabetic retinopathy (refDR).

METHODS

This non-experimental, cross-sectional study involved training a Google AutoML model to detect refDR from UWF retinal images. A dataset of 2000 images was obtained from the Eye Instrument Center at The Medical City, Pasig City. Images were de-identified UWF images and used in accordance to the ethical standards stated in the 1964 Declaration of Helsinki. The study protocol was approved by the institutional review board of The Medical City.

Data Management and Labelling

The dataset was comprised of images with all stages of diabetic retinopathy from no apparent retinopathy to advanced proliferative disease. Eyes with non-diabetic retinal pathology were excluded from data gathering. Manual retrieval of de-identified UWF retinal images from the Eye Instrument Center database was done for maximum image quality of the dataset. UWF images of retinas (including central and peripheral views) with previous official readings by vitreoretina specialists were included in the initial data gathering. Images were segregated and labelled according to differing DR severity while being cross-referenced with their respective official readings. The official readings by the specialists were used as ground truth for the study. The readings with DR severity were done in accordance with the Early Treatment Diabetic Retinopathy Study (ETDRS). Images with pathologies not related to diabetic retinopathy were excluded (e.g. retinal vein and artery occlusions, age-related macular degeneration, hypertensive retinopathy, etc.)

The required training sample size for machine learning models applied to medical imaging data is not completely known. Based on published classification-based studies in DR detection using fundus imaging, and multiple performance metrics for AI programs, the sample size for the training set should be at least 1,000 images for non-referable DR, and 1,000 images for referable DR. We estimated that the minimum number of images needed to train the Google AutoML model was at least 2,000.[1] The 2,000-image dataset was created with a 50-50 split of images with non-referable diabetic retinopathy (non-refDR) and refDR. RefDR defined as moderate non-proliferative DR or worse, while non-refDR was defined as no DR or mild non-proliferative DR.[2]

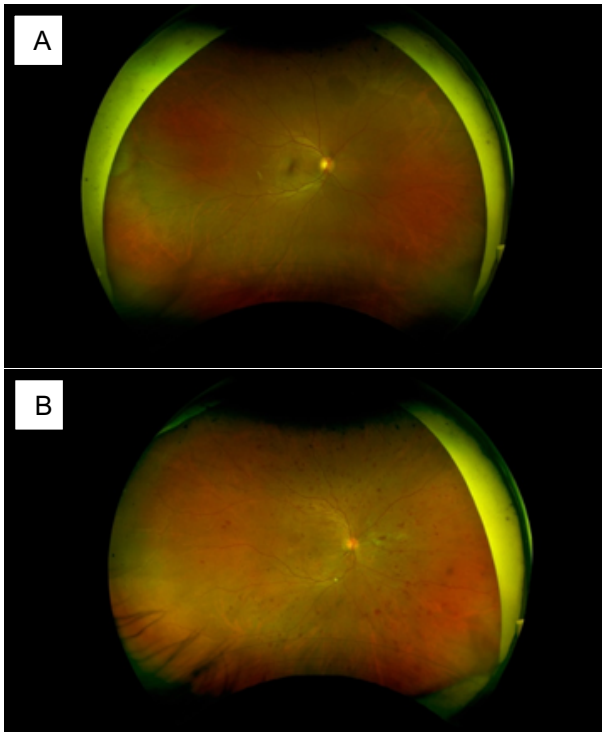

Figure 1. Sample ultrawide field retinal photographs used in the custom AutoML dataset. (A) Non-referable diabetic retinopathy, (B) Referable diabetic retinopathy.

