npj Digital Medicine

https://doi.org/10.1038/s41746-026-03118-y

Article in Press

Uncertainty-inspired open-set model for identifying infantile fundus abnormalities

Received: 21 January 2026

Accepted: 2 August 2026

Published online: 15 August 2026

Cite this article as: Zhao, X., Wu, Z., Zhu, X. *et al.* Uncertainty-inspired open-set model for identifying infantile fundus abnormalities. *npj Digit. Med.* (2026). https://doi.org/10.1038/s41746-026-03118-y

Xinyu Zhao, Zhenquan Wu, Xuemei Zhu, Mengrui Zhang, Kaixuan Cui, Shaobin Chen, Sihui Zhang, Yarou Hu, Shirou Wu, Ruyin Tian, Yi Chen, Xiaorong Cheng, Na Duan, Ling Jiang, Na Li, Wenbin Wei, Wei Chi, Guishuang Ying, Baiying Lei, Jianhong Liang & Guoming Zhang

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/.

ARTICLE IN PRESS

Uncertainty-inspired open-set model for identifying infantile fundus abnormalities

Xinyu Zhao[1,2]; Zhenquan Wu[1,2*]; Xuemei Zhu[3]; Mengrui Zhang[1]; Kaixuan Cui[1,2]; Shaobin Chen[4]; Sihui Zhang[1]; Yarou Hu[1,2]; Shirou Wu[1]; Ruyin Tian[1]; Yi Chen[1]; Xiaorong Cheng[1]; Na Duan[5]; Ling Jiang[6]; Na Li[7]; Wenbin Wei[8]; Wei Chi[1,2]; Guishuang Ying[9]; Baiying Lei[4*]; Jianhong Liang[3*]; Guoming Zhang[1,2*]

These authors contributed equally: Xinyu Zhao, Zhenquan Wu, Xuemei Zhu.

Affiliations:

1. Shenzhen Eye Hospital, Shenzhen Eye Medical Center, Southern Medical University, Shenzhen, China.
2. Jinan University, Guangzhou, China.
3. Peking University People's Hospital, Eye diseases and Optometry Institute, Beijing Key Laboratory of Diagnosis and Therapy of Retinal and Choroid Diseases, Clinical Centre of Optometry, College of Optometry, Peking University Health science center, Beijing, China.
4. National-Regional Key Technology Engineering Lab. for Medical Ultrasound, Guangdong Key Lab. for Biomedical Measurements and Ultrasound Imaging, Marshall Lab. of Biomedical Engineering, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen, China.
5. Huizhou Third People's Hospital, Huizhou, China.
6. Fuzhou University affiliated Provincial Hospital, Fujian Provincial Hospital, Fuzhou, Fujian Province, China.
7. Kunming Maternal and Children's Hospital, Kunming, Yunnan, China.
8. Beijing Tongren Eye Center, Beijing Key Laboratory of Intraocular Tumor Diagnosis and Treatment, Beijing Ophthalmology & Visual Sciences Key Lab, Medical Artificial Intelligence Research and Verification Key Laboratory of the Ministry of Industry and Information Technology, Beijing Tongren Hospital, Capital Medical University, Beijing, China.
9. Center for Preventive Ophthalmology and Biostatistics, Department of Ophthalmology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA.

ARTICLE IN PRESS

ARTICLE IN PRESS

Corresponding authors:

1. Guoming Zhang, Shenzhen Eye Hospital, Shenzhen Eye Medical Center, Southern Medical University, 18 Zetian Road, Futian District, Shenzhen, 518040, China; zhang-guoming@163.com.
2. Jianhong Liang, Peking University People's Hospital, Eye diseases and Optometry Institute, Beijing Key Laboratory of Diagnosis and Therapy of Retinal and Choroid Diseases, Clinical Centre of Optometry, College of Optometry, Peking University Health science center, Beijing, China; drliangjianhong@126.com.
3. Baiying Lei, National-Regional Key Technology Engineering Lab. for Medical Ultrasound, Guangdong Key Lab. for Biomedical Measurements and Ultrasound Imaging, Marshall Lab. of Biomedical Engineering, School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen, 518060, China; leiby@szu.edu.cn.
4. Zhenquan Wu, Shenzhen Eye Hospital, Shenzhen Eye Medical Center, Southern Medical University, 18 Zetian Road, Futian District, Shenzhen, 518040, China; wzq169cn@163.com.

