AMERICAN ACADEMY
OF OPHTHALMOLOGY®

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

Amitha Domalpally, MD, PhD,[1,2] Robert Slater, PhD,[1] Nancy Barrett, MS,[2] Rick Voland, MS,[2] Rohit Balaji, BA,[2] Jennifer Heathcote, BA,[2] Roomasa Channa, MD,[1] Barbara Blodi, MD[2]

Purpose: The curation of images using human resources is time intensive but an essential step for developing artificial intelligence (AI) algorithms. Our goal was to develop and implement an AI algorithm for image curation in a high-volume setting. We also explored AI tools that will assist in deploying a tiered approach, in which the AI model labels images and flags potential mislabels for human review.

Design: Implementation of an AI algorithm.

Participants: Seven-field stereoscopic images from multiple clinical trials.

Methods: The 7-field stereoscopic image protocol includes 7 pairs of images from various parts of the central retina along with images of the anterior part of the eye. All images were labeled for field number by reading center graders. The model output included classification of the retinal images into 8 field numbers. Probability scores (0—1) were generated to identify misclassified images, with 1 indicating a high probability of a correct label.

Main Outcome Measures: Agreement of AI prediction with grader classification of field number and the use of probability scores to identify mislabeled images.

Results: The AI model was trained and validated on 17 529 images and tested on 3004 images. The pooled agreement of field numbers between grader classification and the AI model was 88.3% (kappa, 0.87). The pooled mean probability score was 0.97 (standard deviation [SD], 0.08) for images for which the graders agreed with the AI-generated labels and 0.77 (SD, 0.19) for images for which the graders disagreed with the AI-generated labels ($P < 0.0001$). Using receiver operating characteristic curves, a probability score of 0.99 was identified as a cutoff for distinguishing mislabeled images. A tiered workflow using a probability score of < 0.99 as a cutoff would include 27.6% of the 3004 images for human review and reduce the error rate from 11.7% to 1.5%.

Conclusions: The implementation of AI algorithms requires measures in addition to model validation. Tools to flag potential errors in the labels generated by AI models will reduce inaccuracies, increase trust in the system, and provide data for continuous model development. *Ophthalmology Science 2022;2:100198 © 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/).*

Supplemental material available at www.ophthalmologyscience.org.

Ophthalmology is witnessing an explosion of research in artificial intelligence (AI), with deep learning models performing various image-based tasks such as classifying diseases, predicting outcomes, and segmentation.[1] Fundus photography is one of the most commonly used imaging modalities for deep learning models, with a multitude of open source data sets.[2] However, the application of AI models in clinical care or research is still in its infancy, with a number of publications pointing to issues with applications in clinical setting.[3,4]

A large part of AI research is data curation and the development of an imaging pipeline before model training can be implemented. An adequately planned data curation process will preserve time and resources and is an important step toward the development of robust AI models.[5] To be AI ready, images need to be standardized with metadata and readily associated with corresponding labels. The format for and metadata capture using ophthalmic imaging are currently not standardized.[6] Although 3-dimensional imaging, such as OCT scans, is nuanced with proprietary formats, even simple 2-dimensional imaging, such as fundus photography, is not uniform. Fundus photographs are traditionally in TIFF or JPEG formats, which have no intrinsic information such as patient demographic data, type of image, quality metrics, annotations, measurements, labels, or imaging protocol, including field of view and area of the retina captured. The most basic information, such as laterality of the eye (right eye vs. left eye), can be missing in data sets, preventing further model development.[7] The addition of metadata to images, particularly in large data

https://doi.org/10.1016/j.xops.2022.100198
ISSN 2666-9145/22

sets, requires significant human resources. The use of deep learning to automate sorting of retinal images on the basis of imaging characteristics, such as laterality and field of view, has been particularly useful.[7–9]

Diabetic retinopathy clinical trials require the use of the well-established 7-field stereoscopic color photograph imaging protocol for the assessment of the ETDRS Severity Scale.[10] As part of this protocol, 7 different regions of the retina are imaged, with each region identified by a specific field number. This field number information is critical for curating images for the development of AI and the organization of images for grader assessment. Accurate identification of retinal fields in eyes with pathology requires knowledge of retinal anatomy and vessel orientation as well as familiarity with lesions such as laser scars and hemorrhages that can obscure vessels. Additionally, the ability to account for photographer variability in field definitions is critical. Labeling 16 images per eye (7 pairs of stereoscopic fundus images and red reflex images) in a high-volume setting, such as a reading center, is an arduous task. We developed an AI algorithm to assist with automated labeling of fields with field numbers with the aim of creating an efficient pipeline for data curation. We discuss the process of implementing an AI model in a tiered system for large-scale labeling of images.