AutoML Model Training

A Google Cloud AutoML Vision account was used to upload and train the model with a 50/50 split of referable and non-referable DR images (N=2,000). As per system requirements, the dataset was automatically split into 80% for training, 10% for validation, and 10%. The model was trained with 16 node hours. Two external UWF image sets were then used to test the model's performance separately: 225 images from Jacoba *et al.* and 256 images from Liu *et al.,* with labels adjusted to this study's definitions of refDR.[3,4] A comparison of the datasets used is presented in **Table 1.**

Table 1. Diabetic Retinopathy Severity between Datasets

Diabetic Retinopathy Severity, n(%)	**Custom AutoML Model Training Set (2022) (N = 2000)**	**Jacoba *et al.* (2022) Testing Set (N = 225)**	**Liu *et al.* (2022) Testing Set (N = 256)**
No DR	898 (44.9%)	80 (35.56%)	74 (28.91%)
Mild NPDR	102 (5.1%)	41 (18.23%)	73 (28.51%)
Moderate NPDR	171 (8.55%)	24 (10.67%)	71 (27.73%)
Severe NPDR	126 (6.3%)	34 (15.11%)	28 (10.94%)
PDR	349 (17.45%)	32 (14.22%)	4 (1.56%)
PRP	354 (17.7%)	-	-
Ungradable	-	-	6 (2.34%)

DR – Diabetic retinopathy; NPDR – Non-proliferative diabetic retinopathy; PDR – Proliferative diabetic retinopathy; PRP – Presence of pan-retinal photocoagulation; Ungradable - Image quality is low and cannot be diagnosed and graded.

Analysis

Google AutoML provided detailed performance statistics, including the area under the precision-recall curve (AUPRC). Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated with confidence intervals for external testing using VassarStats (vassarstats.net, NY, USA). In particular, sensitivity and specificity were calculated to meet the US FDA guidelines requirements of 0.85 and 0.825, respectively.[5]

RESULTS

The model was trained on a 2,000-image dataset comprised of 1,000 referable and non-referable DR images each. The external validation sets had similar compositions. The first external testing set of 225 images by the Jacoba *et al.* had 113 images labelled refDR, while 112 were non-refDR. The second external testing set by Liu *et al.* had 109 refDR images and 147 non-refDR images.[3,4] The native calculations by Google AutoML showed an AUPRC of 0.998 (Figure 2).

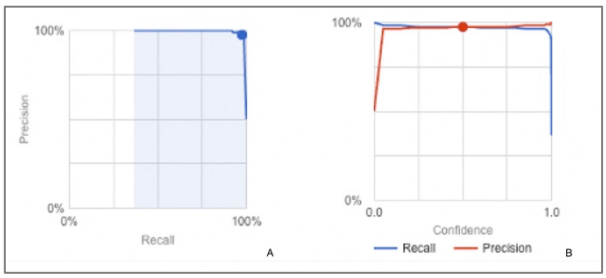

Figure 2. (A) Area under the precision-recall curve: 0.998, (B) Confidence threshold was set at 0.5 to balance both precision and recall at 97.5%.

Jacoba et al. (2022) 225-image testing set

The confusion matrix and performance metrics are summarized in **Table 2.** Overall accuracy of the custom AutoML model against the Jacoba *et al.* testing set is 85.33%. Of the 113 referable DR images of the testing set, the custom AutoML model correctly identified 92 images as referable DR. Moreover, the model identified 100 images as non-referable of the possible 112. There were 21 false positive cases and 12 false negative cases recorded. Referable DR prevalence was recorded as 46.22% (95% CI, 39.61% - 52.30%). Against the published testing set by Jacoba *et al.*, this study's custom AutoML model had a sensitivity of 88.46% (95% CI, 80.34% - 93.63%) and specificity of 82.64% (95% CI, 74.46% - 88.70%). Predictive values are as follows: positive predictive value of 81.42% (95% CI, 72.77% - 87.88%) and a negative predictive value of 89.29% (95% CI, 81.67% - 94.10%).

Table 2. Performance of the Custom AutoML model against the Jacoba *et al.* (2022) Testing Set

Custom AutoML Model	**Jacoba *et al.* (2022) Testing Set (Ground Truth)**						
	RefDR[a]	Non-refDR	Total	Sensitivity (CI 95%)	Specificity (CI 95%)	PPV (CI 95%)	NPV (CI 95%)
RefDR[a]	92	21	113	88.46% (80.43% - 93.63%)	82.64% (74.46% - 88.70%)	81.42% (72.77% - 87.88%)	89.28% (81.67% - 94.10%)
Non-refDR	12	100	112				
Total	104	121	225				