ARTICLE IN PRESS

Abstract

Accurate and safe identification of infantile fundus abnormalities is critical for large-scale screening, yet remains challenging because of heterogeneous disease presentations and overconfident predictions by conventional artificial intelligence (AI) models. In this study, we aim to develop an uncertainty-inspired open-set (UIOS) system to address these challenges. Using a multicenter dataset from 39 centers comprising 319,998 fundus images from 15,647 infants, the UIOS-YOLOv8 (UIOS-Y) model integrates multi-class classification with uncertainty estimation to enable uncertainty-inspired prediction and triage for clinical decision-making. The UIOS-Y model achieved high performance in internal testing (AUC 0.995). The UIOS-Y+θ model, by excluding high-uncertainty prediction, achieved an AUC of 0.997 and enabled effective triage of clinically ambiguous cases for clinician review. Robust performance was maintained across five independent external datasets (AUCs ranging from 0.969 to 0.982). UIOS also effectively detected out-

ARTICLE IN PRESS

of-distribution images, enhancing safety in real-world deployment. Notably, it outperformed resident physicians and general-purpose AI systems, including ChatGPT and Gemini (all $p < 0.001$), particularly in complex or ambiguous cases. These findings underscore that dedicated uncertainty-inspired open-set modeling is critical for reliable identification of infantile fundus abnormalities despite of the availability of general-purpose AI chatbots, and highlights its value in real-world AI-assisted pediatric ophthalmic screening.

Introduction

Infantile fundus diseases are among the leading causes of irreversible vision loss in early life.[1] The first year of life represents a critical period for visual system maturation and the establishment of binocular vision.[2,3] The World Health Organization has emphasized that early-onset irreversible vision loss in infants can lead to delays in motor, language, emotional, social, and cognitive development, resulting in lifelong consequences.[4] Without timely and targeted interventions, the global burden of vision loss is expected to continue to increase.[5] Hence, routine screening and early intervention are essential to mitigate the risk of lifelong vision loss in infants.[3]

However, due to limited healthcare resources and insufficient clinical support, especially during the early infancy, screening systems for infant eye abnormalities are often incomplete or unavailable. As a result, fundus abnormalities in infants frequently remain undetected and untreated in a timely manner. Improving access to timely and accurate eye care for infants is thus critical for advancing health equity and achieving sustainable development goals.

ARTICLE IN PRESS

Artificial intelligence (AI) has rapidly advanced across various ophthalmic domains.[6-10] It holds substantial promise for improving diagnostic efficiency and expanding screening coverage. Our team has previously developed AI systems for identifying retinopathy of prematurity (ROP)[11,12] and multiple fundus abnormalities.[13] However, both our prior AI models and existing approaches are largely limited to single-task settings.[14] These models are typically developed under closed-set assumptions, whereby all test samples are presumed to belong to predefined disease categories. In real-world clinical settings, however, atypical presentations, rare diseases, image artifacts, and previously unseen conditions are common. Under such circumstances, these AI models may produce overconfident yet incorrect predictions when encountering out-of-distribution (OOD) samples, thereby posing potential clinical risks.[15] Moreover, constructing comprehensive training datasets that encompass the full spectrum of infantile fundus abnormalities is practically infeasible. These current limitations underscore the urgent need for innovative AI strategies that can support the real-world clinical application.

Recent advances in medical imaging have highlighted the importance of uncertainty awareness and open-set recognition.[15-20] Uncertainty-inspired AI systems are increasingly regarded as more suitable for clinical decision support, particularly in rare diseases or ambiguous cases.[18-20] Open-set learning frameworks enable models to accurately identify known diseases while simultaneously handling unfamiliar or unreliable inputs, without requiring exhaustive labeled datasets. Although such approaches have shown promise in adult retinal disease classification,[15-17] their application in infantile fundus diseases remains limited due to differences between infantile and adult fundus

ARTICLE IN PRESS

characteristics, thereby hindering the translation of these methods into pediatric ophthalmology in real world.

In this study, we aim to develop an uncertainty-inspired open-set (UIOS) system for the identification of infantile fundus abnormalities. As illustrated in Fig. 1, when presented with fundus images from known disease categories, the UIOS model produces confident diagnostic outputs with low uncertainty. In contrast, for OOD samples, atypical presentations, or low-quality images, the UIOS model generates high uncertainty scores, flagging these cases for clinician review. We further compared the performance of the UIOS model with that of clinicians and AI chatbots to evaluate the gap between domain-specific and general-purpose AI models and demonstrate its clinical applicability.[21,22] This design enhances clinical reliability and provides a more scalable, practical solution for AI-assisted screening of infantile fundus abnormalities in real-world settings.

Results

Dataset distribution

The dataset was collected from a total of 39 centers, including 36 centers within an Eyecare-Cloud-based platform[23] and three additional centers contributing external test data. In addition, two publicly available datasets were also used for testing. Platform-derived images were downloaded and underwent standardized processing, including renaming, de-identification, and exclusion of non-fundus images (e.g., fundus fluorescein angiography [FFA] and anterior segment images) and low-quality images. External

ARTICLE IN PRESS

datasets were provided directly by collaborating centers and processed using the same standardized pipeline.

These fundus images were acquired using multiple imaging devices, primarily including RetCam (Clarity Medical Systems, Inc., USA), the ROP Screening System (SW-8000, China), and the Nautilus imaging system (RS-B002, China). In total, the dataset comprised 319,998 fundus images from 15,647 infants, the majority of whom were younger than 1 year. These images encompassed 11 clinical categories, a summary of which was presented in Table S1. The internal dataset was divided into a training set (13,047 individuals; 122,946 images), a validation set (10,465 individuals; 40,980 images), and a test set (10,448 individuals; 40,990 images) (Fig. S1). External datasets, used primarily for independent testing, comprised 2,030 individuals and 115,082 images. These data were collected from three independent clinical centers, including Fujian Provincial Hospital (FJPH), Kunming Maternal and Children's Hospital (KMCH), and Peking University People's Hospital (PUPH), as well as two publicly available datasets, FARFUM-RoP[24] and the ROP dataset from Ostrava (hereafter referred to as Ostrava-ROP) [25]. For brevity, these datasets are denoted as external testing set 1-5 respectively.

