AMERICAN ACADEMY OF OPHTHALMOLOGY®

How Foundational Is the Retina Foundation Model? Estimating RETFound's Label Efficiency on Binary Classification of Normal versus Abnormal OCT Images

David Kuo, MD,[1] *Qitong Gao, PhD,*[2] *Dev Patel, MS,*[2] *Miroslav Pajic, PhD,*[2,3] *Majda Hadziahmetovic, MD*[1,2]

Objective: While the availability of public internet-scale datasets of images and language has catalyzed remarkable progress in machine learning, medical datasets are constrained by regulations protecting patient privacy and the time and cost required for curation and labeling. Self-supervised learning or pretraining has demonstrated great success in learning meaningful representations from large unlabeled datasets to enable efficient learning on downstream tasks. In ophthalmology, the RETFound model, a large vision transformer (ViT-L) model trained by masked autoencoding on 1.6 million color fundus photos and OCT B-scans, is the first model pretrained at such scale for ophthalmology, demonstrating strong performance on downstream tasks from diabetic retinopathy grading to stroke detection. Here, we measure the label efficiency of the RETFound model in learning to identify normal vs. abnormal OCT B-scans obtained as part of a pilot study for primary care-based diabetic retinopathy screening in North Carolina.

Design: The 1150 TopCon Maestro OCT central B-scans (981 normal and 169 abnormal) were randomly split 80/10/10 into training, validation, and test datasets. Model training and hyperparameter tuning were performed on the training set guided by validation set performance. The best performing models were then evaluated on the final test set.

Subjects: Six hundred forty-seven patients with diabetes in the Duke Health System participating in primary care diabetic retinopathy screening contributed 1150 TopCon Maestro OCT central B-scans.

Methods: Three models (ResNet-50, ViT-L, and RETFound) were fine-tuned on the full training dataset of 915 OCT B-scans and on smaller training data subsets of 500, 250, 100, and 50 OCT B-scans, respectively, across 3 random seeds.

Main Outcome Measures: Mean accuracy, area under the receiver operator curve (AUROC), area under the precision recall curve (AUPRC), F1 score, precision, and recall on the final held-out test set were reported for each model.

Results: Across 3 random seeds and all training dataset sizes, RETFound outperformed both ResNet-50 and ViT-L on all evaluation metrics on the final held-out test dataset. Large vision transformer and ResNet-50 performed comparably at the largest training dataset sizes of 915 and 500 OCT B-scans; however, ResNet-50 suffered more pronounced performance degradation at the smallest dataset sizes of 100 and 50 OCT B-scans.

Conclusions: Our findings validate the benefits of RETFound's additional retina-specific pretraining. Further research is needed to establish best practices for fine-tuning RETFound to downstream tasks.

Financial Disclosure(s): Proprietary or commercial disclosure may be found in the Footnotes and Disclosures at the end of this article. *Ophthalmology Science 2025;5:100707 © 2025 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/).*

Diabetic retinopathy is a leading cause of preventable vision loss among adults in the United States as its late vision-threatening complications, such as diabetic macular edema and proliferative diabetic retinopathy, can be effectively managed if detected and treated promptly. Unfortunately, due to several factors, including socioeconomic and geographic barriers to care, less than half of all patients with diabetes in the United States undergo the recommended annual screening for diabetic retinopathy recommended by the American Diabetes Association and the American Academy of Ophthalmology.[1,2]

In recent years, several deep learning systems have been developed and published for diabetic retinopathy screening, demonstrating robust diagnostic performance across multiple independent validation datasets; however, to our knowledge, none has publicly available code or model weights.[3–10]

Thus, in an effort to establish primary care–based retinal telescreening for diabetic retinopathy and other referrable retinal pathologies, we turned our attention to current state-of-the-art open-source models that could be adapted to our task and limited dataset, emphasizing the recent advancements in foundation models.

