Democratizing Clinican-driven AI: Self-training with Code-free AutoML and Public Data

Authors

*Edward Korot MD[1,2,3], Mariana Batista MD[2,6,7], Josef Huemer MD[2], Sara Beqiri BS[2,5], Hagar Khalid MD[2,4], Madeline Kelly MS[2,5,9], Mark Chia MD[2], Emily Mathijs MS[8], Robbert Struyven MS[2], Magdy Moussa MD[4], Pearse A Keane MD [2]

*Corresponding Author:
Edward Korot, MD
835 Fairview Ave NE Unit 1
Grand Rapids MI, 49503

Manuscript Word Count: 2,892

E-mails: ekorot@gmail.com (Edward Korot), mari.batista.124@gmail.com (Mariana Batista), josef.huemer1@nhs.net (Josef Huemer), sara.beqiri.17@ucl.ac.uk (Sara Beqiri), hagar.khalid@gmail.com (Hagar Khalid), madge.kelly@hotmail.com (Madeline Kelly), mark.a.chia@gmail.com (Mark Chia), mathijse@msu.edu (Emily Mathijs), robbertstruyven@gmail.com (Robbert Struyven), magdymoussa60@gmail.com (Magdy Moussa), pearsek@gmail.com (Pearse A Keane)

1. Retina Specialists of Michigan, 5030 Cascade Rd SE, Grand Rapids, MI 49546
2. Moorfields Eye Hospital, 162 City Rd, London EC1V 2PD, UK
3. Stanford University Byers Eye Institute, 2452 Watson Ct, Palo Alto, CA 94303
4. Ophthalmology Department, Faculty of Medicine, Tanta University Hospital, El-Gaish St, Tanta, Gharbia, Egypt, 31111
5. University College London Medical School, Gower St, London WC1E 6BT, UK
6. Federal University of Sao Paulo, 822 Botucatu St, Sao Paulo 04023062, Brazil
7. Instituto da Visão (IPEPO), 1083 Borges Lagoa St, Sao Paulo 04038032, Brazil
8. Michigan State University College of Osteopathic Medicine, 965 Wilson Rd, East Lansing, MI 48824
9. UCL Centre for Medical Image Computing, 90 High Holborn FL1, London WC1V 6LJ, UK

Key Points

Question: Can clinicians without coding expertise or access to well-labeled private datasets use self-training and AutoML to create high-performing machine learning models?

Findings: The models designed without coding, private datasets, or extensive labeling demonstrated high performance comparable to bespoke and FDA approved models for similar tasks.

Meaning: These findings demonstrate the tools enabling democratization of clinician-driven AI.

Abstract

Importance: Democratizing AI to enable model development by clinicians with a lack of coding expertise, powerful computing resources, and large well-labeled datasets.

Objective: Determine whether resource-constrained clinicians can utilize self-training via AutoML and public datasets to design high-performing machine learning models.

Design: In this diagnostic study, a self-training method without coding is employed on public datasets. An AI model was trained to classify referable diabetic retinopathy as an exemplar use-case. This study was conducted in 2021.

Setting: Datasets were comprised of retinal images from patients in France, the United Kingdom, the United States, and Egypt.

Participants: This study used the freely-accessible datasets EyePACS (n=58,689) and Messidor-2 (n=1748). The Messidor-2 images were assigned adjudicated labels available on Kaggle. Four images from Messidor-2 were deemed ungradable and excluded, leaving 1744 images. 300 images randomly selected from the EyePACS dataset were independently re-labeled by three blinded retina specialists using the international classification of diabetic retinopathy protocol for diabetic retinopathy grade and diabetic macular edema presence. 19 images were deemed ungradable which left 281 images.

Exposures: Using public datasets, we trained a “teacher model” with labeled images using supervised learning. Next, we utilized the resulting predictions, termed “pseudo-labels”, on an unlabeled public dataset. Finally, a “student model” was trained with the existing labeled images and the additional pseudo-labeled images.

Main Outcomes and Measures: The analyzed metrics for our models included area under the receiver operating curve (AUROC), accuracy (ACC), sensitivity, specificity, and F1 score.

Results: Teacher model performance on our internal validation datasets ranged from AUROC (0.886-0.939) while student model AUROC ranged from (0.916-0.951).

Teacher AutoML model external validation performance was AUROC, ACC: (0.964), (93.3%) while the student model was (0.950), (96.7%) respectively and our manually coded bespoke model was (0.985), (96.5%).

Conclusion and Relevance: This study suggests that self-training using AutoML is an effective method to increase both model performance and generalizability while decreasing the need for costly expert labeling. Our approach advances the democratization of AI by enabling clinicians without coding expertise or access to large well-labeled private datasets to develop their own AI models

Introduction

Machine learning (ML), specifically deep learning (DL) algorithms, are promising tools for medical image interpretation, and have been increasingly investigated in imaging-heavy specialties such as radiology, dermatology, and ophthalmology [1]. However, not many models make it into production [2,3]. *Clinician-driven* ML has a higher likelihood of implementation, since clinicians on the front lines are best suited to select relevant use-cases and design ML for patient-relevant endpoints. Despite recent advances, supervised learning (SL), the most utilized form of DL, relies on large amounts of labeled data [4] [5]. The scarcity of expert clinician time is a significant challenge when performing medical image labeling [6–8]. Given that ML model performance is highly reliant on high-quality and reproducible ground truth labels, this entails time-consuming and costly labeling efforts for healthcare ML[9] [10] [11]. Accordingly, *clinician-driven* DL projects are chiefly limited by the need for (1) coding expertise, (2) powerful computing resources, and (3) time and cost of generating high-quality adjudicated labels [12].