Normal cases (Class 1) constituted the largest proportion of the dataset, reflecting real-world screening scenarios, whereas abnormality categories showed varying degrees of imbalance at both the individual and image levels. Retinal fold (RTF, Class 2), retinoblastoma (Class 3), different stages of ROP (Classes 4-7), laser scars (Class 8),

ARTICLE IN PRESS

and retinal hemorrhage (Class 11) were consistently represented across internal and external datasets, supporting reliable performance evaluation. In contrast, rare conditions, such as Coats' disease (Class 9) and chorioretinal coloboma (Class 10), were represented by fewer individuals and images, consistent with their low clinical prevalence. External test sets exhibited heterogeneous distributions in both individual numbers and image numbers across centers, providing a rigorous assessment of model generalizability under diverse real-world clinical settings (Table S1).

Model performance in the internal test dataset

In the internal test dataset, UIOS-YOLOv8 (UIOS-Y) and UIOS-ResNet-50 (UIOS-R) models achieved an average area under the receiver operating characteristic (ROC) curve (AUC) of 0.995 and 0.982, respectively, across the 11 categories. The average accuracy, precision, recall, specificity, and F1-score of the AI models were displayed in Table S2. Images with uncertainty scores exceeding the UIOS model's threshold were flagged for ophthalmologist review. Overall, 11.8% of the samples exceeded the predefined threshold (Table S3). After exclusion of uncertain samples, the UIOS-Y+θ model achieved an average AUC of 0.997, whereas the UIOS-R+θ model achieved 0.975. (Fig. 2, Figs. S2-S5). Performance-coverage trade-off curves further demonstrated the performance of the UIOS-Y model across internal and external datasets (Fig. S6). Fig. S7 and Fig. S8 presented representative image classification results, comparing YOLOv8 with UIOS-Y and ResNet-50 with UIOS-R, respectively.

ARTICLE IN PRESS

Fig. 3 demonstrated that UIOS-Y achieves more reliable image-level classification than UIOS-R in cases with complex and coexisting ROP features. Misclassifications by UIOS-R primarily occurred when multiple pathological characteristics, such as aggressive ROP(A-ROP) overlapping with stage 2-3 ROP or severe plus disease, were present within the same image, highlighting the limitations of single-label classification in clinically ambiguous scenarios. Importantly, all such cases exhibited uncertainty scores exceeding the predefined threshold, enabling the system to flag these images for clinician review and thereby reducing the risk of diagnostic errors. Comparison of prediction confidence distributions across YOLOv8, ResNet-50, UIOS-Y, and UIOS-R further reveals differences in confidence concentration and uncertainty handling among models, highlighting that integrating uncertainty estimation with classification enhances the robustness and safety of AI-assisted clinical decision-making.

Model performance in the external datasets

Across multiple external independent datasets (FJPH, KMCH, PUPH, FARFUM-RoP, and NJPROP), all models demonstrated strong discriminatory performance, although their generalizability varied across centers. UIOS-Y consistently achieved high AUCs across external cohorts, ranging from 0.969 to 0.993, indicating robust performance in diverse clinical settings. In contrast, UIOS-R exhibited greater variability in generalization, with AUCs ranging from 0.913 to 0.970 across datasets. The proportions of samples exceeding the predefined threshold were 24.5% (4,418), 31.7% (1,369), 26.5% (22,756),

ARTICLE IN PRESS

40.0% (584), and 24.8% (1,320), respectively (Table S3). After excluding uncertain samples, the UIOS-Y+θ model achieved AUCs ranging from 0.974 to 0.998, whereas the UIOS-R+θ model achieved AUCs ranging from 0.966 to 0.997(Fig. 2, Figs. S2-S5, Table S2).

Comparison of UIOS models with AI chatbots

A total of 110 images, with 10 images per class, were independently evaluated by the UIOS-Y model, the UIOS-Y+θ model, one retinal specialist, one attending doctor, one resident doctor, Gemini-3.0 Pro, and ChatGPT-5.2. As shown in Fig. 4 and Table S4, the UIOS-Y models demonstrated superior and more consistent classification performance compared with clinician groups and AI chatbots. The UIOS-Y and UIOS-Y+θ models achieved the highest overall accuracies, comparable to those of retinal specialists and attending doctors (all $p > 0.05$), while significantly outperforming resident doctors as well as Gemini-3.0 Pro and ChatGPT-5.2 (all $p < 0.001$) (Table S4). Notably, the UIOS-Y and UIOS-Y+θ models maintained stable performance across multiple categories, while clinicians, particularly less experienced doctors, and AI chatbots showed greater variability and reduced accuracy in complex classes.