Methods

Wisconsin Reading Center receives 7-field stereoscopic color fundus photographs from multiple sites around the world for clinical trial assessments. A complete submission includes 14 retinal images (7 pairs of stereoscopic images) per eye along with a pair of images of the anterior segment, known as red reflex images (Fig 1). The submission system automatically assembles information on study identifiers, subject identifiers, visit code, and laterality of the eye (right eye vs. left eye) into the image file name. However, information regarding the specific field of the retina captured is not readily available. The addition of a field label is usually completed manually using software that adds the field number to the metadata to allow sorting of images.

Institutional review board approval was obtained at each clinical site as part of the clinical trial, and written informed consent was obtained from all study participants. The research was also approved by the University of Wisconsin Institutional Review Board and conducted according to the Declaration of Helsinki.

Training and Testing Data Sets

Images from various clinical trial submissions that had 7-field stereoscopic images with retinal fields previously labeled by Wisconsin Reading Center graders were included in the training data set. The graders were masked to all patient demographics. The graders labeling the fields had ≥ 2 years of experience in evaluating diabetic retinopathy and a good understanding of field

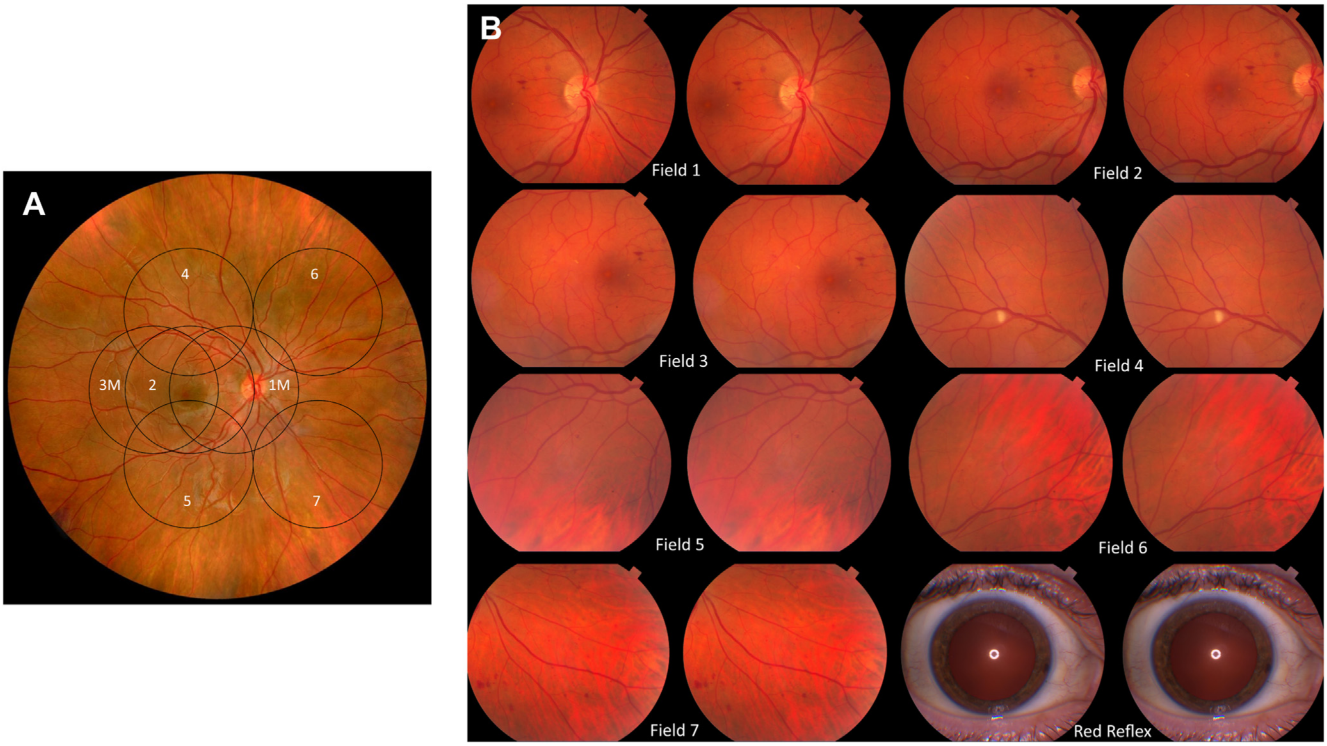

Figure 1. **A,** An ultrawide retinal image showing 7 circles corresponding with the 7-field image. **B,** Images of each circle taken using a 30° camera are submitted. Two images are submitted per field to provide a stereoscopic view for graders; 14 retinal images are provided per eye. The proof sheet is representative of images submitted as part of the 7-field stereoscopic imaging protocol, including 7 regions of the retina and 1 pair of images of the anterior part of the eye (red reflex). Each image is labeled to identify the field number on the basis of the region of the retina photographed. An artificial intelligence model was developed to classify and label the field numbers.

