

<https://doi.org/10.1038/s41746-024-01109-5>

RETFound-enhanced community-based fundus disease screening: real-world evidence and decision curve analysis

Check for updates

Juzhao Zhang^{1,2,3,4,6}, Senlin Lin^{1,2,3,6}, Tianhao Cheng^{5,6}, Yi Xu^{1,2,3}, Lina Lu^{1,2,3}, Jiangnan He¹, Tao Yu¹, Yajun Peng¹, Yuejie Zhang⁵✉, Haidong Zou^{1,2,3,4}✉ & Yingyan Ma^{1,2,3,4}✉

Visual impairments and blindness are major public health concerns globally. Effective eye disease screening aided by artificial intelligence (AI) is a promising countermeasure, although it is challenged by practical constraints such as poor image quality in community screening. The recently developed ophthalmic foundation model RETFound has shown higher accuracy in retinal image recognition tasks. This study developed an RETFound-enhanced deep learning (DL) model for multiple-eye disease screening using real-world images from community screenings. Our results revealed that our DL model improved the sensitivity and specificity by over 15% compared with commercial models. Our model also shows better generalisation ability than AI models developed using traditional processes. Additionally, decision curve analysis underscores the higher net benefit of employing our model in both urban and rural settings in China. These findings indicate that the RETFound-enhanced DL model can achieve a higher net benefit in community-based screening, advocating its adoption in low- and middle-income countries to address global eye health challenges.

Eye health is a critical aspect of public health. The Global Burden of Disease Study revealed that in 2020, ~43 million individuals globally were blind, with approximately 295 million and 258 million individuals suffering from moderate to severe and mild visual impairments respectively¹. Visual impairment not only significantly undermines visual function, thereby impacting quality of life², but also lowers productivity and considerably increases all-cause mortality rates³. Previous studies have confirmed that effective eye disease screening can reduce the prevalence of blindness⁴. The development of diagnostic products for eye diseases assisted by artificial intelligence (AI) has matured in recent years⁵, with a growing number of such products receiving medical device approval and entering the market. Research indicates that the adoption of AI for eye disease screening can significantly alleviate the dependency on ophthalmologists and improve the cost-efficiency of screenings^{6–8}. However, during community eye disease screening, the inability to perform pupil dilation, coupled with inferior image capture conditions, personnel, and environment compared with clinical settings, results in compromised image quality. The accuracy of conventional AI products developed for community eye disease screening

significantly diminishes in comparison with the developmental phase^{9,10}, failing to meet the screening needs of the community¹¹. Similar concerns have been broadly observed in other medical domains, such as tumour imaging¹². Therefore, improving the accuracy of AI in practical screening tasks is a pressing concern¹³.

On 13 September 2023, Zhou et al. published a research report wherein they employed Self-Supervised Learning (SSL) techniques on 1.6 million unlabelled fundus images to train a Vision Transformer (ViT), establishing a novel foundational model for retinal image recognition, RETFound¹⁴. This represents a fundamental shift from the preceding technological pathways. Traditional AI model development requires extensive labelled data¹⁵ and these models are often designed for specific tasks, exhibiting limited generalisation for various clinical applications. Conversely, RETFound was engineered to learn generalisable representations from unlabelled retinal images, thereby providing a foundation for rapid adaptation in multiple application realms. According to Zhou et al., RETFound can be rapidly deployed to downstream tasks through transfer learning, such as automatic diagnosis of diabetic retinopathy and glaucoma, prognostic forecasting of

¹Shanghai Eye Disease Prevention & Treatment Center/ Shanghai Eye Hospital, School of Medicine, Tongji University, Shanghai, China. ²National Clinical Research Center for Eye Disease, Shanghai, China. ³Shanghai Engineering Center of Precise Diagnosis and Treatment of Eye Diseases, Shanghai, China. ⁴Department of Ophthalmology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China. ⁵School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China. ⁶These authors contributed equally: Juzhao Zhang, Senlin Lin, Tianhao Cheng.

✉e-mail: yjzhang@fudan.edu.cn; zouhaidong@sjtu.edu.cn; mYy_29@163.com

age-related macular degeneration, and systemic disease prediction. In various eye disease diagnostic tasks performed for internal validation on public datasets, its accuracy surpassed traditional models trained via supervised learning on ImageNet (SL-ImageNet) and proved superior to models entirely based on self-supervised learning on either ImageNet (SSL-ImageNet) or retinal images (SSL-Retinal)¹⁴. To what extent can the accuracy of community eye disease screening be enhanced by employing deep-learning (DL) models based on RETFound? What value can the RETFound framework potentially bring to community eye disease screening? The key questions concerning the extended application and promotion of the RETFound framework are yet to be answered.

For low-income and middle-income countries, the scarcity of health resources necessitates empirically validated evidence before the dissemination and application of new technologies to help formulate the most appropriate health policies. In this study, we developed a RETFound-enhanced DL model for multiple eye diseases using an image dataset from real-world community screening in Shanghai, China. The screening accuracy of the model was then compared with commercial models and traditional AI models to ascertain the degree to which the RETFound framework influences the screening accuracy. Furthermore, based on the prevalence of eye disease in urban and rural areas of China, a decision curve analysis (DCA) was applied to estimate and compare the net benefits of the different models.

Results

Performance on community-based eye disease screening

The test set incorporated in this study consisted of 1890 images, including 287 AMD, 337 DR, 151 PM, and 1115 normal fundus images. The proportion of images with blind eye disease in the test set exceeded the population prevalence. The comparative results are as follows (Table 1).

(1) There was a significant difference in the sensitivity to AMD among the RETFound model, Model S, and Model Y ($X^2 = 21.09, p < 0.001$). Furthermore, post-hoc pairwise comparisons indicated that the sensitivity of the RETFound model was significantly higher than that of Model S ($X^2 = 14.28, p < 0.001$) and Model Y ($X^2 = 18.33, p < 0.001$). No significant differences were observed between the two commercial models ($X^2 = 0.61, p = 0.61$).

(2) Regarding the sensitivity of DR, there was a significant difference among the RETFound model, Model S, and Model Y ($X^2 = 68.18, p < 0.001$). Furthermore, post-hoc pairwise comparisons suggested that the sensitivity of the RETFound model was significantly superior to that of Models S ($X^2 = 39.92, p < 0.001$) and Y ($X^2 = 69.76, p < 0.001$). No significant difference is observed between the two commercial models ($X^2 = 5.36, p = 0.02$, exceeding α').