Heatmap visualization for model explainability

Heatmap visualizations highlighted characteristic regions associated with diagnostically relevant features, including retinal folds, retinoblastoma lesions, different

ARTICLE IN PRESS

stages of ROP, laser scars, exudation in Coats' disease, chorioretinal coloboma, and hemorrhage (Fig. 5), thereby further validating the strong capability of the UIOS-Y model to localize clinically meaningful features.

Out-of-distribution image identification

Across seven kinds of image datasets outside the training categories (Table S5), including FFA images, optical coherence tomography angiography (OCTA) images, color fundus images with different imaging tools, and anterior segment images with leukocoria, the UIOS model processed all 6,299 images from 1,978 participants. Of these, 5,421 images (86.1%) exceeded the uncertainty threshold, indicating effective identification of OOD samples. This capability may help reduce potential diagnostic errors in real-world clinical settings.

Among these OOD samples, some images with uncertainty scores below the threshold were correctly diagnosed by the UIOS-Y model, and these were predominantly normal images. Conversely, the model was still capable of generating accurate predictions for certain images with uncertainty scores above the threshold (Fig. S9). A greater discrepancy in imaging modality was associated with a higher proportion of images exceeding the threshold (e.g., OCTA and scanning laser ophthalmoscopy [SLO]). By contrast, images from different modalities but with lesion characteristics similar to those used for model development in this study showed a moderately lower proportion exceeding the threshold (e.g., FFA images with ROP in Row 1 of Table S5). Notably, even

ARTICLE IN PRESS

within the same imaging modality, images with markedly different disease features showed a high proportion exceeding the threshold (e.g., RetCam-captured fundus images in Row 7 of Table S5). When presented with RetCam fundus images containing cataract, retinal exudates, or morning glory syndrome (Fig. S10), 89.5% of images exceeded the threshold under the UIOS-Y model. Given that these above-threshold predictions were incorrect or incomplete, elevated uncertainty effectively flagged such cases, thereby reducing the risks of diagnostic errors. These findings further highlight the advantage of the UIOS-Y model in open-set recognition and OOD detection.

We analyzed the uncertainty density distributions across internal, external, and OOD datasets (Fig. S11). Uncertainty scores for internal datasets were tightly clustered at low scores, indicating stable and high-confidence predictions. In contrast, external datasets show broader distributions with a noticeable rightward shift, reflecting increased uncertainty when applied to heterogeneous clinical data. The OOD dataset demonstrated the highest uncertainty. These patterns indicate that the UIOS-Y model can effectively distinguish between familiar and unfamiliar or clinically challenging cases, thereby enhancing the safety and reliability of real-world application.

WeChat application and web-based platform

We conducted a preliminary deployment analysis to better reflect real-world clinical workflows in future applications. The “Fundus disease screening platform” WeChat application was developed based on the UIOS-Y model (Fig. 6 and Video S1). Users can

ARTICLE IN PRESS

easily complete the process by following the instructions within the application interface. The application allows users to review previous identification records for follow-up and to monitor disease progression. An administrator account is available for clinicians to review and document medical data. In addition, a web-based platform was developed (Fig. S12, Video S2). Users can choose between the WeChat application and the web-based platform based on their preference and convenience in future applications.

Under the online deployment framework (WeChat Mini Program to FastAPI) using the same nine images, the latency reduced to 131.57/91.08 ms (P95: 199.95 ms; ~7.60 images/s), while CPU RAM usage was 1264.0/1246.8 MB (Δ RAM: +31.4/+14.3 MB) and GPU memory usage was 394.4/394.4 MB (Δ GPU: +65.5/+65.5 MB; peak: 402.8 MB). Under the Gradio-based web deployment (batch size = 9), the end-to-end inference latency achieved a mean/median of 168.96/179.80 ms (P95: 337.65 ms; ~5.92 images/s), with post-inference CPU RAM usage of 1097.38/1076.50 MB (Δ RAM: +57.19/+56.00 MB) and GPU memory usage of 195.20/195.20 MB (Δ GPU: +72.81/+65.70 MB).

Discussion

In this study, we developed UIOS models for the identification of infantile fundus abnormalities. These models facilitate the early detection of warning signs, enabling timely medical intervention and potentially preserving both vision and life. The UIOS models demonstrated robust performance across multiple external test datasets and

ARTICLE IN PRESS

public datasets. In addition, they achieved markedly higher classification accuracy than general-purpose AI chatbots, with performance comparable to that of retinal specialists.

With the continuous advancement of AI technology across various fields,[26] identifying common diseases is no longer a major challenge for conventional AI models. The primary difficulty lies in diagnosing OOD images, including complex lesions, rare diseases, or previously unseen abnormalities.[15] For in-distribution or simple lesions, the UIOS model provides confident predictions with uncertainty scores well below the threshold, similar to conventional AI models. However, when confronted with OOD lesions, traditional AI models often produce confident predictions constrained to predefined categories. In contrast, the UIOS models generate predictions accompanied by elevated uncertainty scores, prompting clinician review. The UIOS+θ model represents the task-specific integration of classification, uncertainty estimation, and class-wise rejection calibration into a unified and clinically deployable framework for infantile fundus image analysis. This design better aligns with clinicians' hierarchical diagnostic reasoning and may reduce AI-induced diagnostic errors. Notably, images with uncertainty scores exceeding the threshold were predominantly those with poor image quality or blurring, indistinct lesion features, or multiple concurrent lesions, all of which inherently require additional clinical review thereby supporting broader clinical adoption of this framework.