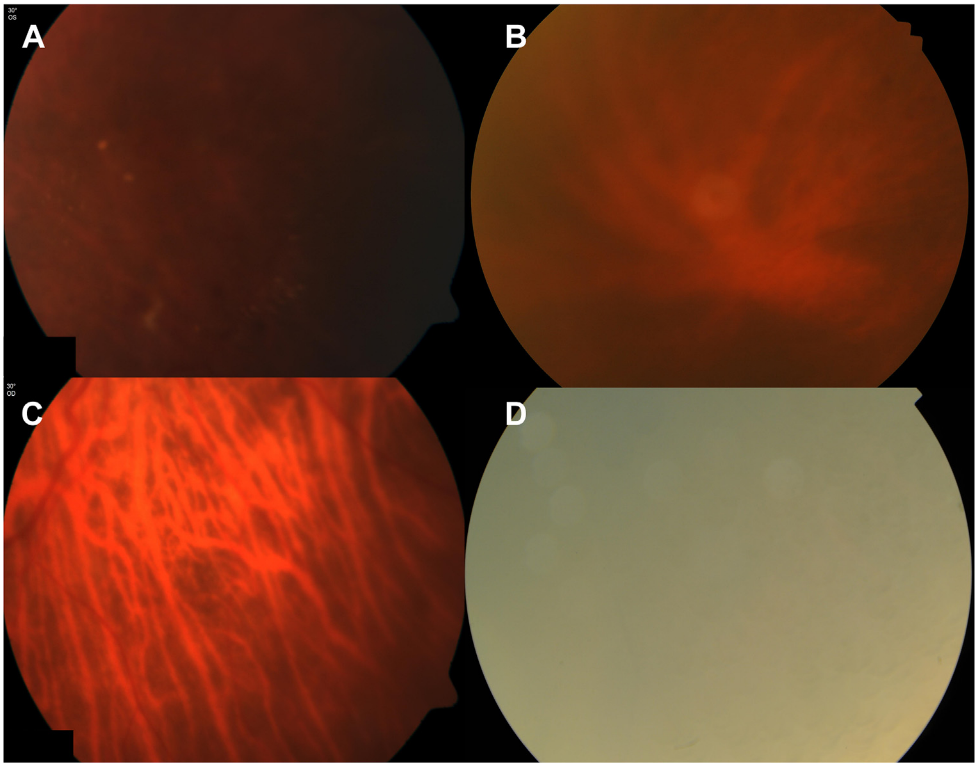

Figure 2. Examples of poor-quality images for identifying the field of the retina. The retinal vasculature is not visible in the images because of poor contrast and dark images (**A**), poor focus and central artifact (**B**), dense choroidal vessels (**C**), and media haze and overexposed images (**D**).

definitions. The training data set included 17 529 images from 229 unique subjects, recruited from 42 clinical sites, and included 162 right eyes and 176 left eyes. Multiple visits from the same subject were included. Images were captured using 3 different digital fundus cameras, including Topcon TRC 50, Zeiss FF450, and Kowa VX-20. The data were split as 80% for training and 20% for internal validation. The data output by the AI model included both field numbers and probability scores. The probability scores ranged from 0 to 1, with 0 indicating the least probability of a correct label and 1 indicating the highest probability of a correct label.

The testing data set included 3004 images from 240 eyes, 40 subjects, and 10 sites. The testing data set included images captured using Topcon TRC 50 (75.3%), Zeiss FF450 (4.5%), and Kowa VX-20 (20.2%). It was ensured that the clinics in the training data set were not repeated in the testing data set because a single clinic can submit images for multiple clinical trials.

After testing was completed, the algorithm was deployed on the first 11 382 images that were submitted to the reading center.

Ground Truth. All images in the testing data set were independently labeled by 2 graders (R.B. and J.H.) for the red reflex image and 7 stereoscopic retinal fields (Fig 1). In addition, the graders documented the confidence score for each image using a 3-level approach: (1) a high confidence score indicates that the field definition adheres to protocol requirements, (2) a moderate confidence score indicates that the field definition departs from protocol requirements but that the field can be identified, and (3) a low confidence score indicates that the field definition departs significantly from protocol requirements or the images are of significantly poor quality and that the field cannot be identified. Examples of poor-quality images are provided in Figure 2, and the reasons for these range from operator issues, such as poor field definitions, to patient-related issues, such as media haze. The

Table 1. Confusion Matrix Showing Comparison of Grader-Attributed Field Number with Artificial Intelligence–Generated Field Number

		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 2. Performance Measures of the Multiclass Artificial Intelligence Model for Field Numbers

