Implementation of a Large-Scale Retinal Image Curation Workflow Using Deep Learning Framework

Abs# 3708844

ARVO 2022

Rohit Balaji[1], Jen Heathcote[1], Robert Slater[2], Nancy Barrett[1], Rick Voland[1], Vesna Tomic[1], Jared McDonald[1], Barbara A. Blodi[1], Amitha Domalpally[1,2]

[1]WISCONSIN READING CENTER
[2]A-EYE Unit
Department of Ophthalmology and Visual Sciences
UNIVERSITY OF WISCONSIN
SCHOOL OF MEDICINE AND PUBLIC HEALTH

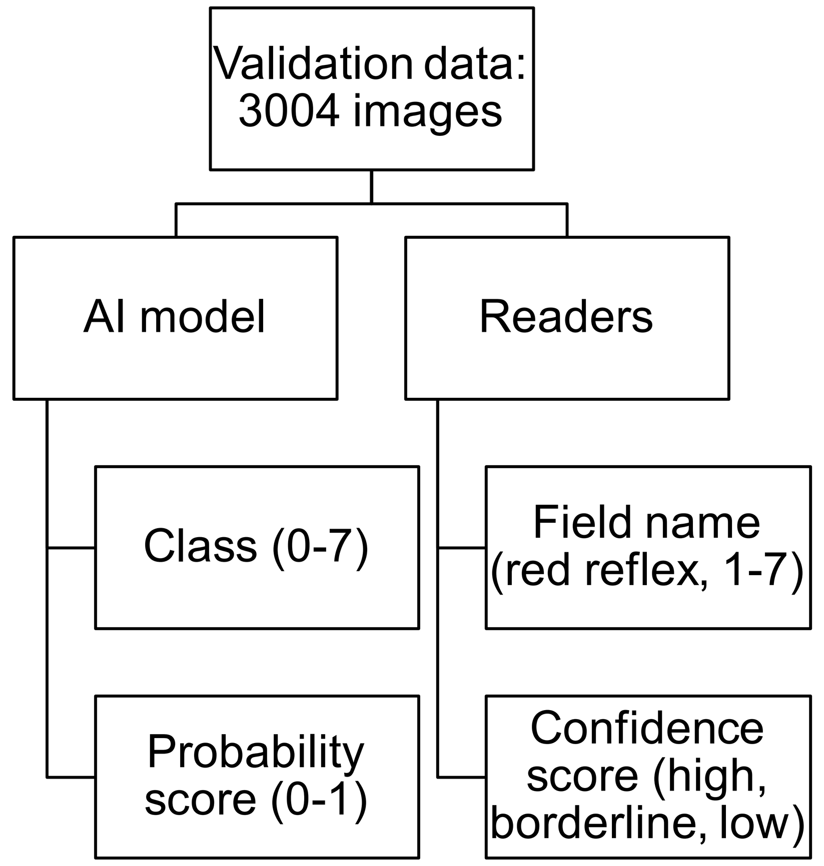
Research to Prevent Blindness

Background and Purpose

- Stereoscopic 7F images are used for evaluating diabetic retinopathy (DR) in clinical trials (Figure 1)
- Data curation and organizing the images is labor intensive
- We sought to implement a deep learning framework to ease the workload of identifying and naming individual 7M fields in order to augment grader workflow
- We explored the potential of a tiered system using AI classification with human review to increase accuracy

Methods

- The training dataset included 17,529 images submitted for clinical trials.
- Two models were trained in an identical fashion for 8 classes using images sized to 256x256 – one model for the right eye and one model for the left eye. EfficientNetB0 architecture from Tensorflow was used.



- AI results and human results were compared for accuracy and reliability of the AI system to identify 7M retinal fields
- We explored the use of probability score to create a flagging system for inaccurate images