Several studies have applied the concepts of uncertainty and open set to the intelligent diagnosis of adult fundus diseases, achieving widespread recognition.[15-20]

ARTICLE IN PRESS

However, these approaches have predominantly relied on a ResNet-50 backbone and evidential deep learning for classification and uncertainty estimation. Given that YOLOv8 represents a more recent mainstream vision architecture with superior feature extraction and multi-scale representation capabilities,[27,28] we adopted a YOLOv8 backbone while retaining Dirichlet evidential modeling at the output layer to generate class probabilities and uncertainty estimates. Regarding thresholding, prior approaches typically relied on a single global uncertainty threshold to assess prediction reliability. In real-world screening scenarios, however, the estimation of such a global threshold is often dominated by common and easy-to-classify samples, resulting in an overly permissive threshold and insufficient reliability control for rare and challenging samples. To address this limitation, we proposed a class-adaptive thresholding mechanism, in which class-specific uncertainty thresholds were derived from ROC analysis on the validation set and subsequently applied during testing to assess prediction reliability. This enabled more reliable prediction assessment and enhanced the clinical applicability of the AI system in routine practice. With the UIOS+θ model, most normal and mildly abnormal images fell below the threshold, whereas a higher proportion of severe lesions exceeded it (Table S3) and were therefore referred to clinical review. This demonstrated effective triage capability, reducing unnecessary clinician workload while ensuring that critical cases are not missed.

Heatmap visualizations showed that the UIOS-Y model consistently focused on clinically relevant regions across diverse infantile fundus abnormalities. The strong

ARTICLE IN PRESS

alignment between model attention and established lesion characteristics supports the interpretability of the UIOS-Y model and enhances clinical confidence in its decision-making, thereby facilitating safer integration into real-world practice.

Recently, numerous studies have explored the application of AI chatbots in assisting intelligent diagnosis and management of ocular diseases.[29-31] However, their capability in identifying infantile fundus abnormalities remains largely unexplored. From these comparisons, we found that the UIOS models significantly outperformed Gemini-3.0 Pro and ChatGPT-5.2. The diagnostic performance of AI chatbots was comparable to that of resident ophthalmologists. These findings suggest that, although AI chatbots demonstrate substantial potential as intelligent assistants across diverse domains, their performance remains limited in highly specialized fields, particularly pediatric ophthalmology. This gap highlights the importance of developing task-specific AI systems to identify abnormalities in infants. However, this is only an exploratory analysis based on 110 images, and as AI chatbot capabilities continue to evolve, further validation using larger, real-time datasets is warranted.

WeChat is the most widely used digital platform in China and is extensively integrated in both daily life and professional practice.[32-34] We aim to integrate the AI system into WeChat to substantially enhance accessibility and user convenience, thereby enabling broad adoption across diverse user groups. This approach enhances practical usability and provides researchers a novel and accessible way to apply AI models.[35] In parallel, a

ARTICLE IN PRESS

user-friendly web-based platform is under development to support clinicians and international users who do not routinely use WeChat. This strategy is designed to provide a flexible and inclusive framework for facilitating screening of infantile fundus abnormalities by healthcare professionals.

With the rapid proliferation of AI models, many AI models have been developed and validated on relatively small datasets,[36,37] limiting their performance and real-world applicability. Consequently, large-scale data are essential for robust model development. However, the limited availability of expert clinician poses a substantial bottleneck for high-quality medical image labeling.[38] To address this challenge, we initially trained an AI system using a small-sample dataset of approximately 1,000 images per category , which was subsequently used to pre-classify large-scale data. These pre-labeled data were then reviewed and corrected by clinicians, significantly reducing the time required for manual labeling of large-scale datasets in medical AI applications. Using this workflow, we constructed the largest infantile fundus image dataset to date, comprising 319,998 images from 15,647 infants, and developed our AI system. Leveraging this large-scale dataset substantially enhances model robustness compared with earlier AI models,[39,40] and provides a valuable reference for scalable data labeling.

This study is based on real-world data accumulated over 20 years of routine clinical practice. Unlike AI models trained on short-term datasets, our models capture real-world disease prevalence, variability in image quality and severity, and inter-center

ARTICLE IN PRESS

heterogeneity encountered in daily clinical settings. As a result, the model is better aligned with clinical reality, enhancing robustness, generalizability, and translational potential for clinical deployment, and providing a reliable foundation for sustainable AI-assisted screening and management of infantile fundus diseases.