(3) For PM sensitivity, there was a significant difference among the RETFound model, Model S, and Model Y (Fisher's $p < 0.001$). Moreover, post-hoc pairwise comparisons indicated that the sensitivity of the RETFound model was significantly higher than that of Models S (Fisher's $p < 0.001$) and Y (Fisher's $p < 0.001$). No significant difference is noted between the two commercial models ($X^2 = 5.32, p = 0.02$, exceeding α').

(4) For specificity, there is a significant difference among the RETFound model, Model S, and Model Y ($X^2 = 143.97, p < 0.001$). Additionally, post-hoc pairwise comparisons revealed that the sensitivity of the

RETFound model significantly outperformed that of Models S ($X^2 = 96.11, p < 0.001$) and Y ($X^2 = 140.48, p < 0.001$). No significant difference is identified between the two commercial models ($X^2 = 4.93, p = 0.03$, exceeding α').

Net benefit in community-based eye disease screening

The DCA results indicated that in rural areas of China (Fig. 1) and urban areas of China (Fig. 2), whether for single-disease screening or multi-fundus disease screening, employing the RETFound-enhanced model can achieve higher net benefits than current commercial models. Specifically, in rural areas, the model exhibits maximal net benefit at probability thresholds ranging from 2% to 40%. In urban areas, optimal net benefit is achieved with thresholds between 4% and 71%. Although the two commercial models have advantages in terms of accuracy for various eye diseases, their net benefits for screening are relatively similar. Moreover, a comparison between Figs. 1 and 2 reveal that the net benefit of screening in urban areas surpasses that of screening in rural areas. This is attributed to the higher prevalence of diseases in urban areas, and the implementation of eye disease screening is anticipated to identify more patients.

Comparison of the RETFound-enhanced model and CNNs

As shown in Fig. 3, the convergence speeds of the models were similar, and the loss on the training set stabilised after 15 training cycles. Compared to the RETFound-enhanced model, the two CNN models exhibited a lower final loss in the training set.

The validation results for the three models are presented in Table 2 and Fig. 4. Both the RETFound-enhanced and traditional CNN models showed similar ability in identifying DR (from AUC 0.9564 to AUC 0.9759) using the fundus photographs from the internal validation set. The sensitivity of all models for recognising DR was higher than 90%, and the sensitivity of the RETFound-enhanced model was as high as 96.22%. Their performance was also good in recognising a normal fundus, and the specificity of all models was above 80%.

However, in the external validation set, the CNN models exhibited a substantial drop in performance when compared with a new dataset with significantly different data distributions. By contrast, the RETFound-enhanced model was able to identify disease-related patterns (Supplementary Fig. 1) and demonstrated efficient performance when dealing with data that were different from the training set source. The AUROC and Youden's indices for all three models have decreased, suggesting that during the generalisation process, the accuracy of all three models has been compromised. However, RETFound exhibits the smallest decrease, while the two CNN models show a similar level of decline. This indicates that the RETFound-enhanced model may have stronger generalisation and generic recognition capabilities than traditional healthcare AI approaches that train CNNs based on SL methods.

Discussion

This is one of the first studies to evaluate the real-world value of the RETFound framework for eye disease screening. Using a foundation model pretrained via SSL to develop deep learning models for real-world community-based eye disease screening, our method demonstrated the highest sensitivity for the three main blinding retinal diseases. The results suggest

Table 1 | Performance on community-based eye disease screening

Model Name	Evaluation metrics	AMD (n = 287)	DR (n = 337)	PM (n = 151)	Normal fundus (n = 1115)
RETFound-enhanced Model	Sensitivity	75.96 (218)	95.25 (321)	100 (151)	/
	Specificity	/	/	/	92.47 (1031)
Model S	Sensitivity	61.32 (176)	78.93 (266)	59.60 (90)	/
	Specificity	/	/	/	77.67 (866)
Model Y	Sensitivity	59.23 (170)	71.22 (240)	72.19 (109)	/
	Specificity	/	/	/	73.63 (821)