RefDR – Referable diabetic retinopathy; Non-refDR – Non-referable diabetic retinopathy; PPV – Positive predictive value; NPV – Negative predictive value

Discrepancies between model performance versus ground truth in this set were due to the following: (a) presence of media opacity representing 57.58% of the 33 missed cases, (b) image artifacts representing 45.46%, and non-DR retinal lesions representing 9.09%. A summary of non-diabetic retinopathy findings in cases with missed diagnoses is seen in **Table 3.**

Table 3. Summary of non-diabetic retinopathy findings in cases with missed diagnoses

Image Findings, n(%)	**Jacoba *et al.* (2022) testing set** (N = 33)	**Liu *et al.* (2022) testing set** (N = 49)
Image artifacts[a]	15 (45.46%)	42 (91.30%)
Media opacity	19 (57.58%)	12 (24.50%)
Non-DR retinal lesions	3 (9.09%)	4 (8.16%)
None	4 (12.12%)	0 (0%)

DR – Diabetic retinopathy

[a] Image artifacts refer to objects in the photo not usually studied by retinal photography, which include camera lens blemishes, glare, lids and lashes, etc.

Liu et al. (2022) 256-image testing set

The confusion matrix and performance metrics are summarized in **Table 4.** Using the external 256-image testing set by Liu *et al.*, the overall accuracy of the custom AutoML model was 80.86%. True positives detected were 90, while true negatives were detected at 117. False negatives were 19, while false positives were at 30. Referable DR prevalence was recorded at 42.58% (95% CI, 40.66% - 53.18%). Against the external published testing set by Liu *et. al.*, the study's custom AutoML had a sensitivity of 82.57% (95% CI, 73.86% - 88.92%) and specificity of 79.59% (95% CI, 71.99% - 85.61%). Predictive values are as follows: positive predictive value of 75% (95% CI, 66.11% - 82.25%) and a negative predictive value of 86.02% (95% CI, 78.79% - 91.16%).

Table 4. Performance of the Custom AutoML model against the Liu *et al.* (2022) Testing Set

Liu *et. al.* (2022) Testing Set (Ground Truth)							
Custom AutoML Model	RefDR	Non-refDR	Total	Sensitivity (CI 95%)	Specificity (CI 95%)	PPV (CI 95%)	NPV (CI 95%)
RefDR	90	30	120	82.57% (73.86% - 88.92%)	79.59% (71.99% - 85.61%)	75% (66.11% - 82.25%)	86.01% (78.79% - 91.17%)
Non-refDR	19	117	136				
Total	109	147	256				

RefDR – referable diabetic retinopathy; Non-refDR – Non-referable diabetic retinopathy; PPV – Positive predictive value; NPV – Negative predictive value

Of the 49 discrepancies in this set, image artifacts were found to represent 91.30%, while presence of media opacity amounted to 24.50%. The remaining 8.16% of discrepancies were attributed to the presence of non-DR retinal lesions. A summary of discrepancies is seen in **Table 3.**

DISCUSSION

The results of this study's custom Google AutoML model are promising. Against the published external training set of 225 UWF images by Jacoba *et al.*, with a sensitivity and specificity of 88.57% and 82.64%, respectively, our custom Google AutoML model met the minimum requirements of the US FDA. Interestingly, our custom AutoML model approaches US FDA requirements when tested against the 256 UWF image set by Liu *et al.*, with a sensitivity and specificity of 82.57% and 79.59%, respectively.

Discrepancy between performances per testing set may be attributed to multiple factors. Due to the nature of image-based machine learning models such as Google AutoML, images uploaded may have multiple confounders that are not necessarily included in this study's limitations as per the definition of referable diabetic retinopathy. It seems that variables that obstruct the view of the retina in these testing sets occur frequently in cases with discrepancies in diagnosis. Media opacities such as cataracts were frequently seen in these cases, amounting to 57.58% and 24.50% of discrepancies in the Jacoba *et al.* and Liu *et al.* testing sets, respectively. Variables such as image artifacts (obstructions with lids, lashes, camera lens glare, and debris) were also frequently observed in cases with incorrect diagnoses, representing 45.46% in the Jacoba *et al.* testing set and 91.30% in the Liu *et al.* set. Further research into methods optimizing UWF-based AI performance in consideration of these variables is warranted.