First coined in 2021, foundation models, "any model that is trained on broad data (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks" have transformed the field of machine learning.[11]

https://doi.org/10.1016/j.xops.2025.100707
ISSN 2666-9145/25

Table 1. Class Distribution across Training, Validation, and Final Held-Out Test Datasets

	Train	Validation	Test	Total
Normal	783	98	100	981
Referable	133	17	19	169
Total	916	115	119	1150

Among several emergent behaviors, foundation models have shown the capability to make accurate predictions on new tasks with only a very small number of labeled examples (few-shot learning) and even learn to perform new tasks without any updates to model weights (in-context learning).[12] These developments are of particular interest in health care, where large datasets can be especially challenging to curate.

Recently, Zhou et al[13] released the Retina Foundation model (RETFound), a large vision transformer (ViT-L) model trained by masked autoencoding (MAE) on 1.6 million color fundus photos and OCT B-scans, the first publicly available model pretrained at such scale for ophthalmology. Evaluated on a number of downstream datasets such as the Indian Diabetic Retinopathy Image Dataset (IDRiD), Asia Pacific Tele-Ophthalmology Society 2019 Blindness Detection Dataset (APTOS-2019), and Messidor-2, RETFound demonstrated strong performance across a range of downstream tasks from diabetic retinopathy grading to glaucoma screening to stroke detection.[13–16]

Furthermore, Zhou et al report impressive label efficiency (i.e., requiring a relatively small amount of training data and labels to achieve a target performance level) across 4 tasks in the publication accompanying RETFound's release: heart failure prediction from color fundus photos, myocardial infarction prediction from color fundus photos, diabetic retinopathy grading using the MESSIDOR-2 dataset, and diabetic retinopathy grading using the IDRiD dataset. For MESSIDOR-2, RETFound was able to achieve an area under the receiver operator curve (AUROC) comparable to the next best performing model, with only 45% of the full MESSIDOR-2 dataset or roughly 786 out of 1748 color fundus photos split essentially evenly between photos with diabetic retinopathy and photos without diabetic retinopathy. Similarly, for IDRiD, RETFound was able to achieve an AUROC comparable to the next best performing mode, with only 50% of the full IDRiD dataset or 258 of 516 color fundus photos, roughly one-third of which have diabetic retinopathy and two-thirds of which do not have diabetic retinopathy. With code and model weights freely available under a Creative Commons license, RETFound presents a promising new tool to spur further advances in machine learning for ophthalmology.[13,14,16]

In this study, we explore the foundational capabilities of the RETFound model. Specifically, given the lack of label efficiency experiments for OCT datasets, we assess the model's label efficiency in classifying normal versus abnormal OCT B-scans within the context of primary care clinic–based diabetic retinopathy screening.

Methods

This study was reviewed and approved by the Duke University's institutional review board. All eligible patients were invited to