	Accuracy	Kappa	Sensitivity	Specificity	Precision	F1 Score	AUROC
Red reflex	1.00	0.99	0.99	1.00	1.00	0.99	1.00
Field 1	0.99	0.97	0.95	1.00	0.99	0.97	1.00
Field 2	0.97	0.88	0.82	0.98	0.87	0.89	0.99
Field 3	0.95	0.77	0.78	0.98	0.82	0.80	0.97
Field 4	0.95	0.79	0.77	0.98	0.86	0.81	0.96
Field 5	0.94	0.76	0.82	0.96	0.77	0.79	0.98
Field 6	0.98	0.91	0.96	0.98	0.88	0.92	0.99
Field 7	0.98	0.89	0.90	0.99	0.90	0.90	0.99

AUROC = area under the receiver operating characteristic curve.

graders first identified fields 1 and 2, which are the easiest to identify on the basis of the inclusion of the optic nerve, followed by the identification of peripheral fields. In some cases, an image series of a particular eye had some images of poor quality that the grader could still label on the basis of the exclusion of previously labeled fields. The order of images in the folder varied depending on the site and was not factored into field renaming.

Model Training

Two distinct models were trained in an identical fashion with identical parameters, for which only the training data differed: one was trained on the left eye and the other on the right eye. This was done because information about eye laterality was known. No augmentation was done on the training data. The EfficientNetB0 architecture (including the top classifier) from Tensorflow was used, with 8 classes as the output. The model used average pooling. The Adam optimizer, with a learning rate of 0.001, β_1 of 0.9, and β_2 of 0.999, was used to optimize the sparse categorical cross-entropy loss. The training results were similar, with a lower learning rate.

Training was repeated until the level of validation loss did not improve for 3 epochs. The weights that returned the least (best) level of loss were then saved for each eye modality. Training was processed using a single Tesla V100.

All images were center cropped to be square in their original format and then resized to 256 × 256 pixels with 3 color channels (red, green, and blue). A batch size of 8 was used, with no image augmentation.

Model Evaluation

The model output of field number was compared with the reading center graders' assigned field numbers on the testing data set. All metrics for comparison were collected, including accuracy, precision, sensitivity, specificity, kappa, area under the curve, and F1 score. A confusion matrix was generated to visualize the agreement and distribution of mislabeled images across the 8 field numbers. The distribution of probability scores, ranging from 0 to 1, in eyes with and without disagreement was summarized as means and medians. Because of their skewed distribution with long tails, an enlarged scale was used to visualize the distribution of probability scores. Stacked scores of all 8 classes were used to generate receiver operating characteristic curves to determine the probability scores for workflow implementation.

Results

The deep learning model output included field numbers and probability scores. The grader output on the testing data set included field numbers and confidence scores. The comparison of the model output with the graders' field numbers showed agreement for 2651 (88.3%; kappa, 0.87) images (Table 1). This implied that implementing an automated workflow with the model would result in mislabeling of 12% of the data set. The performance measures of the multiclass AI model for each field number are shown in Table 2. Although the overall model accuracy was high for every field number, there was variability within each class.

Rather than a fully automated renaming system, we decided to use a tiered system with human override of mislabeled images. To assist in flagging eyes with potential errors, we explored the use of AI model-generated probability scores as a potential marker. The mean AI-generated

Table 3. Distribution of Probability Scores in Images with Incorrect and Correct Labels

	Correct Labels, n (%)	Incorrect Labels, n (%)	Probability Score for Correct Labels, Mean (SD)	Probability Score for Incorrect Labels, Mean (SD)
Red reflex	297 (100)	0 (0)	0.99 (0.05)	NA
Field 1	365 (99)	3 (0.8)	0.99 (0.03)	0.39 (0.08)
Field 2	360 (87)	53 (13)	0.99 (0.06)	0.84 (0.18)
Field 3	301 (82)	65 (18)	0.94 (0.13)	0.79 (0.18)
Field 4	301 (86)	49 (14)	0.97 (0.09)	0.76 (0.20)
Field 5	318 (77)	97 (23)	0.95 (0.12)	0.76 (0.19)
Field 6	368 (88)	50 (12)	0.99 (0.05)	0.79 (0.20)
Field 7	341 (90)	36 (10)	0.96 (0.10)	0.70 (0.18)

NA = not available; SD = standard deviation.

Table 4. Association of Grader Confidence Scores with Artificial Intelligence Probability Scores

Grader Confidence	Correct Labels for Field Number, n (%)	Incorrect Labels for Field Number, n (%)	Prediction Score for Correct Labels, Mean (SD)	Prediction Score for Incorrect Labels, Mean (SD)
High confidence	2484 (93)	195 (7)	0.98 (0.07)	0.76 (0.20)
Moderate confidence	145 (56)	115 (44)	0.89 (0.17)	0.79 (0.19)
Low confidence	22 (34)	43 (66)	0.78 (0.20)	0.80 (0.20)

SD = standard deviation.

probability scores were higher when the graders agreed with the AI prediction (i.e., correct label [0.94—0.99]) and lower when the grader disagreed with the AI prediction (i.e., incorrect label [0.39—0.84]), indicating their potential utility in the identification of mislabeled images (Table 3). This difference in the mean probability scores between correct and incorrect labels remained similar across all fields.