Dataset construction also influences outcomes in AI. While the required training sample size for AI models is not completely known, the general consensus concerning image-based AI systems is that more images used in datasets for training will yield more consistent results. In 2007, Ting *et al.* developed

a deep learning system AI for diabetic retinopathy which included 494,661 standard 45° retinal images. Their study's performance metrics were acceptable at 90.5% sensitivity and 91.6% specificity.[6] On the other hand, the size of the external test set may also play a factor. Free-to-use research datasets like the Messidor-2 set contain 1,748 standard 45° retinal images. This research dataset was used as a testing/validation set for by Gulshan *et al.* in 2016, wherein a deep learning AI using 128,175 images as its training dataset was used to identify standard images for the presence of referable DR. Its performance with the Messidor-2 set was 96.1% sensitivity and 93.6% specificity.[7]

The type of retinal image used to train and test AI platforms may also present as a confounding factor to discrepancies. While multiple AIs have been studied and employed for screening and detection of DR, only two such AI platforms have been approved for clinical use by the US FDA: (a) IDxDR (Digital Diagnostics, Coralville, IA, USA) and (b) EyeNUK (Eyenuk Inc., LA, CA, USA).[5,8-9] These AI platforms were constructed and deployed by using a database of standard retinal photographs. In lieu of advancements in retinal imaging systems, UWF retinal image-based databases are currently being created and studied worldwide, however, none so far are being approved for clinical use by the US FDA. A published study using UWF-image datasets in AI by Silva *et al.* in 2022 used a training dataset of 3,999 images, yielding a performance of 0.79 sensitivity and 0.83 specificity against an external test set; performance metrics of three UWF-based AI systems are detailed in **Table 5**. Recent UWF-based dataset adoption in AI approaches the performance metrics established by standard 45° retinal photography and therefore should be studied continuously. [1,10-11]

An advantage of this pilot study on creating a no-code custom AutoML for referable DR is that the current 2000-image dataset can be further built upon as the Google Cloud platform offers options to the end-user to edit current models continuously. For example, it would be easy to include more UWF retinal images to achieve higher numbers in the current dataset. Training the updated model is a simple endeavor. Inclusion and exclusion criteria can be modified. Inclusion of confounders such as non-diabetic retinal lesions in the custom AutoML model may offer better performance. Additional external testing sets can further be used to evaluate the custom AutoML model and see improvement in overall performance. While the performance of the current custom AutoML model can be further optimized, our study's results are promising enough that this model can potentially be applied in the investigator's institutional teleophthalmology program as a possible adjunctive screening method for referable DR diagnosis, though care must still be taken in ensuring the patients themselves are directed to proper vitreoretinal specialists for management.

Table 5. Comparison with different artificial intelligence systems for detection of diabetic retinopathy

Referable DR	**Custom AutoML**[a] (N = 225)	**Custom AutoML**[b] (N = 256)	**IDx**[c,d] (N = 1,784)	**EyeNUK**[c,d] (N = 1,333 eyes)	**Wang *et al.* 2018**[e] (N = 754 eyes)	**Tang *et al.* 2021**[e] (N = 925 eyes)	**Silva *et al.* 2022**[e] (N = 192 eyes)
Sensitivity	0.88	0.83	0.87	0.89 – 1.00	0.90	0.82 – 0.87	0.79
Specificity	0.83	0.79	0.90	0.85 – 0.97	0.53	0.83 – 1.00	0.83
PPV	0.81	0.75	0.73	0.20 – 0.67	0.52	0.98 – 1.00	0.90
NPV	0.89	0.86	0.67	0.99 – 1.00	0.91	N/A	0.67
Prevalence	46.2%	42.6%	66.2%	2.4% - 19.4%	21.2%	61.8%-94.6%	66.1%

a- Performance against the Jacoba et. al. 2022 testing set
b- Performance against the Liu et. al. 2022 testing set
c- Data based on US-FDA submissions for IDx and EyeNUK
d- Evaluation of standard 45° retinal images
e- Evaluation of ultrawide field retinal images

Google AutoML's user-friendly platform allows users without technical expertise to develop effective AI models for DR detection. This study validates the potential of no-code AI in ophthalmologic practice, highlighting its applicability in low resource settings. Further research and optimization are warranted to enhance the model's performance and integration into clinical workflows.