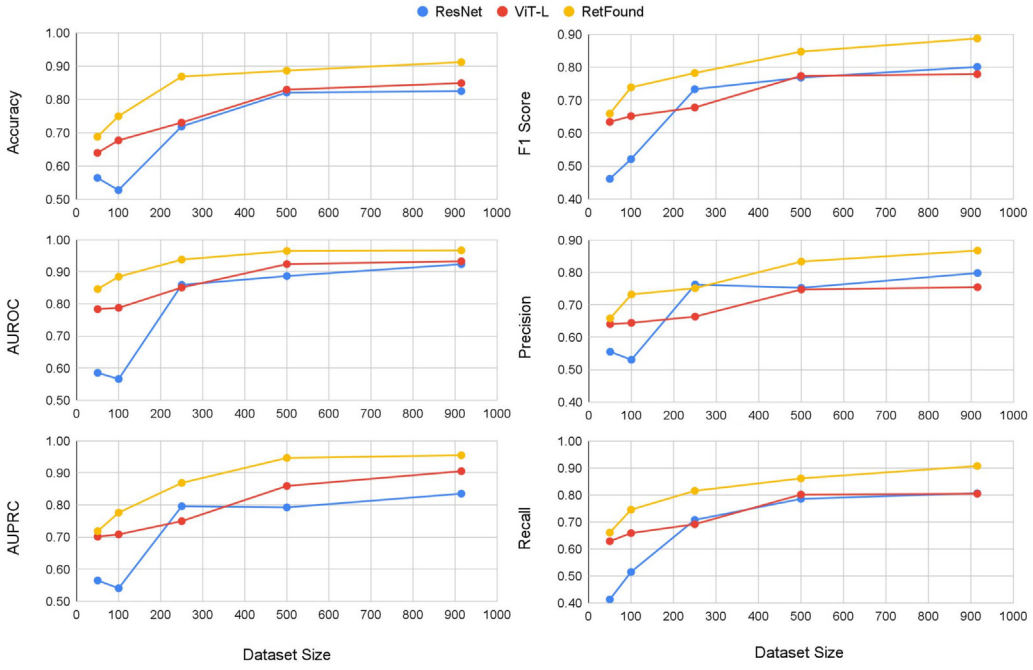

Figure 1. Accuracy, AUROC, AUPRC, F1 score, precision, and recall on the held-out test dataset averaged across 3 random seeds, by model and training dataset size. AUPRC = area under the precision recall curve; AUROC = area under the receiver operator curve; ViT-L = large vision transformer.

Table 2. Test Set Performance (Percent) When Trained on 916 OCT Central B-Scans

Model	Accuracy	AUROC	AUPRC	F1	Precision	Recall
Resnet	82.5 ± 5.1	92.4 ± 3.2	83.5 ± 1.8	80.1 ± 3.4	79.8 ± 5.9	80.7 ± 5.1
ViT-L	84.9 ± 2.3	93.3 ± 1.0	90.5 ± 0.7	77.9 ± 2.6	75.4 ± 3.3	80.5 ± 2.0
RetF	91.2 ± 2.3	96.7 ± 0.6	95.5 ± 0.8	88.7 ± 2.0	86.7 ± 3.2	90.8 ± 2.5

AUPRC = area under the precision recall curve; AUROC = area under the receiver operator curve; ViT-L = large vision transformer.

participate in the study and verbally consented by their primary care provider. Images were taken by trained, certified medical assistants. Patient data were deidentified, and precautions were taken as per Duke University's institutional review board protocol to ensure the security of protected health information and other study data. The protocol followed the tenets of human research as presented in the Declaration of Helsinki.

In total, 1150 TopCon Maestro OCT central B-scans (981 normal and 169 abnormal) were curated from 647 patients with diabetes in the Duke Health System participating in primary care diabetic retinopathy screenings and randomly split 80/10/10 into training, validation, and final held-out test datasets (Table 1).

Two models, ImageNet-pretrained ResNet-50 (timm/resnet50.a1_in1k) and ImageNet-pretrained ViT-L (timm/vit_base_patch16_224.augreg_in21k_ft_in1k), were selected for comparison against the RETFound model.

Despite being first introduced in 2015, deep residual networks or ResNets have stood the test of time.[17] Indeed, Wightman et al[18] demonstrated that with modern optimization and data augmentation techniques, a vanilla ResNet-50 can achieve performance comparable to that of newer model architectures such as EfficientNets and Vision Transformers (ViTs) without extra data or distillation and thus provides a strong baseline comparator (the term "vanilla" has become ubiquitous in computing and technology to describe configurations or implementations that lack customization. In these contexts, it emphasizes simplicity, standardization, and ease of maintenance).

Similarly, the ViT has increasingly replaced convolutional neural networks as a foundational model architecture for computer vision today.[19–23] In fact, RETFound itself is simply a ViT-L model that has undergone additional pretraining via MAE on a large dataset of 1.6 million color fundus photos and OCT B-scans.[24] Thus, ImageNet-pretrained ViT-L allows us to measure the contribution of RETFound's retina-specific pretraining to its performance independent of model architecture.