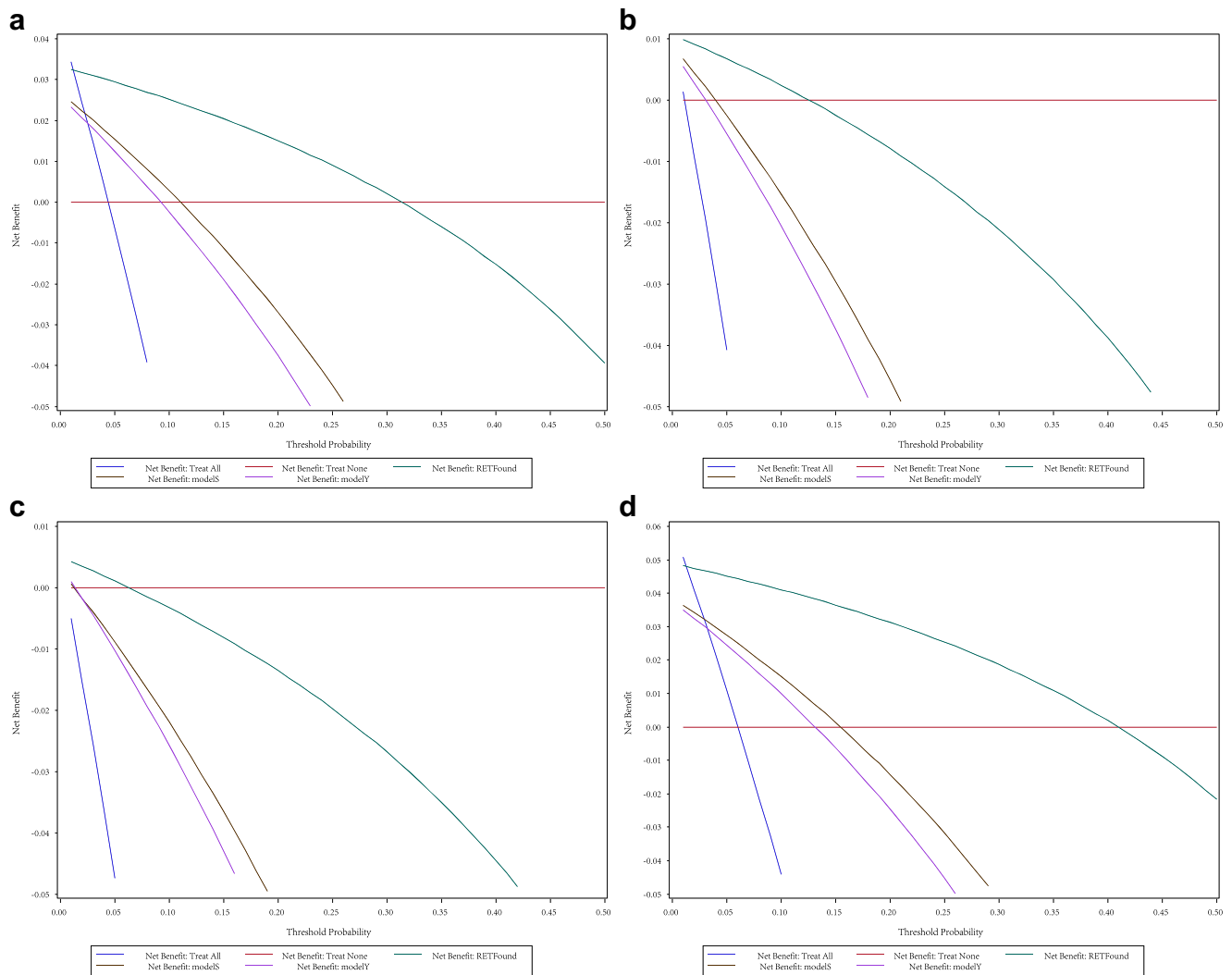


Fig. 1 | Comparison of using RETFound-enhanced model and commercial models in rural areas of China. a AMD single-disease screening. **b** DR single-disease screening. **c** PM single-disease screening. **d** multi-disease combined screening.

that this SSL paradigm can be time-efficient and achieve higher net benefits than current commercial models and conventional SL-based CNNs.

China is one of the countries with the most severe vision impairment and blindness globally. In 2019, China reported 50.59 million individuals with moderate or severe vision impairment, and 8.69 million experiencing blindness¹⁶. The commercial models used as the control groups in this study have broad applications in China. However, a significant reduction in diagnostic accuracy compared to the R&D stage was observed upon implementation in real-world community eye disease screenings^{9,10,17,18}. The decline in the accuracy of commercial models is primarily attributed to the image sets used during the training stage, which are typically derived from high-quality images obtained from clinical institutions^{19–23}. However, community eye disease screening often fails to meet the requirements of clinical settings. Consequently, the images obtained during community eye disease screening are of inferior quality to those obtained from clinical institutions, resulting in heterogeneity between the working image sets and training sets of commercial models. This heterogeneity particularly affects the diagnostic accuracy for mild conditions. Mild conditions typically exhibit small and concealed lesions, necessitating high-quality images to clearly delineate the differences between lesions and surrounding areas; however, community imaging lacks such conditions, often resulting in under- or over-exposure and blurred details (Fig. 5), which in turn diminishes the detection capability of previous models for mild lesions during community screening (Supplementary Table 2). Moreover, the

prevalence of low-quality images is primarily attributed to non-mydratric conditions in community screenings. Advancements in non-mydratric camera technology have the potential to significantly expand the scope of community screenings, from which our model stands to benefit considerably^{24,25}.

Unlike CNNs, foundation models, such as RETFound, are pretrained on a large amount of unlabelled data and capture a wide range of data patterns through SSL. The foundation model is called “foundation” because it is trained on large amounts of data and allows us to finetune it for downstream tasks with minimal additional data. Prior to RETFound, there was no foundational model in the field of ophthalmology. Specifically, when the aim is to adjust and optimise models for a population in a certain community or a certain type of blinding eye disease, traditional medical AI methods utilise CNN models and pretrain them using ImageNet in a supervised manner. This often requires large amounts of training data and repetitive refinement to achieve satisfactory results for downstream tasks. In addition, CNN models may overfit the training data (i.e., extract features that are not relevant to the lesion), resulting in poor performance when adapted to new environments with different data characteristics. In this study, with only several thousand images, minimal GPU resources, and a shorter model training time, the RETFound framework swiftly adapted to our downstream tasks and exhibited good diagnostic accuracy (Supplementary Fig. 2). Additionally, it demonstrated enhanced generalisation capabilities when contrasted with models like EfficientNet and ResNet.