Commercial interest disclosures: NONE for all authors

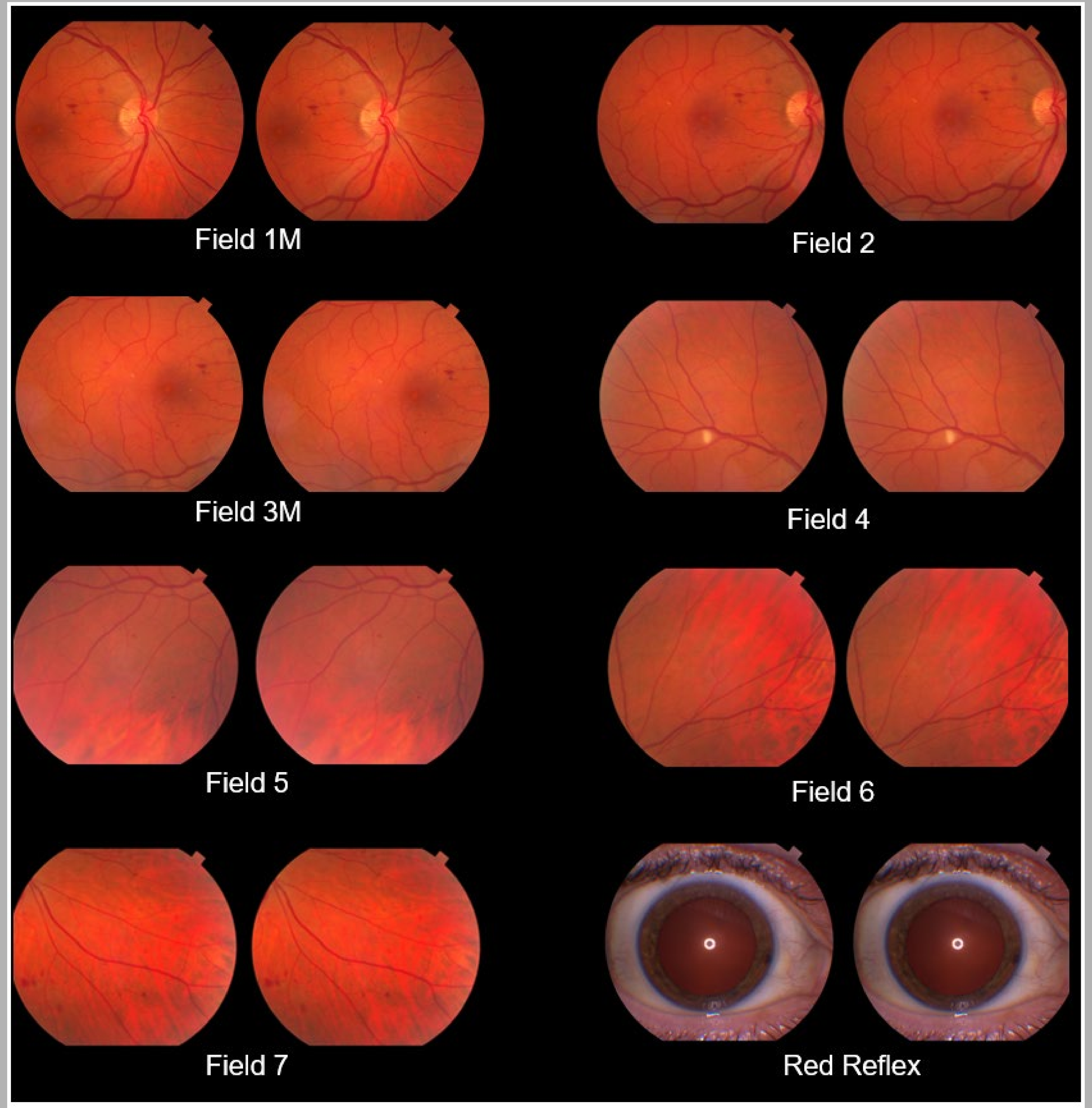


Figure 1. Typical stereoscopic 7-modified field images

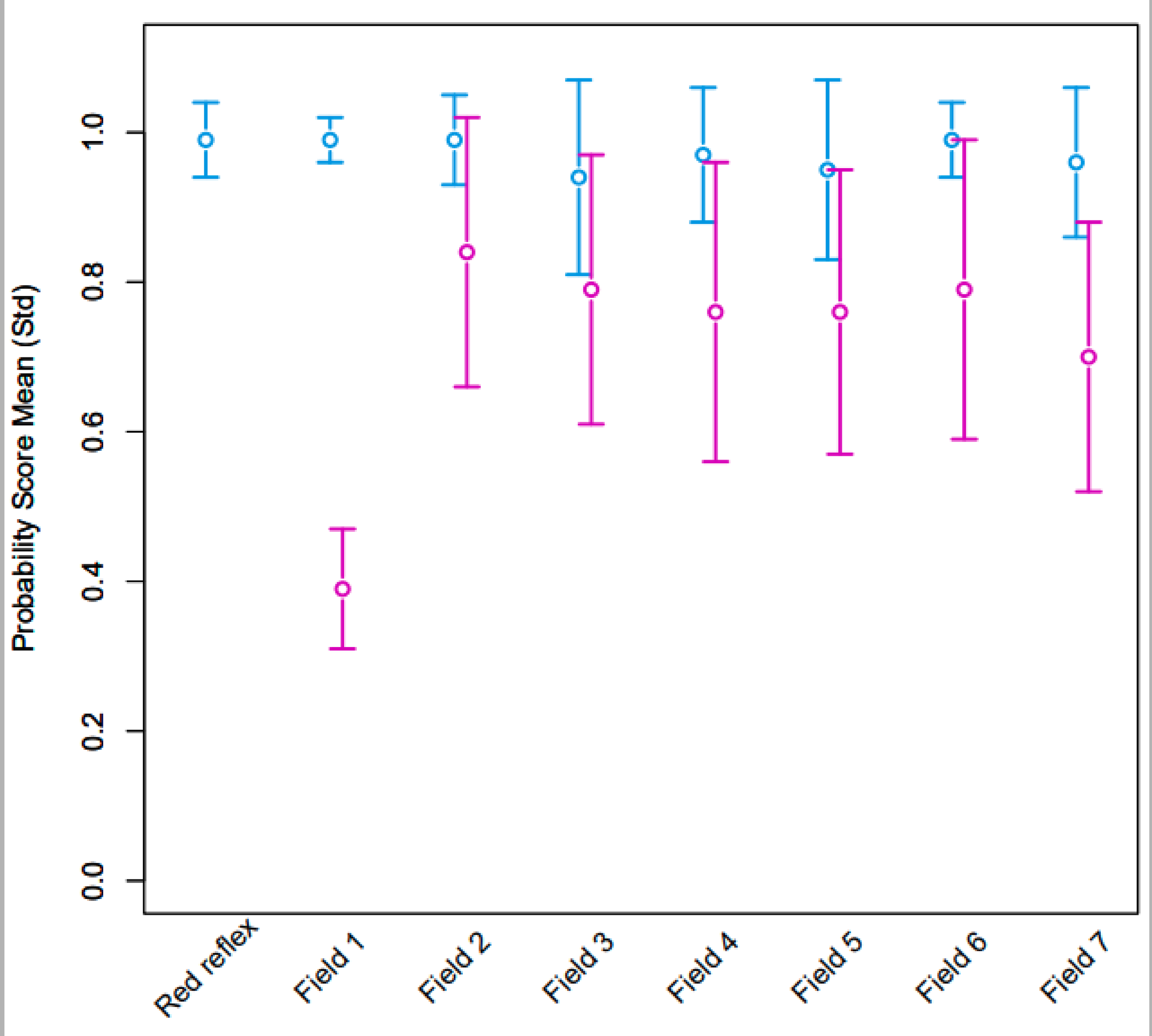


Figure 2. Scatterplot representing probability scores (0-1) for correctly labeled images (blue) and incorrectly labeled images (pink).

		Artificial Intelligence Model Generated Field Number								
		Red Reflex	Field 1	Field 2	Field 3	Field 4	Field 5	Field 6	Field 7	Total
Grader Attributed Field Number	Red Reflex	297	0	0	1	0	0	0	2	300
	Field 1	0	365	14	0	2	1	1	2	385
	Field 2	0	0	360	24	2	5	0	2	393
	Field 3	0	1	24	301	19	29	4	9	387
	Field 4	0	1	7	15	301	44	11	11	390
	Field 5	0	1	4	17	11	318	29	7	387
	Field 6	0	0	0	2	0	9	368	3	382
	Field 7	0	0	4	6	15	9	5	341	380
	Total	297	368	413	366	350	415	418	377	3004

Table 1. Confusion matrix demonstrating Cross-Classification of field images by AI and by Human Grader (per image)