Machine learning model training and hyperparameter tuning for each of the 3 models (ImageNet-pretrained ResNet-50, ImageNet-pretrained ViT-L, and RETFound) were performed on smaller and smaller subsets of the training set, and the best performing models, as determined by validation set performance, were evaluated on the final held-out test set to measure the label efficiency of the RetFound model in learning to identify normal vs. abnormal OCT B-scans.

More specifically, across 3 random seeds, each model was fine-tuned to convergence (roughly 100 epochs) on training datasets of 915, 500, 250, 100, and 50 OCT B-scans, respectively. In line with the data preprocessing steps for the RETFound model, each OCT B-scan was resized with bicubic interpolation to 224 by 224 pixels and normalized using the ImageNet default mean and standard deviation. Fine-tuning was then performed with RandAugment data augmentation, Layer-wise Adaptive Rate Scaling optimizer, weighted cross-entropy loss (for class imbalance), and a modest grid search over learning rates.[13,25,26] Given the class imbalance in our dataset, model checkpoints with the highest validation set F1 score were selected for evaluation on the final held-out test set, for which mean accuracy, AUROC, area under the precision recall curve (AUPRC), precision, recall, and F1 score were reported.

Results

Performance metrics of the best performing fine-tuned ResNet-50, ViT-L, and RETFound models on the final held-out test set are summarized in Figure 1 and Tables 2 to 6.

In brief, across 3 random seeds and all training dataset sizes, RETFound outperformed both ResNet-50 and ViT-L on all evaluation metrics on the final held-out test dataset, achieving a mean accuracy of 91.2 ± 2.3%, mean AUROC of 96.7 ± 0.6%, mean AUPRC of 95.5 ± 0.8%, mean F1 score of 88.7 ± 2.0%, mean precision of 86.7 ± 3.2%, and mean recall of 90.8 ± 2.5% when trained on the full dataset of 915 OCT B-scans.

Furthermore, RETFound trained on only 250 OCT B-scans (roughly 27% of the full training dataset) achieved comparable results to ResNet-50 and ViT-L trained on the full dataset of 916 OCT B-scans with a mean accuracy of 86.8 ± 2.1%, mean AUROC of 93.8 ± 1.7%, mean AUPRC of 86.9 ± 5.9%, mean F1 score of 78.2 ± 4.0%, mean precision of 75.2 ± 3.6%, and mean recall of 81.6 ± 4.4%.

Table 3. Test Set Performance (Percent) When Trained on 500 OCT Central B-Scans

Model	Accuracy	AUROC	AUPRC	F1	Precision	Recall
Resnet	82.0 ± 5.0	88.7 ± 2.2	79.2 ± 4.0	76.8 ± 2.2	75.3 ± 4.0	78.6 ± 1.1
ViT-L	82.9 ± 3.7	92.4 ± 1.1	85.9 ± 3.1	77.3 ± 0.5	74.7 ± 0.3	80.2 ± 1.4
RetF	88.6 ± 2.6	96.5 ± 1.6	94.7 ± 2.3	84.7 ± 6.8	83.4 ± 9.0	86.2 ± 4.6

AUPRC = area under the precision recall curve; AUROC = area under the receiver operator curve; ViT-L = large vision transformer.

Table 4. Test Set Performance (Percent) When Trained on 250 OCT Central B-Scans

Model	Accuracy	AUROC	AUPRC	F1	Precision	Recall
Resnet	71.9 ± 9.7	85.9 ± 3.9	79.6 ± 5.2	73.3 ± 9.2	76.3 ± 6.7	70.8 ± 11.4
ViT-L	73.0 ± 2.8	85.1 ± 1.6	75.0 ± 2.3	67.8 ± 1.2	66.4 ± 0.9	69.2 ± 1.5
RetF	86.8 ± 2.1	93.8 ± 1.7	86.9 ± 5.9	78.2 ± 4.0	75.2 ± 3.6	81.6 ± 4.4

AUPRC = area under the precision recall curve; AUROC = area under the receiver operator curve; ViT-L = large vision transformer.