A major barrier to clinician-driven ML is a lack of technical and coding expertise. Our group has demonstrated a potential solution, automated machine learning (AutoML). This framework enables ML model building *without coding* by largely automating the steps of the ML pipeline including dataset management, neural architecture search, and hyperparameter tuning [*13*]. A number of publicly-available AutoML platforms enable domain experts, including health care professionals, without coding expertise to train their own high-performing DL models [14–16]. Our aforementioned work demonstrates high

performance and the ability to reproduce published seminal models demonstrating novel signals such as sex prediction from retinal fundus photographs [17,18].

Semi-supervised learning (SSL) is a potential solution for label scarcity, as it makes use of unlabeled data, which is vastly more abundant [19]. A type of SSL termed self-training, entails initially training a “teacher model” with a labeled dataset using SL, then utilizing the resulting teacher model to generate predictions, termed “pseudo-labels”, on another unlabeled dataset. Next, a “student model” is trained utilizing an expanded dataset consisting of the initially labeled dataset combined with the additional dataset and its pseudo-labels (Figure 1). This process may be repeated (i.e. designating the student model as a teacher model to generate new pseudo-labels, and subsequently training a new student model iteration) in order to both learn generalized task representations and increase algorithm performance [19–21].

We illustrate this framework through the exemplar disease of diabetic retinopathy (DR), which is a leading cause of visual impairment, affecting up to a third of patients with diabetes mellitus [22,23]. According to projections, diabetes will affect 439 million adults by 2030, and 2.4 million eyes per day would require retinal examination worldwide [24,25]. To deal with this public health problem, several diabetic screening programs have been established over time, such as the English National Screening Programme and the Singapore Integrated Diabetic Retinopathy Program (SiDRP) [26–28]. Furthermore, several AI models have been FDA approved for DR classification and referral [29,30].

Herein we demonstrate a major step toward ML democratization by leveraging AutoML, SSL, and public datasets to design models without coding while utilizing limited data and labels. Using this synergistic approach, we address three primary barriers to democratizing clinician-driven ML: coding expertise, computing resources, and large well-labeled datasets. Accordingly, we train a representative model to classify referable DR, demonstrating performance similar to commercial FDA approved algorithms.

Methods

Public Datasets

Public, freely-accessible datasets were utilized both to enable replication of our work, and to demonstrate that the task may be achieved without access to institution-specific datasets and label sets (Figure 2) [31].

Two retinal fundus photo datasets were selected: EyePACS (n=58,689 .jpeg images) and Messidor-2 (n=1748 .png images) [32,33]. “Gold standard” labels from Kaggle, as assigned and adjudicated through an iterative process involving multiple rounds by three retina specialists, were applied to the Messidor-2 images [9] to train the teacher model. Four images were adjudicated as ungradable and were excluded, leaving 1744 images. Messidor fundus photos were obtained using Topcon TRC NW6 non-mydriatic fundus camera with a 45 degree field of view. Patient inclusion criteria and demographics for these datasets are published in accordance with the source datasets. DR grades were assigned per international classification of diabetic retinopathy (ICDR) protocol for both datasets, and diabetic macular edema (DME) was defined by hard exudates within 1 disc diameter of the fovea. In order to replicate the most common referral triage task encountered in screening programs and performed by FDA/CE approved DL models, labels were binarized to referrable DR (RDR), comprising DR grades of moderate, severe, proliferative, and/or the presence of diabetic macular edema (DME), and non referable DR (NRDR), which represents the absence of RDR [7,8,27,30,34–36].

To ensure high label quality for the EyePACS internal validation subset, 300 images were randomly selected from the EyePACS dataset, which were independently re-labeled by three blinded retinal specialists (MB, JH, HK) using the ICDR protocol for DR grade and DME presence. The images were recategorized to RDR and NRDR as described, and the majority grade was assigned for cases of disagreement. 19 images were deemed ungradable, defined as either field of view not encompassing the entire nerve and temporal vascular arcades, or without sufficient image quality to exclude microaneurysm-sized lesions. These were excluded, leaving 281 images; dataset details are in Table 1.

External Validation Dataset

The external validation dataset comprises 210 color fundus photographs of 106 diabetic patients who attended a private medical retina clinic in Tanta, Egypt. Informed consent was obtained from all patients for this research study. A DRI OCT Triton machine version 10.11 (Topcon Corporation, Tokyo, Japan) was used to acquire 55 degree fundus photographs. All data was obtained retrospectively via convenience sampling, and subsequently anonymized. The dataset was labeled using an identical approach to the EyePACS validation subset by three retina specialists.

Model Training

The DL models were trained utilizing AutoML Vision on Google Cloud Platform (GCP). As described in our prior work, this platform provides a graphical user interface (GUI) for

data upload, labeling, and model training without coding [12,13,31]. AutoML entails dataset management, neural architecture search and automated hyperparameter tuning. Images were uploaded to GCP Buckets, and labels were uploaded to GCP via .csv files containing labels, training set splits, and GCP bucket locations. External validations were performed via command line interface batch prediction requests. Patient level splits were maintained for the EyePACS dataset, however no patient-level data was provided for the Messidor dataset, for which we were consequently unable to ensure that patient-level splits were maintained. Each hour of cloud compute represents 8 parallel NVIDIA® Tesla® V100 GPU connected machines. All AutoML model training was specified to use maximum allowable cloud compute hours (800), with the early stopping option enabled. This serves to automatically stop training when no further model improvement is noted. There were no local computer system requirements for the usage of cloud based platforms.