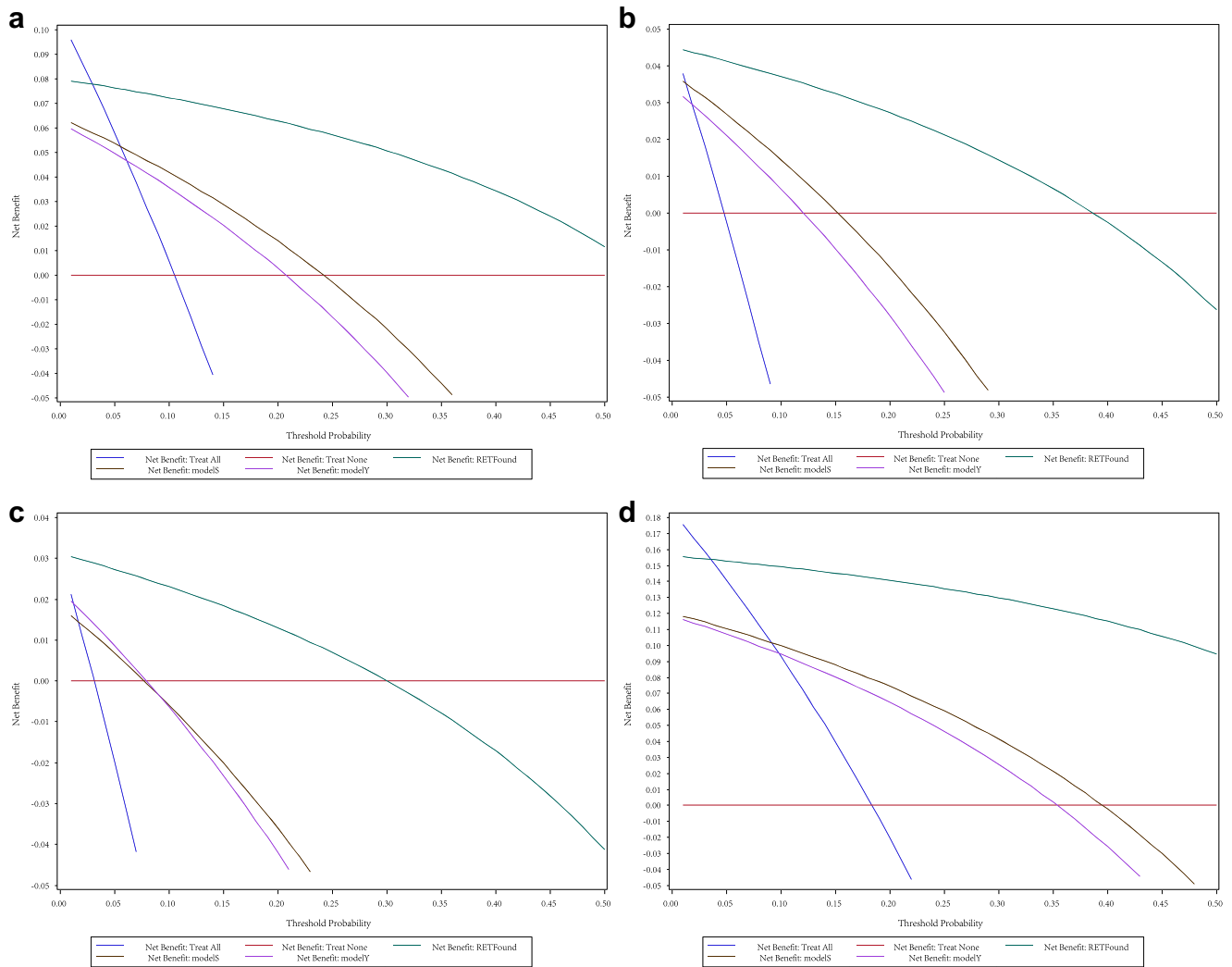


Fig. 2 | Comparison of using RETFound-enhanced model and commercial models in urban areas of China. a AMD single-disease screening. **b** DR single-disease screening. **c** PM single-disease screening. **d** multi-disease combined screening.

Considering the diverse ethnic origins of the public databases used for model training and the SDEDS database employed for validation, this further substantiates the model’s broad applicability and generalisability. We believe that the paradigm of fine-tuning downstream tasks based on pretrained foundation models represents a key shift in medical AI research, providing an efficient alternative to the laborious, costly, and time-consuming data annotation process inherent in the traditional process of training CNNs based on SL methods. This paves the way for population- or region-specific AI model calibration, thereby creating personalised eye disease screening models, which is a unique advantage in ophthalmic applications.

The DCA results of this study indicated different net benefits between urban and rural areas in China when conducting community eye disease screening. This is partly attributable to the higher prevalence of diseases in urban areas, where more patients can be identified through screening. However, the deep learning model developed using the RETFound base model consistently exhibited higher net benefits than commercial models, whether in single-disease screening for AMD, PM, DR, or multi-disease combined screening. Therefore, screening accuracy is pivotal in this aspect. Previous studies by our team suggest that owing to the lower human resource costs in developing countries, the cost-effectiveness of utilising AI for community eye disease screening may not necessarily be better than employing remote screening technologies²⁶. This study validated that the free, open-source RETFound model yields higher accuracy and net benefits

in community eye disease screening than current commercial models. This holds significant value in promoting the utilisation of AI models in community eye disease screening in developing countries, including China. Further research can conduct health economic evaluations of the application of the RETFound framework model in the real world, which we believe will help to further supplement and validate the conclusions of this study.

This study has several notable advantages. First, we pioneered the demonstration that foundational models, such as RETFound, can accelerate the progress of AI in medicine by creating personalised models. In addition to reducing the likelihood of overfitting (which is common in CNNs) to improve generalisability, it also allows healthcare practitioners using AI to focus more on the functionality and application scenarios of AI models rather than on data collection or algorithm development. Second, it emphasises the net benefit as the primary outcome, surpassing the sole focus on screening diagnosis accuracy. This approach provides direct evidence of the practical application of the RETFound foundation model for eye disease screening. Third, all images used in this study were sourced from a long-term SDEDS project conducted in Shanghai, rather than clinical scenarios. This makes it a more suitable environment for evaluating the real-world accuracy of AI models.

This study has some limitations. First, it encompasses only two commonly encountered commercial models in China because of the unavailability of internationally recognised models, such as IDx-DR, OphAI, RetinAI Medical, Retmarker, and Eyenuk. However, it is

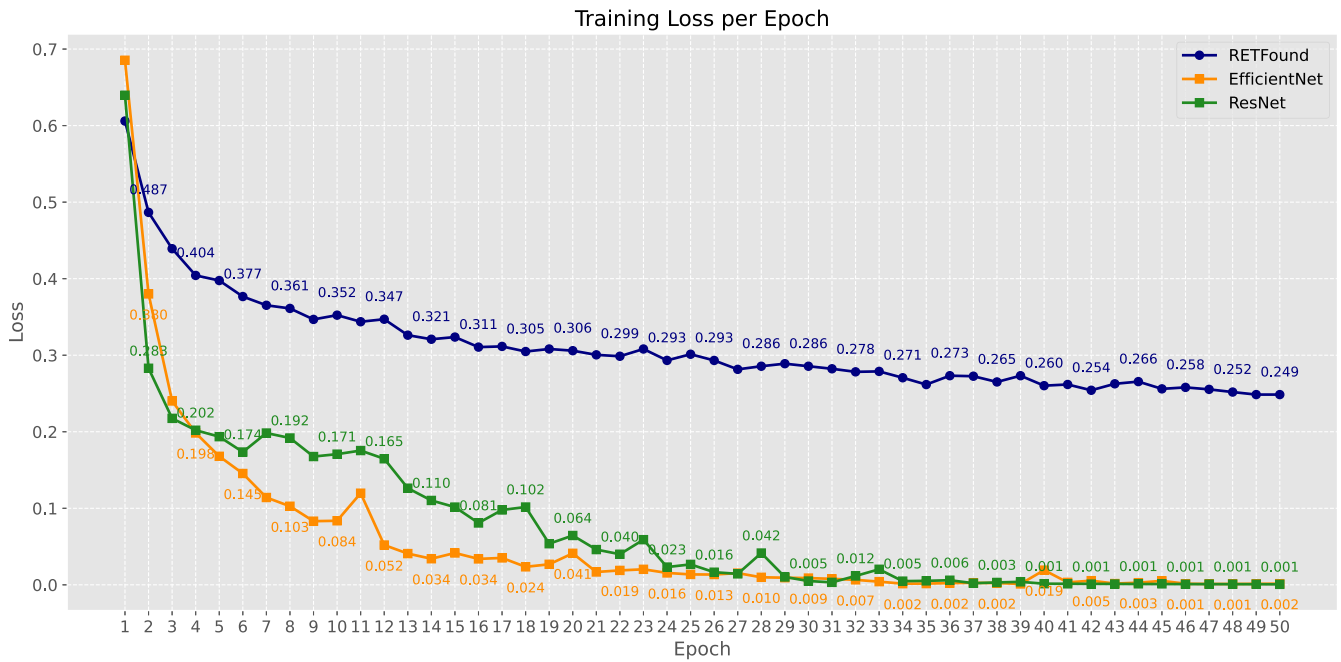


Fig. 3 | Comparison of model training loss over epochs. Each line represents the training loss for three distinct deep learning models across 50 epochs. The blue line corresponds to the RETFound-enhanced model, the orange line to the EfficientNet model, and the green line to the ResNet model.