ResNet-50 and ViT-L performed comparably at the largest dataset sizes of 915 and 500 OCT B-scans but separated at the smallest dataset sizes of 100 and 50 OCT B-scans, with ResNet-50 suffering more pronounced performance degradation: with 100 OCT B-scans, ResNet-50's performance dropped to a mean accuracy of 52.8 ± 4.6%, mean AUROC of 56.6 ± 1.2%, mean AUPRC of 54.1 ± 0.9%, and mean F1 score of 52.1 ± 6.8, compared to ViT-L's mean accuracy of 67.7 ± 6.5%, mean AUROC of 78.8 ± 5.9%, mean AUPRC of 70.8 ± 4.9%, mean F1 score of 65.2 ± 4.8%, mean precision of 53.1 ± 9.9%, and mean recall of 51.5 ± 3.8%.

Discussion

In this study, we compared the performance and label efficiency of the recent RETFound model developed by Zhou et al to that of standard ImageNet-pretrained ResNet-50 and ViT-L models on a small imbalanced OCT classification dataset. Across 3 random seeds and all training dataset sizes, RETFound significantly outperformed both ResNet-50 and ViT-L on all evaluation metrics and, furthermore, was able to match their results despite training on only 27% of the training dataset. These findings validate the benefits of RETFound's retina-specific pretraining and suggest that RETFound should be a strong default model for OCT classification tasks. Additionally, our experiments suggest that 250 OCT B-scans may be a reasonable initial dataset size to target for OCT-based binary classification tasks as evaluation metrics dropped off sharply for all 3 models below that number.

Of the evaluation metrics measured, AUROC was the most robust, and F1 score was the least robust to decreasing dataset size for all 3 models. AUROC provides a summary metric of model performance across all possible classification thresholds. Therefore, a relatively preserved AUROC suggests that there may be additional performance (e.g., better accuracy, precision, recall) gained at lower dataset sizes by optimizing each model's classification threshold, for instance, guided by the Youden index. F1 score is the harmonic mean of precision (or positive predictive value) and recall (or sensitivity) and measures the proportion of true positives to false positives and false negatives. This is particularly relevant in applications where the positive class is rare relative to the negative class, such as screening for disease in a predominantly healthy population. For this study, poor F1 score relative to accuracy and other metrics suggests that the models struggled to handle class imbalance as dataset size decreased. This could be improved by optimizing classification thresholds as discussed previously, and other methods for addressing class imbalance, such as resampling, may also be helpful.

Although our study is limited by its narrow scope (e.g., binary classification on a small OCT dataset from a single institution), we hope that our results from such a simple task can provide a useful approximate lower bound of the amount of labeled data required to achieve reasonable performance with the RETFound model on OCT classification tasks. Further experiments are certainly warranted to evaluate the robustness of RETFound's performance with multiclass classification and segmentation tasks as well as under varying degrees of class imbalance and distribution shift.

Another potential future direction could be to further explore pretraining convolutional neural networks and smaller, more computationally efficient models in general with the RETFound dataset. Zhou et al did explore pretraining a ResNet-50 model with SimCLR and SwAV but found that these (as well as ViT-L pretrained with MOCO-v3 and DINO) did not perform as well as ViT-L pretrained with MAE. However, the authors do note that "asserting the superiority of the MAE requires caution, given the presence of several variables across all models."[13] Indeed, literature certainly suggests that convolutional neural networks can be quite competitive. As mentioned previously, Wightman et al[18] demonstrated that leveraging modern optimization,

Table 5. Test Set Performance (Percent) When Trained on 100 OCT Central B-Scans

Model	Accuracy	AUROC	AUPRC	F1	Precision	Recall
Resnet	52.8 ± 4.6	56.6 ± 1.2	54.1 ± 0.9	52.1 ± 6.8	53.1 ± 9.9	51.5 ± 3.8
ViT-L	67.7 ± 6.5	78.8 ± 5.9	70.8 ± 4.9	65.2 ± 4.8	64.5 ± 3.8	65.9 ± 5.8
RetF	74.9 ± 3.9	88.5 ± 3.1	77.6 ± 5.1	73.9 ± 3.5	73.2 ± 3.7	74.6 ± 3.7

AUPRC = area under the precision recall curve; AUROC = area under the receiver operator curve; ViT-L = large vision transformer.