Teacher Model

The Messidor-2 Dataset with gold standard adjudicated labels applied was randomly split to train, tune, and validation sets (80%, 10%, and 10% respectively). AutoML was used as described to train a DL model, which is henceforth referred to as the teacher model.

Student Model

Following teacher model training, the model was deployed on GCP, and batch prediction was performed via the Google Cloud software developer kit command line

interface to run inference on the EyePACS training dataset (n=58,389). This generated model predictions of RDR and NRDR for EyePACS training dataset images. The resulting predictions were assigned as pseudo-labels to the EyePACS training dataset (n=58,389), which was combined with the teacher model training dataset (n=1,395). Subsequently, a student model was trained via AutoML utilizing the combined train set (n=59,784) with the teacher tune set (n=175) (Figure 2).

Bespoke Model

To compare our code-free AutoML model approach with a traditional DL model designed via coding, we developed a bespoke RDR student model utilizing identical images, data splits, and label sets to the AutoML approach. Models were built in Tensorflow 1.15 with Python 3.7 and sklearn library [37,38].

We followed the subsequent standard pipeline for developing a bespoke coded RDR model; First, we compared the performance of two commonly used model architectures, InceptionV3 and ResNet50 in a general hyperparameter configuration, and selected InceptionV3 as the backbone of the bespoke model due to its superior performance. We then searched for optimized performance by grid-searching the hyperparameters, including the learning rate, momentum of the optimiser, and batch size. We implemented data augmentation on the training and tuning sets including random rotation and flipping, and color jitter, to avoid model overfitting, and to increase generalization. An InceptionV3 model pre-trained on Imagenet weights, with a learning rate of 0.1, SGD momentum of 0.0 and batch size of 32 was trained for 20 epochs. All

model layers were set as trainable. We thus obtained a bespoke model as the baseline for comparison.

Statistical Analysis

Performance metrics are reported at Youden's threshold. Batch prediction results displayed softmax outputs for predictions, which were used to generate receiver operating characteristic curves and calculate the area under the receiver operating characteristic curve (AUROC) using sklearn inbuilt functions. Fisher's exact test was performed and p values were calculated for failure case analysis.

Results

Internal validation

The student model demonstrated improved AUROC and overall performance metrics as compared with the teacher model (Figure 3). Student model AUROC, accuracy (ACC), and F1 score on the Messidor validation dataset were 0.951, 93.7%, 86.7% as compared with teacher model of 0.939, 92.0%, 84.1% respectively. Sensitivity and specificity were 84.8%, 96.9% for the student model and 80.4%, 96.1% for the teacher model respectively. Full performance metrics are available in Table 2.

EyePACS validation dataset AUROC was improved for the student as compared to the teacher model. Student model AUROC, ACC, F1 were 0.916, 74.4%, 52.0% as compared with teacher model 0.886, 84.3%, 51.1% respectively. As compared with Messidor internal validation, there were increased false positives in both student and teacher models (Supplementary Table 1) and thus lower PPV and F1 scores. The decreased number of false negatives in the student model (two) as compared with 18 in the teacher model, represents 16 patients who may have otherwise suffered vision-threatening consequences from undertreatment.

External validation

Student model performance on the Egypt external validation dataset demonstrated similarly high performance as compared with the teacher model. AUROC, ACC and F1 were 0.950, 96.7%, 98.3% for the student model and 0.964, 93.3%, 96.4% for the

teacher model respectively. Sensitivity and specificity were 100%, 41.7% for the student model and 94.4%, 75.0% for the teacher model respectively. A markedly lower specificity in this relatively unbalanced dataset containing over 90 percent RDR, suggests that tuning the threshold for this care setting may be necessary for balanced model performance with respect to desired false negative and false positive rates.

Failure Case Analysis

We performed a post-hoc failure analysis, in order to better characterize cases where models generated incorrect (false positive or false negative) predictions. As we had access to 3 independent grades for each image in the EyePACS and Egypt validation datasets, we characterized the validation set to images with or without grader agreement in respect to RDR vs NRDR (Table Z). In the Egypt validation dataset, incorrect model predictions were significantly less likely to have grader agreement (20.0%) as compared to correct (true positive or true negative) predictions (92.7%) ($p=0.0002$). Similarly, in the EyePACS validation dataset, incorrect predictions demonstrated significantly less grader agreement (60.8%) as compared with correct predictions (87.4%) ($p<.0001$). Together, this suggests that disagreed upon images may have inherent ground truthing difficulty, and have a corresponding increased classification difficulty for both humans and models.

Bespoke DL Model

The AUROC of the bespoke model on the EyePACS validation dataset was slightly worse as compared with the AutoML student model. Bespoke AUROC, ACC, F1 was 0.853, 74.4%, 83.2% as compared with 0.916, 74.4%, 52.0% respectively for the

AutoML model. However, performance of the bespoke model on the Messidor-2 validation dataset was improved as compared with the AutoML student model. Bespoke AUROC, ACC, F1 was 0.985, 96.5%, 94.8% as compared with 0.951, 93.7%, 87.6% respectively for the AutoML model. With regards to the Egypt external validation dataset, the AUROC of the bespoke model was worse as compared with the AutoML student model. Bespoke AUROC, ACC, F1 was 0.890, 94.3%, 50.0% as compared with 0.950, 96.7%, 98.3% respectively for the AutoML model.