This study has several limitations that warrant further investigation. First, although the UIOS model achieved accurate identification of multiple fundus abnormalities, 40.0% of samples in the public dataset exhibited uncertainty scores above the threshold. This may stem from variations in imaging devices and populations. Future work will integrate multicenter data from diverse imaging systems and populations to further improve generalizability. Second, the image-level evaluation framework used in this study may not fully capture clinical decision-making at the eye or patient level, underscoring the need for more clinically relevant evaluation strategies. Third,, despite leveraging the largest dataset to date, imbalance of the data distribution across different fundus abnormalities persisted, leading to superior performance for more common diseases. In addition, comparisons with several modern architectures, such as ConvNeXt, ResNeXt, and Transformer-based models, were not included, potentially limiting the interpretation of the architectural contribution of the current framework. Future efforts should focus on optimizing model architectures and improving performance for rare fundus abnormalities to achieve more equitable and clinically robust results. Fourth, the current OOD evaluations were primarily conducted using several representative images and may not

ARTICLE IN PRESS

fully reflect real-world clinical scenarios. In addition, the current system has not been validated in prospective clinical settings. Therefore, further real-world studies are needed to evaluate the robustness, generalizability, and clinical decision-support efficiency of the framework. Fifth, the clinician-AI chatbot comparison is an exploratory analysis based on 110 images. Further validation using larger real-world datasets is warranted to better assess the performance differences. Finally, the WeChat application and web-based platform remain in early stages of preliminary deployment profiling, and further work is needed to establish effective strategies for real-world implementation.

In conclusion, the UIOS model demonstrates robust performance in identifying infantile fundus abnormalities from fundus images while effectively flagging uncertain samples for clinician review, thereby reducing diagnostic errors. Its strong performance supports its potential clinical utility in real-world AI-assisted pediatric ophthalmic screening.

ARTICLE IN PRESS

Methods

Participant details

The study adhered to the principles of the Declaration of Helsinki and was approved by the Institutional Review Board of Shenzhen Eye Hospital (IRB No. 2025KYPJ098). Informed consent was waived due to the retrospective nature of the study. All images were de-identified prior to the analysis.

ARTICLE IN PRESS

Multicenter fundus images collected between 2005 and 2023 were retrieved from the Eyecare-Cloud platform, which we established for large-scale, multimodal data storage.[23] These images were primarily obtained from multiple clinical centers in Guangdong, Guangxi, and Sichuan provinces and were mainly used for AI system development, validation, and internal testing. In addition, fundus images from three independent clinical centers (FJPH, KMCH, and PUPH) and two publicly available datasets (FARFUM-RoP and Ostrava-ROP) were collected for external testing of the AI system.

All images underwent standardized data analysis, including deduplication and label verification. Images with insufficient quality for lesion classification, including those affected by severe artifacts, overexposure or underexposure, and blurred, were excluded. Only images of adequate quality suitable for lesion classification were included in this study. Image screening was conducted using a human-AI collaborative framework.

ARTICLE IN PRESS

Development of human-AI collaborative models

Based on different clinical task requirements, a hierarchical multi-task pretraining framework was established, progressing from image quality control to coarse classification and finally to fine-grained classification. Specifically, a ConvNeXt-based binary classification model was developed to categorize images into acceptable and unacceptable groups. ResNet-50 was adopted as the backbone network to perform successive pretraining tasks, including normal/abnormal binary classification, and multi-

ARTICLE IN PRESS

class classification covering 11 categories.[41] Normal images, which constitute the largest category, were grouped and subsequently underwent rapid double review by two retinal ophthalmologists with over 10 years of experience in pediatric retinal diseases. Inter-observer agreement was assessed on a randomly selected subset of 110 images (10 per category), yielding a Cohen's kappa coefficient of 0.84. During this process, any remaining unacceptable-quality images were excluded, and misclassified abnormal images were corrected accordingly. For images identified as abnormal, a multi-class classification framework was applied to assign them into 11 categories. The resulting labels were jointly reviewed and confirmed by the two ophthalmologists. Discrepancies were resolved through discussion, and an experienced retinal expert served as adjudicator when consensus could not be reached. Images without a definitive label after expert review were excluded. Therefore, clinicians labeled images pre-classified by AI models rather than original data, substantially reducing the manual labeling burden in large-scale datasets.[42]

Finally, a total of 319,998 fundus images from 15,647 infants were included for the development, validation, internal and external testing of the AI models. The dataset comprised 11 categories: normal (Class 01), retinal fold (Class 02), retinoblastoma (Class 03), A-ROP (Class 04), ROP stages 1–2 (Class 05), ROP stage 3 (Class 06), ROP stages 4–5 (Class 07), laser scars (Class 08), Coats' disease (Class 09), chorioretinal coloboma (Class 10), and retinal hemorrhage (Class 11).

ARTICLE IN PRESS

In addition, this study included 6,299 images from 1,978 participants across multiple modalities as OOD samples for AI model validation. FFA images from 73 participants (597 images), fundus images from 32 participants (331 images), OCTA images from 276 participants (588 images), SLO images from 195 participants (2,564 images), fundus images with Clarus 500 device from 1,107 participants (1,107 images), anterior segment images from 263 participants (1,074 images), and 38 RetCam fundus images from 32 infants (38 images) were collected (Table S5).