Table 2 | Performance on DR classification

Model Name	Evaluation metrics	Internal validation		Youden's index	External validation		Youden's index
		DR (n = 502)	Normal fundus (n = 517)		DR (n = 337)	Normal fundus (n = 1115)	
RETFound	Sensitivity	96.22 (483)	/	0.7939	82.79 (279)	/	0.6127
	Specificity	/	83.17 (430)		/	78.48 (875)	
EfficientNetB3	Sensitivity	90.64 (455)	/	0.7826	67.95 (229)	/	0.5055
	Specificity	/	87.62 (453)		/	82.60 (921)	
ResNet50	Sensitivity	92.83 (466)	/	0.7871	61.72 (208)	/	0.5328
	Specificity	/	85.88 (444)		/	91.56 (1021)	

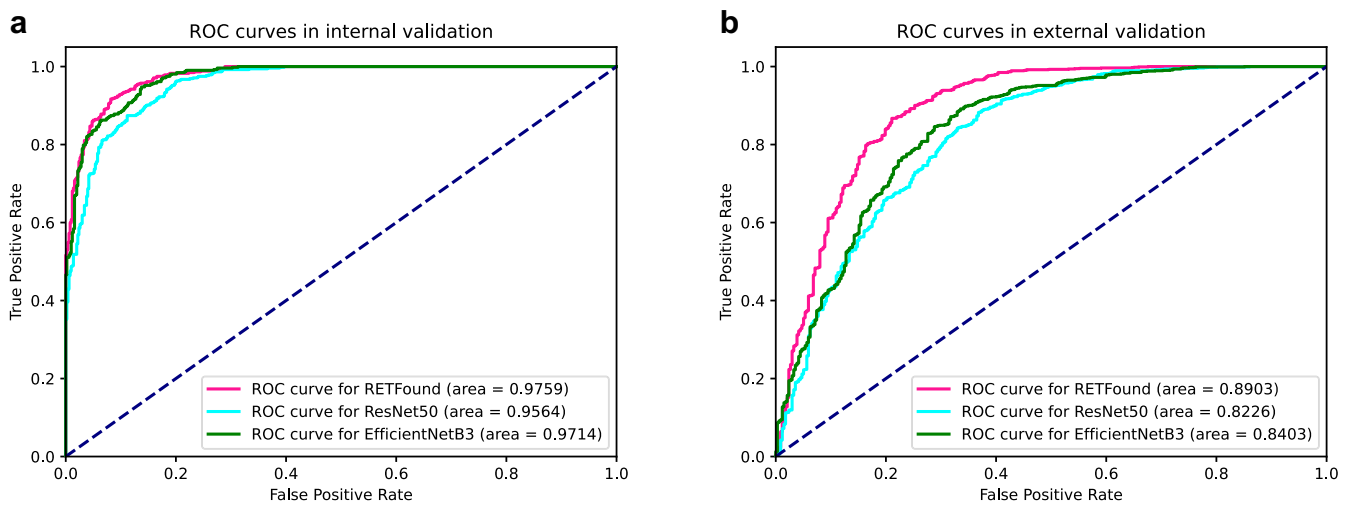


Fig. 4 | ROC curves of each model in internal and external validation. a internal validation. b external validation. ROC receiver operating characteristic curve.

important to emphasise that our primary focus was not to conduct direct comparisons with commercial models. Instead, we focused on exploring the feasibility of developing automated screening tools based on foundational models using real-world and population-specific data. Second,

the assessment of the accuracy and utility of the RETFound framework model in this study was based on the SDEDS Image Dataset without its actual implementation in the field. Nevertheless, our findings indicate that adopting the RETFound framework model can lead to substantial

Fig. 5 | Display of typical images from the SDEDS image dataset. During community eye disease screenings, photographs are often taken by trained general practitioners or nurses under non-darkroom and non-dilated conditions, resulting in image quality that contrasts sharply with those taken by professional optometrists or ophthalmologists in ophthalmic clinical institutions under darkroom and dilated conditions. SDEDS Shanghai Digital Eye Diseases Screening Program, CFPs colourful fundus photographs, AMD age-related macular degeneration, PM pathological myopia, DR diabetic retinopathy.