Discussion

In this work, we demonstrate a self-training AutoML approach with public datasets for DR classification of retinal fundus photographs. This framework simultaneously addresses multiple barriers to the democratization of clinician-driven ML. Our findings suggest that leveraging small datasets with high-quality labels on large public unlabeled datasets improves model performance. The resultant AutoML student model demonstrated improved AUROC on internal validation and a similarly high AUROC on the external validation dataset. Validation performance improved on the EyePACS dataset with images from a variety of imaging hardware, suggesting enhanced generalizability, via incorporating the larger pseudo-labeled dataset [39]. As compared with our bespoke model, the AutoML models demonstrate improved AUROC in both EyePACS and external validation datasets, suggesting AutoML obviates the necessity for manual coding.

In a post-hoc analysis performed of the EyePACS and Egypt validation datasets, we determined that images with grader disagreement were significantly more likely to receive incorrect model predictions. This suggests that edge cases with inherent ground truth uncertainty are more likely to lead to incorrect predictions by the model. Grader variability is a well-known issue with DR grading, a complex process that requires human identification of subtle retinal microvascular abnormalities [9]. While we attempted to gain insights on model outputs through our post-hoc analysis of incorrect predictions, model explainability is inherently more opaque with AutoML. Though this has the potential to improve ML explainability, the knowle;[-[=;. dge of model architectures and

hyperparameters in bespoke models do not inherently provide explainability on a per-image basis. We are working on AutoML explainability including saliency techniques such as integrated gradients and XRAI region based attribution maps [40,41].

Although our student models demonstrated high AUROCs, this average validation set performance may obscure the possibility of inaccurate predictions on important subgroups[42]. Despite good performance achieved in the detection of RDR, the model may underperform in patients with proliferative DR and tractional retinal detachment. Prior studies have shown commercial DR algorithms had low sensitivity to detect these cases in a real-world scenario [43], despite high sensitivity on a publicly available dataset [44]. This is termed hidden stratification, and may be especially meaningful if worse performance occurs in severe disease [45,44]. An analysis of hidden stratification in the context of self-training is beyond the scope of our study given the limited granularity of public datasets. However, we believe this topic represents a promising direction for further investigation.

Our study has several limitations; As patient-level data was not provided for the Messidor-2 dataset, there we were unable to ensure patient-level splits were maintained. To mitigate the potential for falsely increased performance metrics, we performed an additional validation with a portion of the EyePACS dataset regraded and arbitrated by three retina specialists. The Egypt dataset consists of photos from a different camera, field of view, and higher RDR prevalence as compared to our internal validation datasets. On this dataset, the student model had 11 less false negatives as

compared with the teacher model, which in combination with a 100% NPV is amenable to a screening use-case. The data shift was presumably larger for the student model since it was mostly trained on EyePACS images which had the least RDR prevalence. This dataset primarily acts as a “worst case” stress test of data distribution shift and model generalization, and not to compare student and teacher performance. Additionally, the bespoke model may have overfit, as its performance was worse on this dataset.

Self-training has the potential for misclassification of unlabeled data by the teacher model[46]. A possible solution is filtering training data based on teacher model softmax outputs, such that predictions below a threshold would not be used in student model training [10,20,47]. However, exclusively using the most confident predictions may decrease generalizability, which would be detrimental in real-world use, as high variability occurs from image (e.g. quality, device) and patient-related (e.g. ethnicity, concurrent disease) factors [48].

Although AutoML platforms provide free tiers, cloud-based ML model training incurs costs depending on the number of images used. Therefore, utilizing the self-training approach on large unlabeled datasets may become expensive. While many AutoML platforms, including GCP are HIPAA and ISO compliant, institutional research board approvals are necessary before utilizing these platforms on identifiable datasets.

Conclusion

Herein, we leveraged a self-training code-free AutoML approach to address barriers for democratization of clinician-driven ML including the need for coding expertise, the scarcity of data, and the cost of high-quality labeling. Using public DR datasets, we elucidated that self-training is an effective method to increase both model performance and generalizability, while decreasing the need for expensive expert labeling. To effectively address patient-relevant clinical endpoints, medical ML models are best designed by use-case experts such as clinicians. As evidenced by the improved performance of our code-free AutoML models as compared with our manually designed bespoke model, AutoML allows clinicians and researchers without coding expertise to achieve similar results as computer scientists. As the tools for machine learning continue to be democratized, our SSL approach has the potential to address the remaining disparity of expensive clinical labeling, enabling clinicians without coding expertise or access to large well-labeled private datasets to develop their own models.

Acknowledgements

The Messidor-2 dataset was kindly provided by the Messidor program partners (see http://www.adcis.net/en/third-party/messidor/).

This work was supported by a Springboard Grant from the Moorfields Eye Charity (GR000080/R190016A; EK) and UK National Institute for Health Research (NIHR) Clinician Scientist Award (NIHR-CS–2014-12-023; PAK).
This work was supported by grants from the Engineering and Physical Sciences Research Council (EP/S021930/1 & 2410776; RS) and by a UK Research & Innovation Future Leaders Fellowship (MR/T019050/1; PAK).

The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health. The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Author Contributions:

EK designed the study. EK, HK, MM acquired data. EK, MB, JH, HK, RS, MK, SB analyzed data. EK wrote the first draft of the manuscript. PAK, MC, EM, RS, MB, SB contributed to the writing and approval of the manuscript.