Along with field numbers, the graders provided confidence scores on the validation data. Grader confidence scores give an understanding of human certainty while labeling images; a lower image quality is denoted by low confidence scores. Probability scores are a similar metric of certainty for the AI model. We explored whether a relationship exists between grader confidence scores and AI probability scores. In the data set, 2679 (89.1%) images were labeled with a high confidence score by the graders, 260 (8.6 %) with a medium confidence score, and 65 (2.1%) with a low confidence score. The overall AI probability scores were highly correlated with the grader confidence scores, with a mean of 0.96 (standard deviation, 0.10) in high confidence, 0.85 (standard deviation, 0.19) in medium

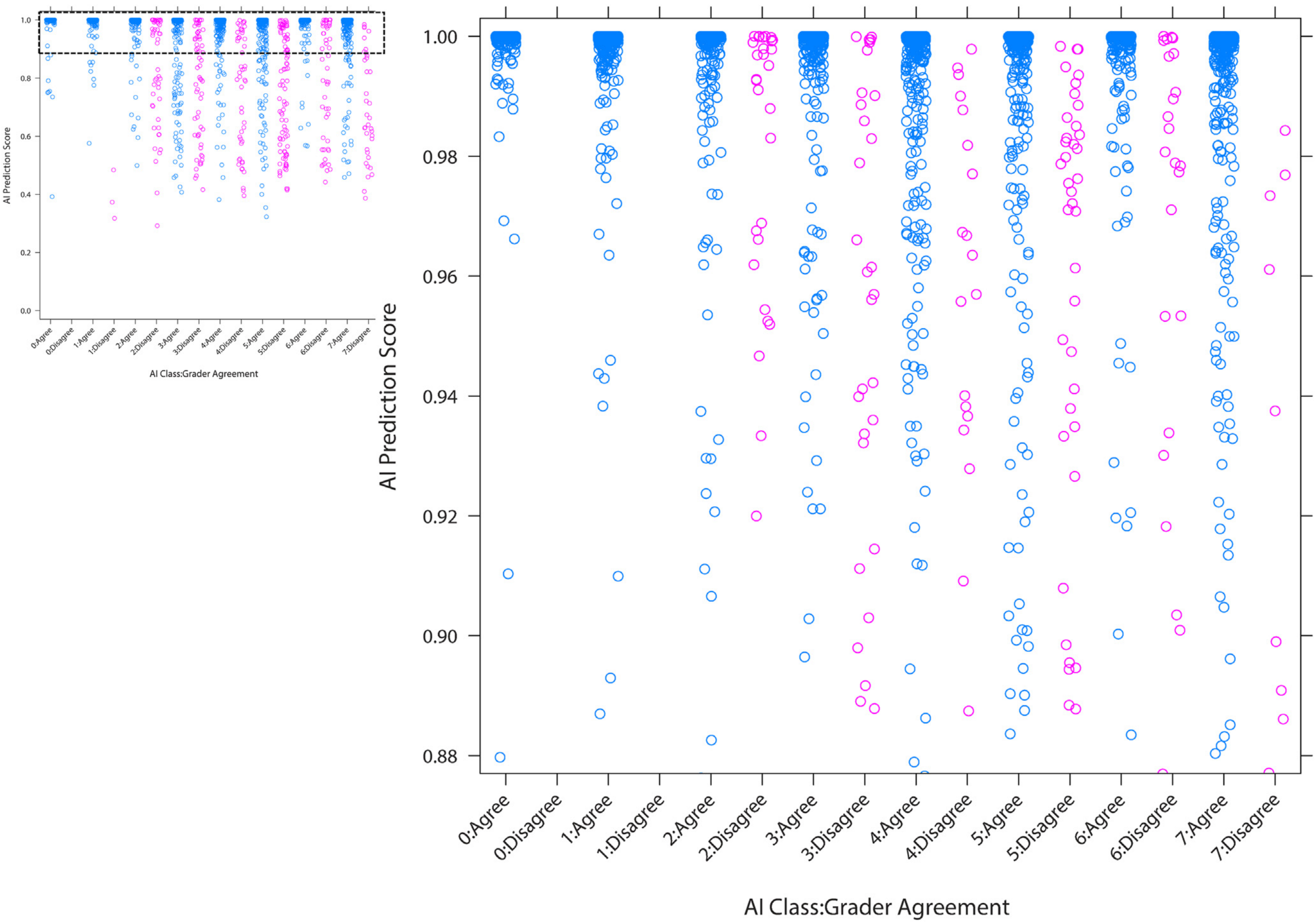

Figure 3. The scatter plots represent the distribution of probability scores for artificial intelligence (AI) prediction that graders agreed (blue) and disagreed with (pink) for each field number. Zero indicates red reflex, and 1 to 7 indicates regions of the retina. The smaller inset represents the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