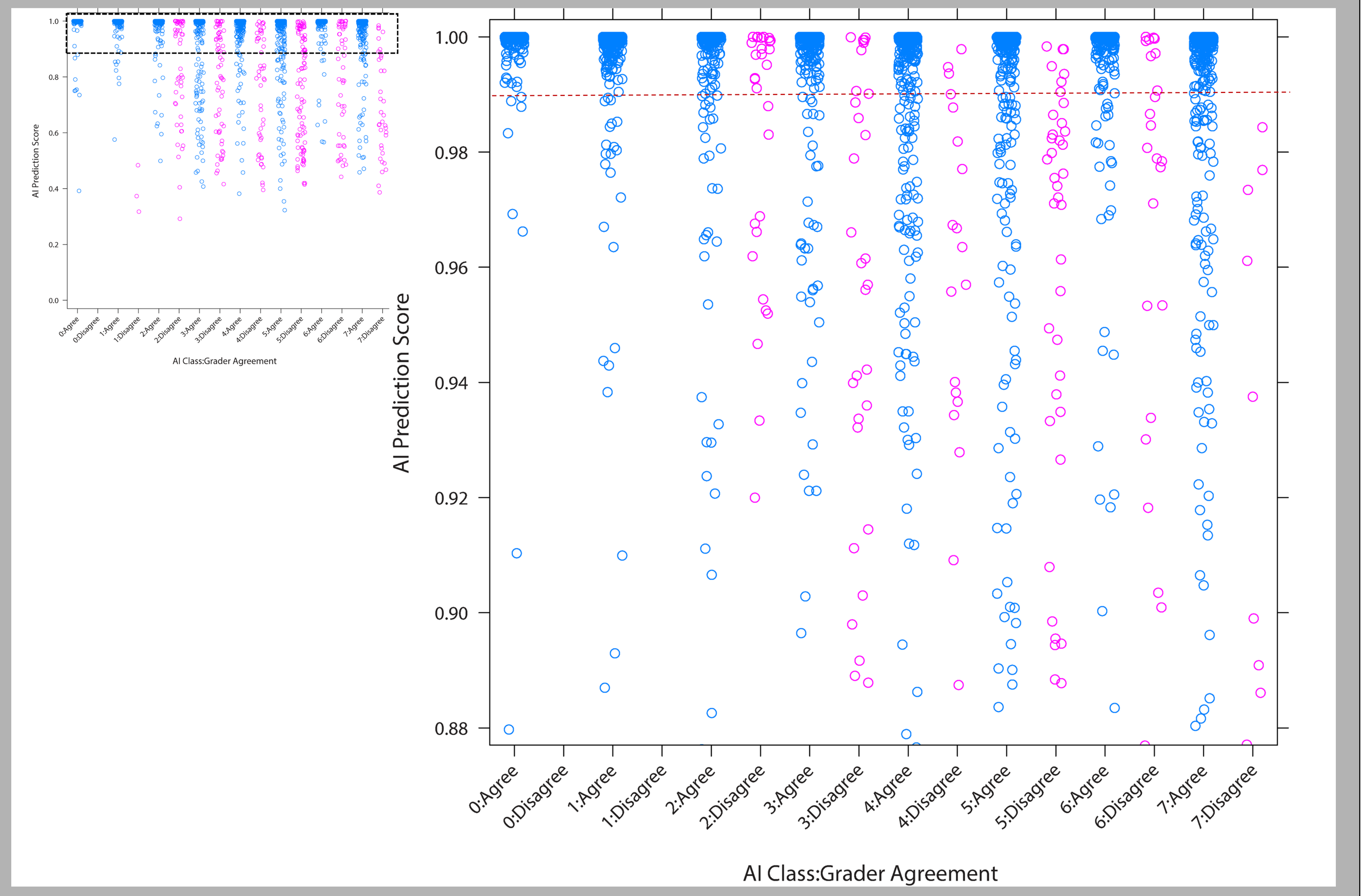


Figure 3. The scatter plots represent distribution of probability score for AI prediction that graders agreed with (blue) and disagreed with (pink) for each field number. 0 indicates red reflex and 1-7 indicates regions of retina. The smaller inset represents the full range of the probability scores. Due to skewed distribution, the larger plot provides a closer look at the higher probably scores. The correct labels (blue) are densely packed in 0.99 – 1.00 range whereas incorrect labels (pink) are distributed evenly across the spectrum.

Results

- The results of the comparison between the AI and human grader are shown in Table 1. Exact Agreement was 88% (Kappa = 0.87).
- Figure 2 shows that probability score could be a useful flag for identifying incorrect labels for grader review.
- Figure 3 shows that most correct labels have a probability score > 0.99
  - A probability cutoff of 0.99 identifies approximately 28% of images for human review. These include both correct and incorrectly labeled images.
  - With human review, incorrect labels drop from 11.7% to 1.5%
- Mismatched AI Class and Grader Field could be due to poor image quality or DR lesions present (Figure 4).

Figure 4. Examples of images with poor image focus/quality and DR lesions affecting ability to identify appropriate field

Conclusions

- AI algorithms can be effectively implemented with human review and adequate study of performance
- Tools to flag potential errors in labels generated by AI models will reduce inaccuracies, increase trust in the system and provide data for continuous model development

Figure 5. Proposed AI-enhanced workflow

References

Fundus photographic risk factors for progression of diabetic retinopathy. ETDRS report number 12. Early Treatment Diabetic Retinopathy Study Research Group. (1991). *Ophthalmology*, *98*(5 Suppl), 823–833.