Competing Interests

EK is a co-founder of Sanro Health, a consultant for Alimera Sciences and previously was a consultant for Google Health, Reti Health, and Genentech. PAK has received speaker fees from Heidelberg Engineering, Topcon, Carl Zeiss Meditec, Haag-Streit, Allergan, Novartis, and Bayer. He has served on advisory boards for Novartis and Bayer and has been a consultant for DeepMind and Optos. All other authors declare no competing interests.

Table 1: Dataset characteristics

	Teacher model		Student model		Teacher and student models		
Datasets	Development set: Teacher		Development set: Student		Validation set		
	Training set	Tuning set	Training set	Tuning set	Internal validation 1	Internal validation 2	External validation
Image Source	Messidor-2		Messidor-2 + EyePACS		Messidor-2	EyePACS	Egypt dataset
Label Source	Google (Kaggle Public)		Google (Kaggle Public) + Teacher model		Google (Kaggle Public)	Expert	Expert
Labeling Approach	Adjudication		Adjudication + Self-training		Adjudication	Majority vote	Majority vote
Number of images	1,395	175	59,784	175	174	281	210
Camera and FOV	Topcon TRC NW6 45 degree		Various cameras 45 degree		Topcon TRC NW6 45 degree	Different cameras 45 degree	Topcon DRI OCT Triton 55 degree
Clinical environment	Public Ophthalmology department		Public Ophthalmology department + Primary care		Public Ophthalmology department	Primary care	Private Ophthalmology Practice

Table 2: Algorithm Performance Metrics

	AUROC	Accuracy	F1 score	Sensitivity	Specificity	PPV	NPV
Teacher model							
Messidor Internal validation 1	0.939	92.0%	84.1%	80.4%	96.1%	88.1%	93.2%
EyePACS Internal validation 2	0.886	84.3%	51.1%	56.1%	89.2%	46.9%	92.2%
Egypt External validation	0.964	93.3%	96.4%	94.4%	75.0%	98.4%	45.0%
Student model							
Messidor Internal validation 1	0.951	93.7%	87.6%	84.8%	96.9%	90.7%	94.7%
EyePACS Internal validation 2	0.916	74.4%	52.0%	95.1%	70.8%	35.8%	98.8%
Egypt External validation	0.950	96.7%	98.3%	100%	41.7%	96.6%	100%
Bespoke model							
Messidor Internal validation 1	0.985	96.5%	94.8%	96.5%	96.9%	89.1%	96.1%
EyePACS Internal validation 2	0.853	74.4%	83.2%	74.2%	75.6%	94.7%	33.3%
Egypt External validation	0.890	94.3%	50.0%	50.0%	97.0%	50.0%	97.0%

Figure 1: Self Training

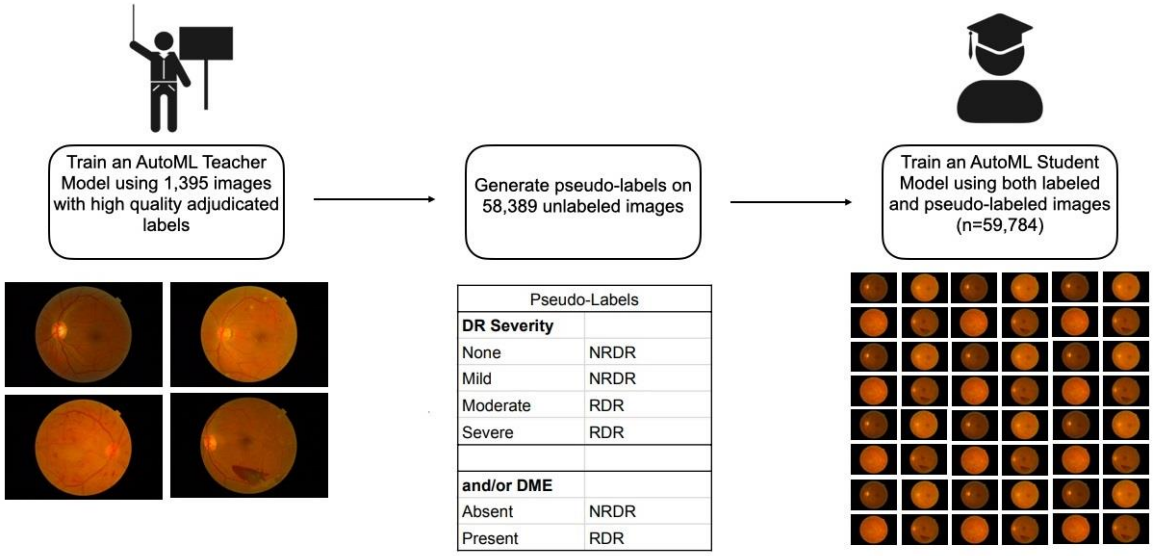


Figure 1 Simplified. Illustration of the self-training approach, which involves three steps: (1) a teacher model is trained on labeled data; (2) the teacher model is used to generate pseudo-labels on unlabeled data; (3) a student model is trained on labeled and pseudo-labeled images