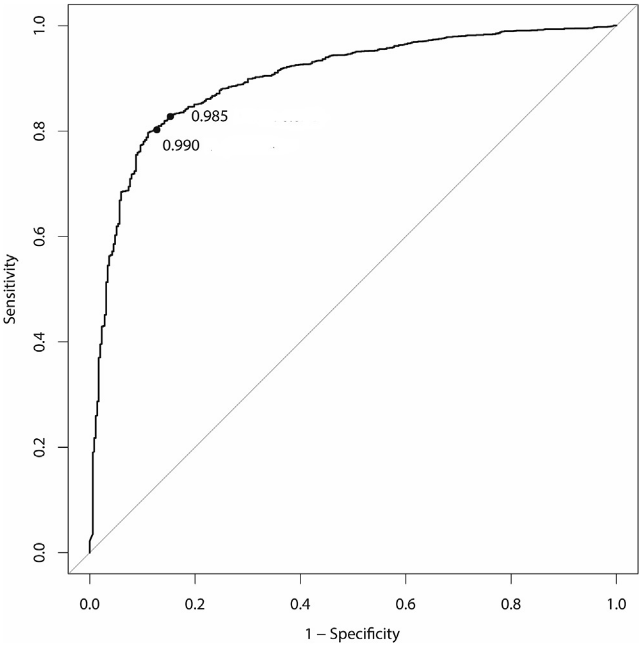

Figure 4. Receiver operating characteristic curve showing the sensitivity and specificity of probability scores. The sensitivity and specificity for the 2 ideal cutoff for probability scores 0.99 and 0.985 are shown in parentheses.

confidence, and 0.79 (standard deviation, 0.20) in low confidence ($P < 0.0001$). We further investigated the relationship between incorrect labels and probability scores on the basis of grader confidence scores (Table 4). The frequency of incorrect labels was 7% in images with a high grader confidence score and 66% in those with a low grader confidence score. In images with high and moderate grader confidence scores, the probability scores were distinguishable for correct and incorrect labels: a probability score of 0.89 to 0.98 in eyes with correct labels versus 0.76 to 0.79 in eyes with incorrect labels. In images with a low grader confidence score, the probability scores were overall low and did not distinguish between correctly and incorrectly labeled images (0.78 vs. 0.80, respectively). This indicated that the probability scoring system was a helpful flag for the identification of errors in good-quality images but faltered in poor-quality images. This implies that, in poor-quality images, the model seems to give a low probability score despite being accurate. Building a quality assessment filter for the model could potentially help reduce these errors.

On the basis of the abovementioned data, it has been established that the probability score cutoff could be a useful metric for the identification of incorrect labels. It is also evident that a probability score cutoff would include a mix of both incorrect and correct labels. The next set of analyses focused on identifying the cutoff to maximize the detection of incorrect labels. Figure 3A, B shows the distribution of correct (blue) and incorrect (pink) labels with probability scores. There were a few incorrect labels, with the highest probability score of 1.0, and conversely, correct labels had a low score of 0.4. The distribution of probability scores was skewed to the upper end of the probability scores and prevented the visualization of a cutoff. The smaller inset Figure 2A shows the full range of probability scores, and Figure 3B depicts an expanded view of the distribution of probability scores, describing the sparse distribution of incorrect labels in the high range of probability scores. Correct labels were densely packed in the region with a range of 0.99 to 1.0, whereas incorrect labels seemed to be evenly distributed.

On the basis of the receiver operating characteristic curve (Figure 4), both 0.985 and 0.99 were contenders for a threshold for triaging images. The sensitivity and specificity were 0.87 and 0.80, respectively, for a probability score of 0.99 and 0.85 and 0.82, respectively, for a probability score of 0.985. We examined the true- and false-positive balance with each of these cutoffs (Table S1, available at www.ophthalmologyscience.org). When applied to the validation data set of 3004 images, a cutoff point of 0.99 would flag 831 (27.7%) images to be reviewed by the grader. Of these, 308 (10.3%) would require a grader override (incorrect label), and the remaining 523 (17.4%) would be reviewed, although they are accurately labeled by AI. This cutoff of 0.99 would permit 45 (1.5%) images with incorrect labels to slip through the system. Implementing this threshold of 0.99 for flagging incorrect labels would reduce the number of mislabeled images from 353 (11.7%) in a fully automated workflow to 45 (1.5%) in a tiered system with human oversight.