Development of large-scale AI models

An 11-class classification model for infantile fundus images was developed using the YOLOv8m framework, with a ResNet-50-based model implemented as a benchmark for comparison.[15] All images were split into training, validation, and test sets in a 6: 2: 2 ratio. The YOLOv8m backbone was initialized from a pretrained classification network, with the original head removed and replaced by a lightweight custom classification head comprising an adaptive average pooling layer, a flattening layer, and a fully connected layer with 11 output units to generate the final logits.

All images were resized to 256×256 pixels and normalized using ImageNet channel statistics (mean = [0.485, 0.456, 0.406], standard deviation = [0.229, 0.224, 0.225]). Random horizontal flipping (probability 0.5) was applied during training to increase robustness to viewing-angle variations, whereas validation and test images were

ARTICLE IN PRESS

processed using only resizing and normalization. Training was performed with a batch size of 64 for a total of 50 epochs using a two-stage optimization strategy. In Stage 1 (epochs 1-5), the backbone was frozen and only the classification head was trained using the Adam optimizer with a learning rate of 3×10^{-4} and a weight decay of 1×10^{-4}. In Stage 2 (from epoch 6 onward), the entire network was unfrozen and jointly optimized using a unified learning rate of 1×10^{-4}. Experiments were conducted on a Windows workstation equipped with an NVIDIA GeForce RTX 4060 Ti GPU (CUDA Toolkit 12.6), 32 GB DDR4 RAM (3200 MHz), and SSD storage.

To enable uncertainty estimation, the output logits were further used to construct a Dirichlet-based evidential representation. Specifically, logits were transformed into non-negative evidence values via the softplus function, and the Dirichlet parameters were computed as $\alpha_k = evidence_k + 1$. The total evidence was defined as $S = \sum_{k=1}^{11} \alpha_k$, and the Dirichlet mean was computed as $b_k = \frac{\alpha_k - 1}{S}$. Sample-level uncertainty was quantified as $u = \frac{C}{S}$, where C denotes the number of classes, with larger u indicating higher predictive uncertainty.

The training objective comprised two components: (1) a temperature-annealed cross-entropy loss, where the temperature varies linearly with the epoch index t according to $T = 0.01 + 0.99 \cdot \frac{t}{50}$, thereby sharpening logits in the early stage to accelerate head convergence; (2) a Dirichlet-Evidential regularization term, comprising a Dirichlet-CE loss and a KL-divergence term toward a uniform prior, with its annealing coefficient increasing

ARTICLE IN PRESS

gradually during training. This regularization weight was set to 0 during the warm-up stage and fixed at 0.5 during full-network fine-tuning.

After model training, class-wise adaptive uncertainty thresholds were constructed using the validation set to determine the reliability of predictions. Specifically, the uncertainty value u was calculated for all validation samples, and an error label (err = 0/1) was assigned according to whether the model's prediction was correct. Samples were then grouped by their true class, and for each class, an ROC curve was generated using u as the score and err as the binary indicator. The optimal threshold θ_c was determined by maximizing the Youden index; if a valid ROC could not be formed (e.g., the model predicted all samples correctly), the 75th percentile of that class's u distribution was used as a fallback threshold. This process yielded a set of 11 class-specific thresholds $\theta_0, \theta_1, \ldots, \theta_{10}$. During testing, the threshold corresponding to the predicted class $\hat{y}$ was selected, and predictions with $u < \theta_{\hat{y}}$ were regarded as reliable, whereas those with $u \geq \theta_{\hat{y}}$ were considered unreliable. Key resources used for the development of the ResNet-50, YOLOv8, and UIOS models were summarized in Table S6.

Comparison of UIOS models with AI chatbots

We compared the UIOS models with Gemini-3.0 Pro, ChatGPT-5.2, and three clinicians (one retinal specialist, one attending ophthalmologist, and one resident physician) in identifying infantile fundus abnormalities. Three clinicians independently evaluated 110 images. For Gemini-3.0 Pro and ChatGPT-5.2, the same images were input individually. All comparators performed an 11-category single-label diagnosis task, and no non-image clinical context was provided. AI models were evaluated using the same protocol as in the testing phases.

ARTICLE IN PRESS

Quantification and statistical analysis

Classification performance was evaluated using accuracy, precision, recall, specificity, F1-score, ROC curve, and AUC. The reported average AUC, accuracy, precision, recall, specificity, and F1-score in this study refer to the macro-averaged metrics. All metrics were calculated based on true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).[43]

Statistical analyses were performed using R (version 4.5.1; R Foundation for Statistical Computing, Vienna, Austria). Categorical variables were summarized as counts and percentages. Performance differences among AI models, AI chatbots, and clinicians were assessed using a logistic mixed-effects model, accounting for the paired design. The binary outcome (true/false) was modeled with each comparator as a fixed effect and image ID as a random intercept to account for within-image correlation. For images exceeding the uncertainty threshold (UIOS-Y+θ), the mixed-effects framework was used, which accommodates incomplete data under the missing-at-random assumption. After confirming a significant overall comparator effect, pairwise comparisons were conducted using estimated marginal means derived from the fitted model. The Tukey method was applied to adjust for multiple comparisons. A p value < 0.05 was considered statistically significant.