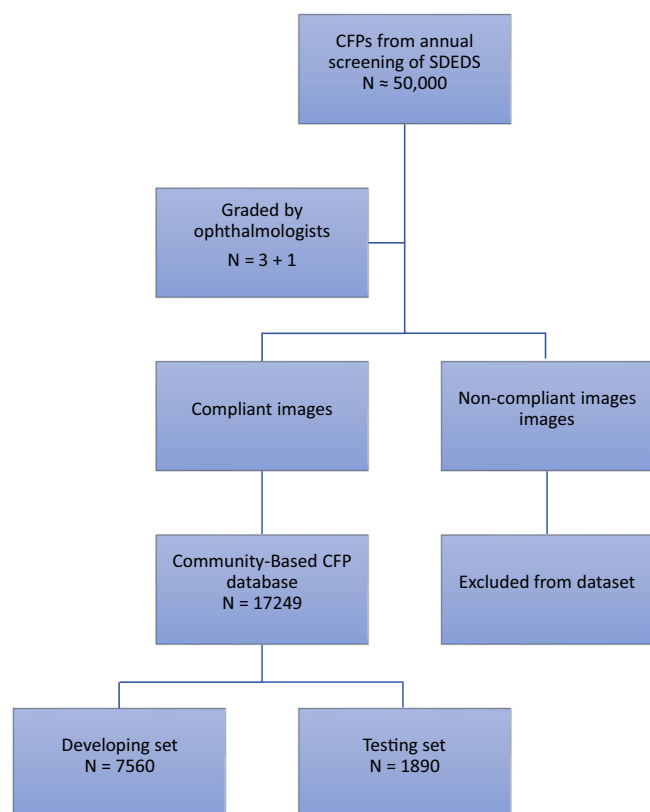
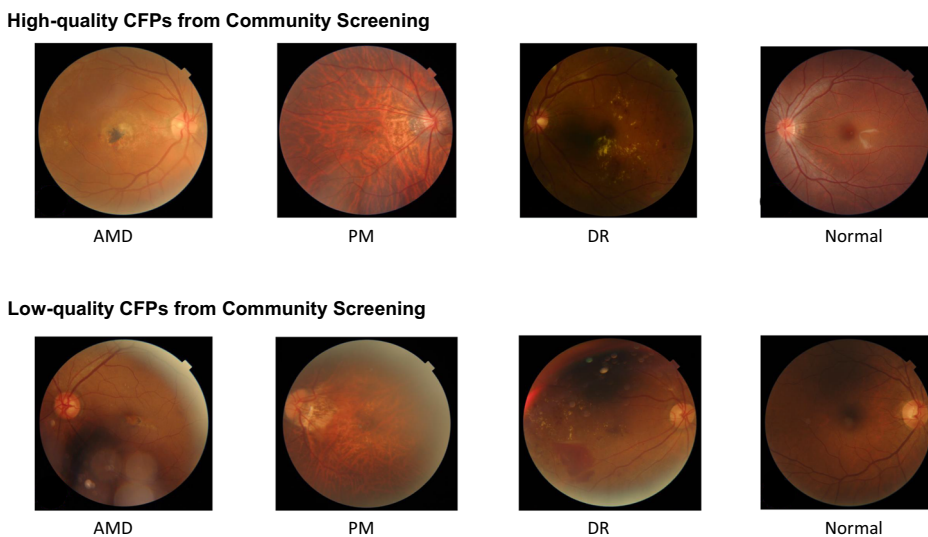


Fig. 6 | SDEDS image dataset inclusion and building workflow. The final selection of all fundus photographs underwent a thorough quality review and was meticulously annotated.

improvements in accuracy and net benefits. This finding strengthens the robustness of our conclusions.

In conclusion, the development of AI methods for eye disease screening using RETFound can yield net benefits that exceed those of commercial models and have better generalisability than traditional CNN models, while the development process is rapid. Thus, we recommend the application of the RETFound framework in real-world community screening for blinding eye diseases and encourage commercial companies to advance their intelligent eye disease screening products based on RETFound.

Methods

The study was divided into three parts. We conducted a community eye disease screening program in Shanghai, China (Shanghai Digital Eye Disease Screening Program [SDEDS]). Since 2021, our team has retrospectively identified images of suspected eye diseases from past screenings and organised ophthalmology experts to conduct image readings for diagnosis (diagnostic criteria in Supplementary Table 1), gradually building a community eye disease screening image dataset (SDEDS dataset; Fig. 5). Currently, this dataset encompasses 17,249 images, including 1432 of age-related macular degeneration (AMD), 1682 of diabetic retinopathy (DR), 2485 of glaucoma, 748 of pathologic myopia (PM), 5334 of tessellated fundus, and 5568 of normal fundus. Image datasets continue to expand dynamically. First, we conducted a cross-sectional study. We developed a DL model enhanced by RETFound, based on transfer learning and the SDEDS dataset. We compared the accuracy of this model in multi-disease eye disease screening with that of two commercial models (anonymous models S and Y) that are widely used in China. The relevant results are derived from real-world operational databases and do not involve company participation, making them unsuitable for disclosing specific company names. Second, we combined the aforementioned accuracy and prevalence of eye diseases in urban and rural areas of China¹ as parameters and constructed a hypothetical cohort of 100,000 individuals. The DCA technique was employed to evaluate the net benefit of implementing the RETFound-enhanced model for individual ocular disease screening in urban and rural areas of China. Third, we conducted a detailed comparison between the RETFound-enhanced DL model and traditional convolutional neural network (CNN) models trained via SL on ImageNet.

Part one: construction and evaluation of RETFound-enhanced DL model

The data used in this study were sourced from the SDEDS dataset (Fig. 6). Each fundus image was independently classified and annotated by three ophthalmologists. In cases of discrepancy, collective deliberation involving ophthalmologists and a senior retinal specialist was convened to determine the final diagnoses. All images were re-evaluated based on the following criteria: the retinal fovea was not fully visible or obscured in over 50% of the total area, blurriness, severe artefacts, low contrast, uneven lighting, and excessive reflectance. Eventually, from the pool of images conforming to the criteria and upon expert review, a random assortment of 7560 images encompassing DR, PM, AMD, and no-eye disease was used as the development dataset. An additional 1890 images, including DR, PM, AMD, and no-eye disease images, were randomly chosen to constitute the test dataset.

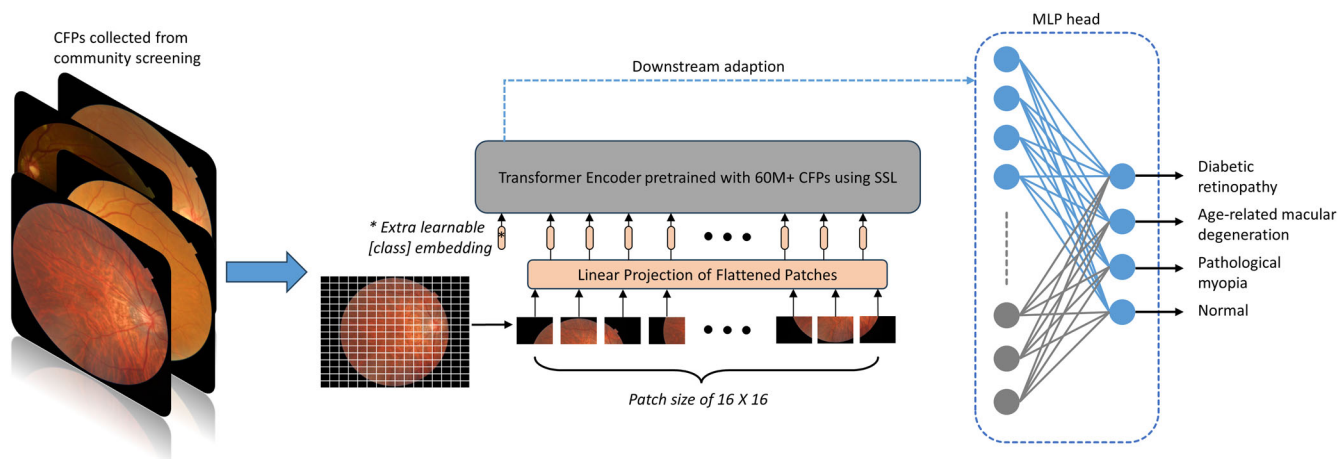


Fig. 7 | Overview of the RETFound-enhanced DL model for community-based fundus diseases screening. The model consists of a pretrained transformer encoder and a MLP head.