After validation, we deployed the model prospectively in a workflow system that would flag any image that received a probability score of < 0.99 on the first set of submissions after model development. Out of the 11 382 images received, 2905 (25.5%) images were identified as having a probability score of < 0.99, requiring grader review. These percentages are in line with the review rate of 27.6% predicted on the basis of the testing data set and support our use of the tiered labeling workflow. The time taken to deploy the algorithm for the 11 382 images was approximately 12 minutes (including accessing files from a network drive and resizing the images) using the Intel i5 processor. The average grader time for the renaming process is 7 minutes for an eye, amounting to 5173 minutes for the same data set. Using the tiered approach, the grader time was reduced to 25% (i.e., 1293 minutes). The average cost saving of employing AI in this large-volume scenario was approximately $2000 for every 10 000 images.

Discussion

In this study, we explored the process of implementing a deep learning model for image curation in a high-volume setting. The focus of this paper was an exercise on the implementation pathway and the explainable AI outputs needed to support the workflow. The purpose of the AI model is to automate and enhance the workflow and reduce the burden on reading center graders, whereas the implementation of this workflow analyzes ways to maximize automation and reduce incorrect field labels. To achieve this balance, we used probability scores to define the threshold of sensitivity and specificity acceptable for the study.

The data used for AI research need to follow the Findability, Accessibility, Interoperability, and Reusability principles.[11] The preparation of data for AI research following the Findability, Accessibility, Interoperability, and Reusability principles includes many components: implementing required ethical reviews, deidentifying protected health information, ensuring comprehensive metadata, structuring images in a homogenized and machine-readable format, and linking the images to the ground truth.[5] The challenges of the ground truth have been previously discussed in publications because these serve as the basis of model validation.[12–14] However, more complex processes, including image standardization and metadata structuring, have only been recently highlighted.[6]

Like most deep learning algorithms, the model developed for this study can perform only a single task. Unlike most binary classifiers, it is unique in being a multilevel classifier, differentiating between 8 classes. However, this deep learning model is limited to classifying field numbers and cannot identify other image characteristics, such as image eye laterality. It is also limited to 35° images, which are not as common as wide-field images but still constitute 30% to 40% of clinical trial submissions for diabetic retinopathy at Wisconsin Reading Center. A comprehensive image curation platform would be a string of AI algorithms that can determine basic metadata from an image, such as type of image, eye laterality, field degree, field position, and image quality metrics.

The model has a high level of accuracy (97%) for classifying field numbers. Fields 1 and 2 had the best accuracy possibly because of the presence of the optic nerve in both the fields, which helped distinguish them from other fields. Fields 3 to 5 had a lower accuracy (94%–95%) than other fields, possibly because of overlapping fields in the temporal retina that can be confused with each other. The most frequently mislabeled field was field 5 (inferotemporal quadrant), with an error rate of 5.5% compared with field 1 (optic nerve centered), which had an error rate of 0.7%. On the basis of the confusion matrix, most errors were due to the misattribution of field 4 (superotemporal) or 3 (temporal) as field 5 by the deep learning model. In comparison, field 7 had a lower error rate but a wider distribution of mislabels, wherein the AI model made an erroneous prediction across all field numbers, even calling red reflex images as field 7 in 2 images. In both these eyes, the red reflex images were zoomed out further than usual. In a multiclass classifier, apart from the frequency of errors, the type and distribution of errors help to fine tune the model. In this case, although the error rate was higher in field 5, it is the field 7 classifier that requires further investigation. A noteworthy difference in field labeling between an AI model and a human grader is that a grader can identify fields by the process of elimination. For humans, once the optic nerve and macular fields are identified, only 1 peripheral field needs to be labeled confidently before the rest fall into place. The deep learning model labels each image as an independent class and does not approach the image series as dependent classes. Increased image input size and treating images as a set rather than independent data have the potential to greatly increase performance. The AI model we developed uses a set of 7 large pictures rather than 1 large picture, which is a current technical limitation.

A major challenge for implementing AI-enabled automated workflows is lack of trust. To address this, various forms of explainable models have been introduced, the most common of which are attribution maps.[4] Reviewing individual images for attribution maps is a good solution in smaller data sets. In high-volume data sets, a more quantitative metric that allows triaging of images is preferred.[15] Probability scores are an easily interpretable metric of model confidence for assigning a class. The model probability scores were highly correlated with the grader confidence scores and served as a useful metric to separate out mislabeled images, especially for good-quality images. The triaging system enables a review of the majority of misclassified images along with an opportunity to review correctly labeled ones. This feedback system helps us understand the model better and can serve toward continued model development.[16] Over time, fine tuning the model by retraining it on corrected image labels will help reach a stage where low probability scores can mostly be limited exclusively to mislabeled images.

We are in an exciting era of AI implementation where integration standards are being defined.[17] In this article, we described our thought process and the factors involved in implementing AI in a workflow using an example of a multiclass classifier model and probability scores for identifying mislabeled retinal images. As AI-based research continues to advance in ophthalmology, we will need to invest more research into tools that foster trust.