Figure 2: Model Training and Validation Data Flow

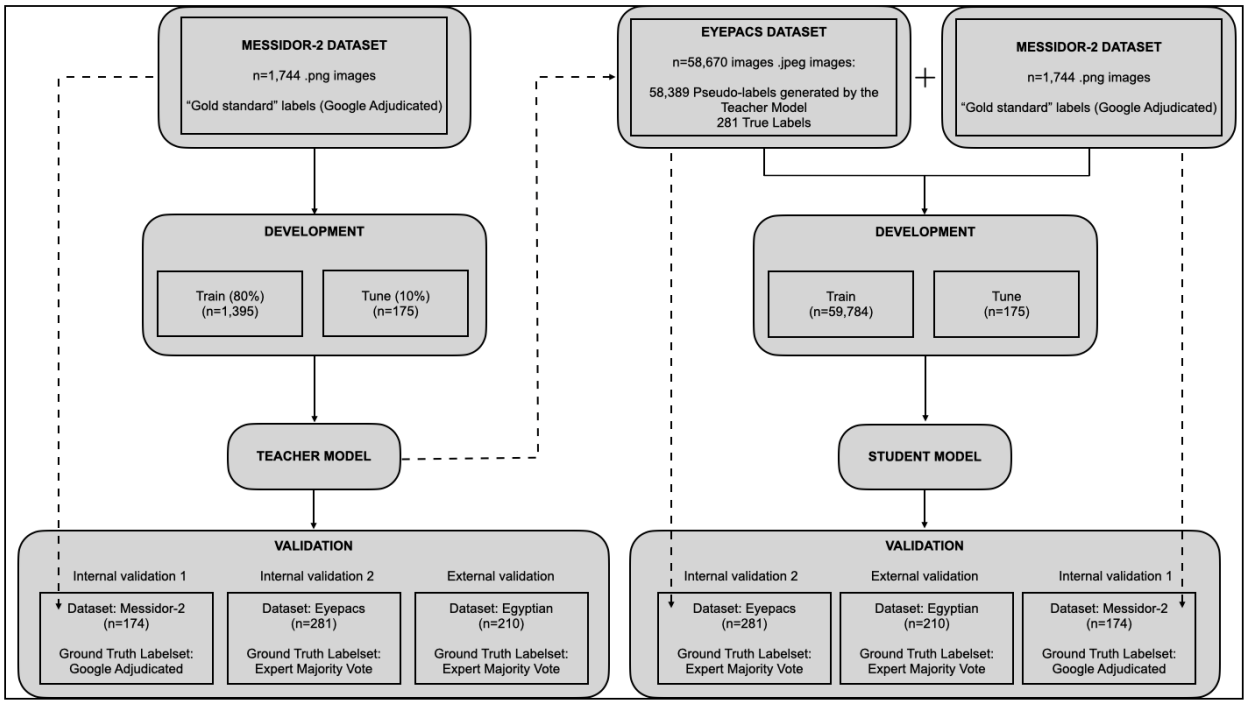
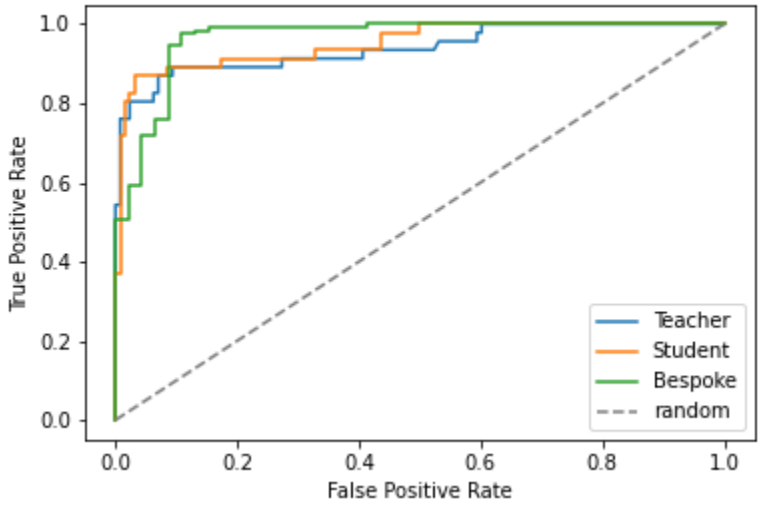


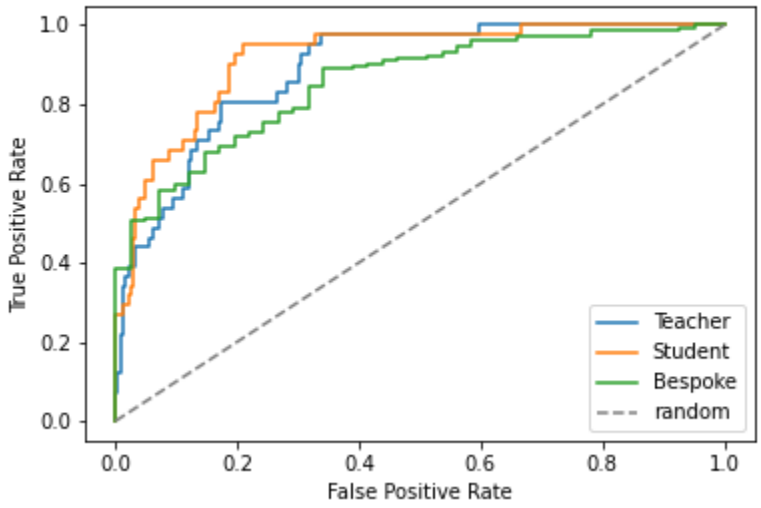
Figure 2: Data splits and flows for self-training and validation of student and teacher models. Once trained, the teacher model is used to apply pseudo-labels to a new dataset (EyePACS). Subsequently, the resulting labels and images are combined with the teacher model training dataset to train a student model. Both models are internally and externally validated on adjudicated or arbitrated validation datasets.