Our study adhered to the principles of the Declaration of Helsinki and was approved by the ethics committee of Shanghai Eye Diseases Prevention and Treatment Centre. This study exclusively utilised retrospective data, with all images undergoing irreversible anonymisation and no active patient engagement, so informed consent was deemed not applicable. No commercial interest was implicated in the design or execution of this study.

The construction process of the RETFound-enhanced eye disease screening model is illustrated in Fig. 7. We employed the encoder component of RETFound, which utilises the ViT-large architecture²⁷ and features 24 transformer blocks with an embedding vector size of 1024. The encoder accepts unmasked patches (with a patch size of 16 × 16) as input and projects them into a feature vector with a size of 1024. The 24 transformer blocks, which comprise multiheaded self-attention and a multilayer perceptron, process these feature vectors to generate high-level features. Subsequently, these high-level features are input into a multilayer perceptron (MLP) head, which produces the final predicted categories.

Of the 7560 images in the development dataset, 6599 were designated as the training set, including 1002 images of AMD, 1177 of DR, 523 of PM, and 3897 of normal fundus, and 941 were allocated to the validation set, encompassing 143 images of AMD, 168 of DR, 74 of PM, and 556 of normal fundus. All images were resized to 256 × 256 pixels using cubic interpolation. All images underwent the same data augmentation procedures as those during model training, including random cropping (with a cropping range of 20% to 100% of the entire image), followed by resizing the cropped image blocks to 224 × 224, random horizontal flipping, and image normalisation. The training objective was to generate classification outputs congruent with the labels. In this study, four categories were used: DR, PM, AMD, and normal fundus. The training was conducted using four NVIDIA GeForce RTX 2080 Ti GPUs, with CUDA version 11.1, powered by an Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50 GHz, in the Ubuntu 18.04 system environment with 86 GB of memory. The batch size was set to 16. A total of 50 training epochs were set, with the initial 10 epochs designated for the warm-up phase (the learning rate gradually increased from 0 to 5 × 10⁻⁴), followed by cosine annealing scheduling (the learning rate gradually descended from 5 × 10⁻⁴ to 1 × 10⁻⁶). After each epoch, the model was evaluated using a validation set. The model weights with the highest AUROC in the validation set were preserved as model checkpoints for testing and DCA.

After the training phase, we determined the precise diagnostic sensitivity and specificity of our model using a test set. Subsequently, we employed the χ^2 test (or Fisher’s Exact test) and post-hoc pairwise comparisons (via the Bonferroni method, $\alpha=0.05/3$) to compare the sensitivity and specificity of our RETFound-enhanced model with those of two commercial models.

Part two: decision curve analysis

The primary outcome was the net benefit. The metrics commonly used to evaluate prediction models include sensitivity and specificity. However, these measures do not provide insight into the practical applicability of the model. The thresholds for sensitivity and specificity necessary to endorse its clinical use are ambiguous. Likewise, the level of miscalibration deterring the use of a prediction model or the criteria for selecting between two models—one with superior calibration and the other with enhanced discrimination—remain undefined. Therefore, decision curves have emerged as a prevalent tool for assessing the clinical utility of prediction models by analysing their net benefits²⁸.

The net benefit is expressed as

$$Net\ benefit = \frac{TruePositiveCount}{n} - \frac{P_t}{1 - P_t} \times \frac{FalsePositiveCount}{n} \quad (1)$$

where P_t represents the probability threshold at which the expected benefit of engaging in subsequent therapy (or further testing) balances the expected benefit of avoidance. In the context of diagnostic testing, doctors are required to discern the precise risk level that merits further intervention. For instance, some may consider a 10% risk of blinding diseases warranting further therapy after an adverse reaction assessment, whereas others may suggest a 20% risk criterion with a more cautious stance. This risk cutoff point was characterised as the probability threshold in the decision curve analysis. One model may be favoured over another if its net benefit exceeds that of the other models at the selected threshold probability²⁸.

We used DCA to compare the net benefits of applying our RETFound-enhanced DL model and two commercial models to real-world scenarios. DCA is a statistical technique for evaluating the clinical outcomes of models and tests. Traditional accuracy metrics, such as the AUROC or Brier score, disregard situational considerations. DCA assesses the net benefit of a model against the two standard strategies of treating all patients or treating none.

Part three: comparison with CNN baselines

CNNs have been the standard for automated medical image diagnosis over the last decade²⁹. Transformers, particularly ViTs, have recently gained prominence. To make further comparisons with traditional CNN models, we designed the following task:

We chose ResNet50 and EfficientNetB3, pretrained using SL on ImageNet-21k, as representatives. Three commonly used public datasets (MESSIDOR-2, APTOS-2019, and IDRiD) were selected. The automatic diagnosis of DR based on fundus images was one of the earliest applications of DL in ophthalmology. Relevant public fundus image datasets are numerous, well recognised, and of excellent quality. Therefore, we focused

Table 3 | Characteristic of three public datasets of DR fundus images

Dataset	DR	Normal	Total
MESSIDOR-2	731	1017	1748
APTOS-2019	1496	1433	2929
IDRiD	279	134	413
Total	2506	2584	5090

on the DR to compare our RETFound-enhanced model with the two CNN models. Basic information about the data is presented in Table 3.

Eighty percent of these public datasets were randomly selected to train the three models, and 20% were used for internal validation. The training process was the same as Part One, and 50 epochs were performed. The model parameters with the highest accuracy in the internal validation set were saved and tested using the DR part of the test dataset in Part One for external validation.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The export of human-related data is governed by the Ministry of Science and Technology of China (MOST) in accordance with the Regulations of the People's Republic of China on Administration of Human Genetic Resources (State Council No.717). A request for the non-profit use of the fundus images in the SDEDS should be sent to corresponding author Yingyan Ma.