REFERENCES

1. Silva P, Lewis D, Cavallerano J, Ashraf M, *et al.* Automated machine learning (AutoML) models for diabetic retinopathy (DR) image classification from ultrawide field (UWF) Retinal Images. *Invest Ophthalmol Vis Sci* 2022;63(7):2095–F0084.

2. Early Treatment Diabetic Retinopathy Study Research Group. Early photocoagulation for diabetic retinopathy. ETDRS report number 9. *Ophthalmology*. 1991;98(5 Suppl):766-85
3. Jacoba CM, Aquino LA, Salva CM, *et al*. Comparisons of Handheld Retinal Imaging Devices with Ultrawide Field (UWF) and Early Treatment Diabetic Retinopathy Study (ETDRS) Photographs for Determining Diabetic Retinopathy (DR) Severity. *Invest Ophthalmol Vis Sci* 2022;63(7):4449–F0128.
4. Liu R, Wang X, Wu Q, *et al*. DeepDRiD: Diabetic Retinopathy-Grading and Image Quality Estimation Challenge. *Patterns (N Y)*. 2022;3(6):100512.
5. Abràmoff MD, Lavin PT, Birch M, *et al*. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med*. 2018;1:39.
6. Ting DSW, Cheung CY, Lim G, *et al*. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA*. 2017;318(22):2211-2223.
7. Gulshan V, Peng L, Coram M, *et al*. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*. 2016;316(22):2402-2410.
8. Abràmoff MD, Folk JC, Han DP*, et al*. Automated analysis of retinal images for detection of referable diabetic retinopathy. *JAMA Ophthalmol*. 2013;131(3):351-357.
9. Bhaskaranand M, Ramachandra C, Bhat S, *et al*. The value of automated diabetic retinopathy screening with the EyeArt system: a study of more than 100,000 consecutive encounters from people with diabetes. *Diabetes Technol Ther*. 2019;21(11):635–643.
10. Wang X, Ji Z, Ma X, *et al*. Automated Grading of Diabetic Retinopathy with Ultra-Widefield Fluorescein Angiography and Deep Learning. *J Diabetes Res*. 2021;2021:2611250.
11. Tang F, Luenam P, Ran AR, *et al*. Detection of Diabetic Retinopathy from Ultra-Widefield Scanning Laser Ophthalmoscope Images: A Multicenter Deep Learning Analysis. *Ophthalmol Retina*. 2021;5(11):1097-1106.
12. Tabuchi H. Understanding required to consider AI applications to the field of ophthalmology. *Taiwan J Ophthalmol*. 2022;12(2):123-129.
13. Hogarty DT, Mackey DA, Hewitt AW. Current state and future prospects of artificial intelligence in ophthalmology: a review. *Clin Exp Ophthalmol*. 2019;47(1):128-139.
14. Rahimy, E. Deep learning applications in ophthalmology. *Curr Opin Ophthalmol*. 2018;29(3):254-260.
15. Poplin R, Varadarajan AV, Blumer K*, et al*. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng*. 2018;2(3):158-164.
16. Bellemo V, Lim G, Rim TH, *et al*. Artificial Intelligence Screening for Diabetic Retinopathy: The Real-World Emerging Application. *Curr Diab Rep*. 2019;19(9):72.
17. Balyen L, Peto T. Promising Artificial Intelligence-Machine Learning-Deep Learning Algorithms in Ophthalmology. *Asia Pac J Ophthalmol (Phila)*. 2019;8(3):264-272.
18. Jacoba, CP, Celi, LA, Silva, PS. Biomarkers for Progression in Diabetic Retinopathy: Expanding Personalized Medicine through Integration of AI with Electronic Health Records. *Semin Ophthalmol*. 2021;36(4):250-257.