Figure 3: Algorithm ROC Curves

A. Messidor Dataset Internal Validation



B. EyePACS Dataset Internal Validation



C. Egypt Dataset External Validation

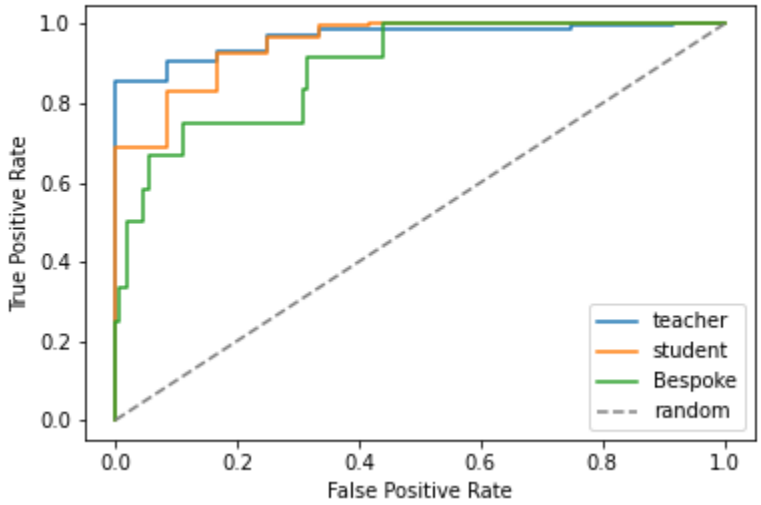


Figure 3: A) Receiver-operating curves for teacher, student, and bespoke manually coded models on the Messidor internal dataset. B) Respective model performance on the EyePACS internal dataset. C) Respective model performance on the Egypt external dataset. Ground truth labels were adjudicated (Messidor) or arbitrated (EyePACS and Egypt).

Supplementary

Table 1: Model Confusion Matrices

Model	Dataset	Confusion Matrix		
		Model Prediction	True Positive	True Negative
Teacher	Messidor (Internal 1)	Positive	37	5
		Negative	9	123
Student		Positive	39	4
		Negative	7	124
Teacher	EyePACS (Internal 2)	Positive	23	26
		Negative	18	214
Student		Positive	39	70
		Negative	2	170
Teacher	Egypt (External)	Positive	187	3
		Negative	11	9
Student		Positive	198	7
		Negative	0	5

References

1. Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol*. 2019;103(2):167-175.

2. Aristidou A, Jena R, Topol EJ. Bridging the chasm between AI and clinical implementation. *Lancet*. 2022;399(10325):620.

3. Gartner Identifies the Top Strategic Technology Trends for 2021. Gartner. Accessed November 27, 2022. https://www.gartner.com/en/newsroom/press-releases/2020-10-19-gartner-identifies-the-top-strategic-technology-trends-for-2021

4. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444.

5. Mahajan D, Girshick R, Ramanathan V, et al. Exploring the Limits of Weakly Supervised Pretraining. *arXiv [csCV]*. Published online May 2, 2018. http://arxiv.org/abs/1805.00932

6. Korot E, Wood E, Weiner A, Sim DA, Trese M. A renaissance of teleophthalmology through artificial intelligence. *Eye* . 2019;33(6):861-863.

7. Bellemo V, Lim ZW, Lim G, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *The Lancet Digital Health*. 2019;1(1):e35-e44.

8. Heydon P, Egan C, Bolter L, et al. Prospective evaluation of an artificial intelligence-enabled algorithm for automated diabetic retinopathy screening of 30 000 patients. *Br J Ophthalmol*. Published online June 30, 2020. doi:10.1136/bjophthalmol-2020-

316594

9. 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(8):1264-1272.

10. Xie Q, Luong MT, Hovy E, Le QV. Self-Training With Noisy Student Improves ImageNet Classification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Published online 2020. doi:10.1109/cvpr42600.2020.01070

11. Nguyen Q, Valizadegan H, Hauskrecht M. Learning classification models with soft-label information. *J Am Med Inform Assoc*. 2014;21(3):501-508.

12. Faes L, Wagner SK, Fu DJ, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. *The Lancet Digital Health*. 2019;1(5):e232-e242.

13. Korot E, Guan Z, Ferraz D, et al. Code-free deep learning for multi-modality medical image classification. *Nature Machine Intelligence*. Published online March 1, 2021:1-11.

14. Yang J, Zhang C, Wang E, Chen Y, Yu W. Utility of a public-available artificial intelligence in diagnosis of polypoidal choroidal vasculopathy. *Graefes Arch Clin Exp Ophthalmol*. 2020;258(1):17-21.

15. Antaki F, Kahwati G, Sebag J, et al. Predictive modeling of proliferative vitreoretinopathy using automated machine learning by ophthalmologists without

coding experience. *Sci Rep*. 2020;10(1):19528.

16. Lee JH, Kim YT, Lee JB, Jeong SN. A Performance Comparison between Automated Deep Learning and Dental Professionals in Classification of Dental Implant Systems from Dental Imaging: A Multi-Center Study. *Diagnostics (Basel)*. 2020;10(11). doi:10.3390/diagnostics10110910

17. 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.

18. Korot E, Pontikos N, Liu X, et al. Predicting sex from retinal fundus photographs using automated deep learning. *Sci Rep*. 2021;11(1):10286.