Table 6. Test Set Performance (Percent) When Trained on 50 OCT Central B-Scans

Model	Accuracy	AUROC	AUPRC	F1	Precision	Recall
Resnet	56.5 ± 5.3	58.5 ± 1.6	56.5 ± 1.5	46.1 ± 8.5	55.5 ± 4.0	41.3 ± 13.3
ViT-L	64.0 ± 4.3	78.4 ± 2.8	70.1 ± 3.2	63.4 ± 3.8	64.0 ± 3.4	62.9 ± 5.0
RetF	68.8 ± 8.4	84.6 ± 6.9	71.8 ± 5.9	65.9 ± 7.4	65.9 ± 7.5	66.1 ± 8.1

AUPRC = area under the precision recall curve; AUROC = area under the receiver operator curve; ViT-L = large vision transformer.

regularization, and data augmentation techniques such as LAMB, label smoothing, stochastic depth, RandAugment, and MixUp can significantly boost the performance of a vanilla ResNet-50 model. Similarly, Smith et al[27] showed that convolutional neural networks, specifically NFNets, can match ViTs at scale when provided comparable compute budgets and pretraining datasets. Finally, Goldblum et al[28] found that supervised ConvNeXt-Base and ConvNeXt-Tiny models excelled across a diverse set of computer vision tasks and datasets, noting that "despite the recent attention paid to transformer-based architectures and self-supervised learning, high-performance convolutional networks pretrained via supervised learning outperform transformers on the majority of tasks we consider."

Footnotes and Disclosures

Originally received: August 25, 2024.
Final revision: December 25, 2024.
Accepted: January 7, 2025.
Available online: January 11, 2025. Manuscript no. XOPS-D-24-00323.

[1] Department of Ophthalmology, Duke University, Durham, North Carolina.

[2] Department of Electrical and Computer Engineering, Duke University, Durham, North Carolina.

[3] Department of Computer Science, Duke University, Durham, North Carolina.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have made the following disclosure(s):

M.H.: Consultant − AbbVie/Allergan, Bausch+Lomb; Participation on a Data Safety Monitoring Board or Advisory Board − Ocular Terapeutix, ADVERUM, Sparing Vision, Astellas.

The research reported in this publication was supported by the National Institutes of Health's grant: R21EY033480, awarded to M.P. and M.H.

HUMAN SUBJECTS: Human subjects were included in this study. This study was reviewed and approved by the Duke University's institutional review board. All eligible patients were invited to participate in the study and verbally consented by their primary care provider. Patient data were deidentified, and precautions were taken as per Duke University's institutional review board protocol to ensure the security of protected health information and other study data. The protocol followed the tenets of human research as presented in the Declaration of Helsinki.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Kuo, Gao, Pajic, Hadziahmetovic

Data collection: Kuo, Gao, Patel

Analysis and interpretation: Kuo, Gao, Patel

Obtained funding: Pajic, Hadziahmetovic

Overall responsibility: Kuo, Gao, Pajic, Hadziahmetovic

Abbreviations and Acronyms:
AUPRC = area under the precision recall curve; **AUROC** = area under the receiver operator curve; **IDRiD** = Indian Diabetic Retinopathy Image Dataset; **MAE** = masked autoencoding; **ViT** = vision transformer; **ViT-L** = large vision transformer.

Keywords:
Diabetic retinopathy screening, Foundation model, Machine learning, OCT.

Correspondence:
Majda Hadziahmetovic, MD, Department of Ophthalmology, Duke University, 2351 Erwin Rd, Durham, NC 27713. E-mail: majda.hadziahmetovic@duke.edu.