Code availability

The code of RETFound model for fine-tuning own model is available at https://github.com/rmaphoh/RETFound_MAE. The code for finetuning our model is available at <https://github.com/Akemimadokami/DL-Model-for-Community-based-Fundus-Disease-Screening>. The analytical procedure for DCA was sourced from the SAS macro available at <https://raw.githubusercontent.com/ddsjoberg/dca.sas/main/dca.sas>.

Received: 14 October 2023; Accepted: 12 April 2024;

Published online: 30 April 2024

References

1. Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study. *Lancet Glob. Health* **9**, e130–e143 (2021).
2. Man, R. E. K. et al. The differential impact of age on vision-related quality of life across the visual impairment spectrum. *Ophthalmology* **128**, 354–363 (2021).
3. Ehrlich, J. R. et al. Association between vision impairment and mortality: a systematic review and meta-analysis. *Lancet Glob. Health* **9**, e418–e430 (2021).
4. Cheng, C. Y. et al. Prevalence and causes of vision loss in East Asia in 2015: magnitude, temporal trends and projections. *Br. J. Ophthalmol.* **104**, 616–622 (2020).
5. Ting, D. S. W. et al. Artificial intelligence and deep learning in ophthalmology. *Br. J. Ophthalmol.* **103**, 167–175 (2019).
6. Liu, H. et al. Economic evaluation of combined population-based screening for multiple blindness-causing eye diseases in China: a cost-effectiveness analysis. *Lancet Glob. Health* **11**, e456–e465 (2023).
7. Morrison, S. L., Dukhovny, D., Chan, R. V. P., Chiang, M. F. & Campbell, J. P. Cost-effectiveness of artificial intelligence-based retinopathy of prematurity screening. *JAMA Ophthalmol.* **140**, 401–409 (2022).
8. Xie, Y. et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. *Lancet Digit. Health* **2**, e240–e249 (2020).
9. Lee, A. Y. et al. Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. *Diabetes Care* **44**, 1168–1175 (2021).
10. Xu, Y. et al. The diagnostic accuracy of an intelligent and automated fundus disease image assessment system with lesion quantitative function (SmartEye) in diabetic patients. *BMC Ophthalmol.* **19**, 184 (2019).
11. Lin, S., Li, L., Zou, H., Xu, Y. & Lu, L. Medical staff and resident preferences for using deep learning in eye disease screening: discrete choice experiment. *J. Med. Internet Res.* **24**, e40249 (2022).
12. Cheung, H. M. C. & Rubin, D. Challenges and opportunities for artificial intelligence in oncological imaging. *Clin. Radiol.* **76**, 728–736 (2021).
13. Zou, H. D., Lu, L. N., Xu, Y. & Lin, S. L. Challenges of artificial intelligence used for eye disease screening in recent China communities. *Zhonghua Yan Ke Za Zhi* **58**, 743–746 (2022).
14. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
15. Zhang, J. & Zou, H. Insights into artificial intelligence in myopia management: from a data perspective. *Graefes Arch. Clin. Exp. Ophthalmol.* <https://doi.org/10.1007/s00417-023-06101-5> (2023).
16. Xu, T. et al. Prevalence and causes of vision loss in China from 1990 to 2019: findings from the Global Burden of Disease Study 2019. *Lancet Public Health* **5**, e682–e691 (2020).
17. Ipp, E. et al. Pivotal evaluation of an artificial intelligence system for autonomous detection of referable and vision-threatening diabetic retinopathy. *JAMA Netw. Open* **4**, e2134254 (2021).
18. Keane, P. A. & Topol, E. J. With an eye to AI and autonomous diagnosis. *NPJ Digit. Med.* **1**, 40 (2018).
19. Cen, L. P. et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nat. Commun.* **12**, 4828 (2021).
20. Pei, X. et al. Efficacy of artificial intelligence-based screening for diabetic retinopathy in type 2 diabetes mellitus patients. *Diabetes Res. Clin. Pract.* **184**, 109190 (2022).
21. Wang, W. et al. *Two-stream CNN With Loose Pair Training For Multimodal AMD Categorization*. p. 156–164 (Springer International Publishing) https://link.springer.com/chapter/10.1007/978-3-030-32239-7_18 (2019).
22. Zhang, W. F. et al. The validation of deep learning-based grading model for diabetic retinopathy. *Front. Med. (Lausanne)* **9**, 839088 (2022).
23. Li, F. et al. Deep learning-based automated detection for diabetic retinopathy and diabetic macular oedema in retinal fundus photographs. *Eye* **36**, 1433–1441 (2022).
24. Hafiz, F. et al. A new approach to non-mydriatic portable fundus imaging. *Expert Rev. Med. Devices* **19**, 303–314 (2022).
25. Mackay, D. D. & Bruce, B. B. Non-mydriatic fundus photography: a practical review for the neurologist. *Pract. Neurol.* **16**, 343–351 (2016).
26. Lin, S. et al. Artificial intelligence in community-based diabetic retinopathy telemedicine screening in urban china: cost-effectiveness and cost-utility analyses with real-world data. *JMIR Public Health Surveill.* **9**, e41624 (2023).
27. Dosovitskiy, A. et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arxiv:2010.11929[cs.CV,cs.AI,cs.LG] <https://arxiv.org/abs/2010.11929> (2020).
28. Vickers, A. J., Van Calster, B. & Steyerberg, E. W. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* **352**, i6 (2016).
29. Zhang, J. & Zou, H. Artificial intelligence technology for myopia challenges: a review. *Front Cell Dev. Biol.* **11**, 1124005 (2023).

Acknowledgements

This study was funded by the Chinese National key research and development program (2021YFC2702100), key disciplinary projects of the 3-year action plan to strengthen the construction of the public health system

in Shanghai (GWVI-11.1-30), Science and Technology Commission of Shanghai Municipality (23ZR1481000), and Shanghai Municipal Health Commission (2022YQ051 and 2024HP44).

Author contributions

J.Z., S.L., H.Z. and Y.M. conceptualised the study. J.Z., Y.X., L.L., J.H., Y.P. and T.Y. collected the data. J.Z., S.L. and Y.M. developed the system, conducted formal analysis, and prepared the first draft of the manuscript. T.C. and Y.Z. assisted in conducting model comparison experiments and analysing the results. H.Z. and Y.M. critically reviewed drafts of the manuscript. All authors approved the final version to be submitted.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41746-024-01109-5>.

Correspondence and requests for materials should be addressed to Yuejie Zhang, Haidong Zou or Yingyan Ma.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024