19. Mey A, Loog M. A soft-labeled self-training approach. *2016 23rd International Conference on Pattern Recognition (ICPR)*. Published online 2016. doi:10.1109/icpr.2016.7900028

20. Triguero I, García S, Herrera F. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowl Inf Syst*. 2015;42(2):245-284.

21. Chen T, Kornblith S, Norouzi M, Hinton G. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv [csLG]*. Published online February 13, 2020. http://arxiv.org/abs/2002.05709

22. Lee R, Wong TY, Sabanayagam C. Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *Eye Vis (Lond)*. 2015;2:17.

23. Zhang X, Saaddine JB, Chou CF, et al. Prevalence of diabetic retinopathy in the United States, 2005-2008. *JAMA*. 2010;304(6):649-656.

24. DeTore J, Rizzolo D. Telemedicine and diabetic retinopathy. *JAAPA*. 2018;31(9):1-5.

25. Sim DA, Mitry D, Alexander P, et al. The Evolution of Teleophthalmology Programs in the United Kingdom: Beyond Diabetic Retinopathy Screening. *J Diabetes Sci Technol*. 2016;10(2):308-317.

26. Nguyen HV, Tan GSW, Tapp RJ, et al. Cost-effectiveness of a National Telemedicine Diabetic Retinopathy Screening Program in Singapore. *Ophthalmology*. 2016;123(12):2571-2580.

27. Scanlon PH. The English National Screening Programme for diabetic retinopathy 2003-2016. *Acta Diabetol*. 2017;54(6):515-525.

28. Huemer J, Wagner SK, Sim DA. The Evolution of Diabetic Retinopathy Screening Programmes: A Chronology of Retinal Photography from 35 mm Slides to Artificial Intelligence. *Clin Ophthalmol*. 2020;14:2021-2035.

29. Ipp E, Liljenquist D, Bode B, et al. Pivotal Evaluation of an Artificial Intelligence System for Autonomous Detection of Referrable and Vision-Threatening Diabetic Retinopathy. *JAMA Netw Open*. 2021;4(11):e2134254.

30. Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med*. 2018;1:39.

31. 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 Digit Health*. 2021;3(1):e51-e66.

32. Decencière E, Zhang X, Cazuguel G, et al. FEEDBACK ON A PUBLICLY DISTRIBUTED IMAGE DATABASE: THE MESSIDOR DATABASE. *Image Anal Stereol*. 2014;33(3):231.

33. Diabetic Retinopathy (resized). Accessed February 14, 2020. https://www.kaggle.com/tanlikesmath/diabetic-retinopathy-resized

34. Tufail A, Rudisill C, Egan C, et al. Automated Diabetic Retinopathy Image Assessment Software: Diagnostic Accuracy and Cost-Effectiveness Compared with Human Graders. *Ophthalmology*. 2017;124(3):343-351.

35. Ting DSW, Cheung CYL, 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.

36. Ipp E, Shah VN, Bode BW, Sadda SR. 599-P: Diabetic Retinopathy (DR) Screening Performance of General Ophthalmologists, Retina Specialists, and Artificial Intelligence (AI): Analysis from a Pivotal Multicenter Prospective Clinical Trial. *Diabetes*. 2019;68(Supplement 1). doi:10.2337/db19-599-P

37. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12(85):2825-2830.

38. Abadi M, Agarwal A, Barham P, et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *arXiv [csDC]*. Published online March 14, 2016. Accessed September 21, 2021. https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45166.pdf

39. Rogers TW, Gonzalez-Bueno J, Garcia Franco R, et al. Evaluation of an AI system for the detection of diabetic retinopathy from images captured with a handheld portable fundus camera: the MAILOR AI study. *Eye* . Published online May 7, 2020. doi:10.1038/s41433-020-0927-8

40. Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. *arXiv [csLG]*. Published online March 4, 2017. http://arxiv.org/abs/1703.01365

41. Kapishnikov A, Bolukbasi T, Viégas F, Terry M. XRAI: Better Attributions Through Regions. *arXiv [csCV]*. Published online June 6, 2019. http://arxiv.org/abs/1906.02825

42. Goel K, Gu A, Li Y, Ré C. Model Patching: Closing the Subgroup Performance Gap with Data Augmentation. *arXiv [csLG]*. Published online August 15, 2020. http://arxiv.org/abs/2008.06775

43. van der Heijden AA, Abramoff MD, Verbraak F, van Hecke MV, Liem A, Nijpels G. Validation of automated screening for referable diabetic retinopathy with the IDx-DR device in the Hoorn Diabetes Care System. *Acta Ophthalmol*. 2018;96(1):63-68.

44. Abràmoff MD, Lou Y, Erginay A, et al. Improved Automated Detection of Diabetic

Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning. *Invest Ophthalmol Vis Sci*. 2016;57(13):5200-5206.

45. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. *Proc ACM Conf Health Inference Learn (2020)*. 2020;2020:151-159.

46. Li M, Zhou ZH. SETRED: Self-training with Editing. *Advances in Knowledge Discovery and Data Mining*. Published online 2005:611-621. doi:10.1007/11430919_71

47. Fazakis N, Kanas VG, Aridas CK, Karlos S, Kotsiantis S. Combination of Active Learning and Semi-Supervised Learning under a Self-Training Scheme. *Entropy* . 2019;21(10). doi:10.3390/e21100988

48. Yip MYT, Lim G, Lim ZW, et al. Technical and imaging factors influencing performance of deep learning systems for diabetic retinopathy. *NPJ Digit Med*. 2020;3:40.