References

1. Lee PP, Feldman ZW, Ostermann J, et al. Longitudinal rates of annual eye examinations of persons with diabetes and chronic eye diseases. *Ophthalmology*. 2003;110:1952−1959.
2. Shi Q, Zhao Y, Fonseca V, et al. Racial disparity of eye examinations among the U.S. Working-age population with diabetes: 2002−2009. *Diabetes Care*. 2014;37:1321−1328.
3. Jani PD, Forbes L, Choudhury A, et al. Evaluation of diabetic retinal screening and factors for ophthalmology referral in a telemedicine network. *JAMA Ophthalmol*. 2017;135:706.
4. 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;316:2402.
5. 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;318:2211.
6. 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 Digital Med*. 2018;1:39.
7. Gulshan V, Rajan RP, Widner K, et al. Performance of a deep-learning algorithm vs manual grading for detecting diabetic retinopathy in India. *JAMA Ophthalmol*. 2019;137:987.
8. Ruamviboonsuk P, Krause J, Chotcomwongse P, et al. Deep learning versus human graders for classifying diabetic

retinopathy severity in a nationwide screening program. *NPJ Digit Med*. 2019;2:25.
9. 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*. 2020;105:723−728.
10. Ruamviboonsuk P, Tiwari R, Sayres R, et al. Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: a prospective interventional cohort study. *Lancet Digit Health*. 2022;4:e235−e244.
11. Bommasani R, Hudson DA, Adeli E, et al. On the opportunities and risks of foundation models. ArXiv. https://crfm.stanford.edu/assets/report.pdf. Accessed August 15, 2024.
12. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, et al., eds. *Advances in Neural Information Processing Systems*. 33. Curran Associates, Inc.; 2020:1877−1901.
13. Zhou Y, Chia MA, Wagner SK, et al. A foundation model for generalizable disease detection from retinal images. *Nature*. 2023;622:156−163.
14. Porwal P, Pachade S, Kokare M, et al. IDRiD: diabetic retinopathy − segmentation and grading challenge. *Med Image Anal*. 2020;59:101561.
15. Karthik SD Maggie. APTOS 2019 blindness detection. https://kaggle.com/competitions/aptos2019-blindness-detection; 2019. Accessed August 15, 2024.
16. Decencière E, Zhang X, Cazuguel G, et al. Feedback on a publicly distributed database: the Messidor database. *Image Anal Stereol*. 2014;33:231.
17. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas: IEEE; 2016:770−778.
18. Wightman R, Touvron H, Jégou H. Resnet strikes back: an improved training procedure in timm. *arXiv*. 2021. https://doi.org/10.48550/arXiv.2110.00476.
19. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv*. 2020. https://doi.org/10.48550/arXiv.2010.11929.
20. Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021:10012−10022.
21. Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021:9650−9660.
22. Bao H, Dong L, Piao S, Wei F. Beit: bert pre-training of image transformers. *arXiv*. 2021. https://doi.org/10.48550/arXiv.2106.08254.
23. Touvron H, Cord M, Jégou H. Deit iii: revenge of the vit. In: *European Conference on Computer Vision*. Tel Aviv: Springer; 2022:516−533.
24. He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans: IEEE; 2022:16000−16009.
25. Cubuk ED, Zoph B, Shlens J, Le QV. Randaugment: practical automated data augmentation with a reduced search space. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020:702−703.
26. You Y, Gitman I, Ginsburg B. Large batch training of convolutional networks. *arXiv*. 2017. https://doi.org/10.48550/arXiv.1708.03888.
27. Smith S, Brock A, Berrada L, De S. ConvNets match vision transformers at scale. *arXiv*. 2023. https://doi.org/10.48550/arXiv.2310.16764.
28. Goldblum M, Souri H, Ni R, et al. Battle of the backbones: a large-scale comparison of pretrained models across computer vision tasks. *arXiv*. 2023. https://doi.org/10.48550/arXiv.2310.19909.