Footnotes and Disclosures

Originally received: March 20, 2022.
Final revision: June 9, 2022.
Accepted: July 7, 2022.
Available online: July 13, 2022. Manuscript no. XOPS-D-22-00052R1.

[1] A-EYE Unit, Department of Ophthalmology and Visual Sciences, University of Wisconsin, Madison, Wisconsin.

[2] Wisconsin Reading Center, Department of Ophthalmology and Visual Sciences, University of Wisconsin, Madison, Wisconsin.

Disclosures:

All authors have completed and submitted the ICMJE disclosures form.

The authors have no proprietary or commercial interest in any materials discussed in this article.

Supported in part by an unrestricted grant from Research to Prevent Blindness, Inc. to the University of Wisconsin Madison Department of Ophthalmology and Visual Sciences.

HUMAN SUBJECTS: Human subjects were included in this study. Institutional review board approval was obtained at each clinical site as part of the clinical trial. The research was approved by the University of Wisconsin Institutional Review Board and conducted according to the Declaration of Helsinki. Written informed consent was obtained from all study participants.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Domalpally, Slater, Blodi

Data collection: Domalpally, Slater, Barrett, Balaji, Heathcote

Analysis and interpretation: Domalpally, Slater, Barrett, Voland, Channa, Blodi

Obtained funding: Study was performed as part of regular workflow at the Wisconsin Reading Center. No additional funding was obtained.

Overall responsibility: Domalpally, Slater, Channa, Blodi

Abbreviations and Acronyms:
AI = artificial intelligence.

Keywords:
Artificial intelligence, Deep learning, Fundus photograph, Image curation, Machine learning, Metadata, Retinal image, Standardization.

Correspondence:
Amitha Domalpally, MD, PhD, A-EYE Unit, Department of Ophthalmology and Visual Sciences, University of Wisconsin School of Medicine and Public Health, 310 N. Midvale Blvd, Suite 205, Madison, WI 53705. E-mail: domalpally@wisc.edu.

References

1. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, et al. Artificial intelligence in retina. *Prog Retin Eye Res*. 2018;67:1−29.
2. Khan SM, Liu X, Nath S, et al. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Dig Health*. 2021;3:e51−e66.
3. Baxter SL, Lee AY. Gaps in standards for integrating artificial intelligence technologies into ophthalmic practice. *Curr Opin Ophthalmol*. 2021;32:431−438.
4. Domalpally A, Channa R. Real-world validation of artificial intelligence algorithms for ophthalmic imaging. *Lancet Dig Health*. 2021;3:e463−e464.
5. Willemink MJ, Koszek WA, Hardell C, et al. Preparing medical imaging data for machine learning. *Radiology*. 2020;295:4−15.
6. Lee AY, Campbell JP, Hwang TS, et al. Recommendations for standardization of images in ophthalmology. *Ophthalmology*. 2021;128:969−970.
7. Liu P, Gu Z, Liu F, et al. Large-scale left and right eye classification in retinal images. In: Stoyanov D, Taylor Z, Ciompi F, et al., eds. *Computational Pathology and Ophthalmic Medical Image Analysis. Lecture Notes in Computer Science*. Cham: Springer International Publishing; 2018:11039.
8. Rim TH, Soh ZD, Tham YC, et al. Deep learning for automated sorting of retinal photographs. *Ophthalmol Retina*. 2020;4:793−800.
9. Lai X, Li X, Qian R, et al. Four models for automatic recognition of left and right eye in fundus images. In: Kompatsiaris I, Huet B, Mezaris V, et al., eds. *MultiMedia Modeling*. Cham: Springer International Publishing; 2019:11295. https://link.springer.com/chapter/10.1007/978-3-030-05710-7_42.
10. 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.
11. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data*. 2016;3:1−9.
12. Abràmoff MD, Cunningham B, Patel B, et al. Foundational considerations for artificial intelligence using ophthalmic images. *Ophthalmology*. 2021;129:e14−e32.
13. Nakayama LF, Gonçalves MB, Ferraz DA, et al. The challenge of diabetic retinopathy standardization in an ophthalmological dataset. *J Diabetes Sci Technol*. 2021;15:1410−1411.
14. Krause J, Gulshan V, Rahimy E, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. *Ophthalmology*. 2018;125:1264−1272.
15. McCrindle B, Zukotynski K, Doyle TE, Noseworthy MD. A radiology-focused review of predictive uncertainty for AI interpretability in computer-assisted segmentation. *Radiol Artif Intell*. 2021;3:e210031.
16. Lee CS, Lee AY. Clinical applications of continual learning machine learning. *Lancet Digit Health*. 2020;2:e279−e281.
17. Wiggins WF, Magudia K, Schmidt TM, et al. Imaging AI in practice: a demonstration of future workflow using integration standards. *Radiol Artif Intell*. 2021;3:e210152.