ARTICLE IN PRESS

Acknowledgments

This work was supported by National Natural Science Foundation of China (82401315, 82301269, 82271103, 82401272), Shenzhen Medical Research Fund (A2403020, C2301005, C2501034), Shenzhen Science and Technology Program (JCYJ20240813152703005, JCYJ20250604184009011), Sanming Project of Medicine in

ARTICLE IN PRESS

Shenzhen (SZSM202311018), Guangdong Basic and Applied Basic Research Foundation (2026A1515012558).

Data availability All data in this study is available from the corresponding author (zhang-guoming@163.com) upon reasonable request.

Code availability Original code in this study is available from the corresponding author (zhang-guoming@163.com) upon reasonable request.

Author contributions G.Z., X.Z., and Z.W. had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Conceptualization, design, and drafting of the manuscript: X.Z., Z.W., G.Z.. Critical review of the manuscript for important intellectual content: K.C., S.C., Y.H., S.W., R.T., Y.C., X.C., N.D., W.W., W.C.. AI model development: M.Z., B.L.. Statistical analysis: S.Z., G.Y.; Multi-center data collection: X.Z., L.J., N.L., J.L.. All authors have read and approved the manuscript.

Conflict of interest The authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the writing process During the preparation of this work, we used ChatGPT to improve language and readability. After using this tool, we reviewed and edited the content as needed and take full responsibility for the content of this publication.

ARTICLE IN PRESS

References:

1 Lim, H. W. *et al.* Causes of Childhood Blindness in the United States Using the IRIS® Registry (Intelligent Research in Sight). *Ophthalmology* **130**, 907-913 (2023).

2 Siu, C. R. & Murphy, K. M. The development of human visual cortex and clinical implications. *Eye Brain* **10**, 25-36 (2018).

3 Chen, W. *et al.* Early detection of visual impairment in young children using a smartphone-based deep learning system. *Nat Med* **29**, 493-503 (2023).

ARTICLE IN PRESS

4 (WHO), W. H. O. https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment, 2023).

5 Wong, E. S. *et al.* Global and Regional Trends in Retinopathy of Prematurity. *JAMA Ophthalmol* (2025).

6 Zhao, X. *et al.* Ultra-wide-field fundus photography and AI-based screening and referral for multiple ocular fundus diseases. *Cell Rep Med* **6**, 102187 (2025).

7 Chen, S. *et al.* FIT-Net: Feature Interaction Transformer Network for Pathologic Myopia Diagnosis. *IEEE Trans Med Imaging* **42**, 2524-2538 (2023).

8 Singh, L. K., Garg, H. & Khanna, M. Histogram of oriented gradients (HOG)-based artificial neural network (ANN) classifier for glaucoma detection. *International Journal of Swarm Intelligence Research (IJSIR)* **13**, 1-32 (2022).

9 Singh, L. K., Khanna, M., Garg, H., Singh, R. & Iqbal, M. A three-stage novel framework for efficient and automatic glaucoma classification from retinal fundus images. *Multimedia Tools and Applications* **83**, 85421-85481 (2024).

10 Singh, L. K., Garg, H., Pooja & Khanna, M. Performance analysis of machine learning techniques for glaucoma detection based on textural and intensity features. *International Journal of Innovative Computing and Applications* **11**, 216-230 (2020).

11 Xie, H. *et al.* Adversarial learning-based multi-level dense-transmission knowledge distillation for AP-ROP detection. *Medical image analysis* **84**, 102725 (2023).

12 Chen, S. *et al.* Dual-Scale Swin Transformer via Feature Alignment and Adversarial Discrimination for Retinopathy of Prematurity Diagnosis. *IEEE J Biomed Health Inform* **29**, 9013-9026 (2025).

13 Liu, Y. *et al.* Automated detection of nine infantile fundus diseases and conditions in retinal images using a deep learning system. *EPMA Journal* **15**, 39-51 (2024).

14 Fei, N. *et al.* Towards artificial general intelligence via a multimodal foundation model. *Nat Commun* **13**, 3094 (2022).

15 Wang, M. *et al.* Uncertainty-inspired open set learning for retinal anomaly identification. *Nat Commun* **14**, 6757 (2023).

16 Peng, Y. *et al.* Enhancing AI reliability: A foundation model with uncertainty estimation for optical coherence tomography-based retinal disease diagnosis. *Cell Rep Med* **6**, 101876 (2025).

17 Zou, K. *et al.* Confidence-aware multi-modality learning for eye disease screening. *Medical image analysis* **96**, 103214 (2024).

18 Luo, Y., Lin, A., Peng, Y., Lin, T. & Chen, H. Foundation Model With Uncertainty Estimation-Based Active Learning for Retinal Image Classification. *IEEE J Biomed Health Inform* **Pp** (2025).

19 Wang, W. *et al.* UA-VLFM: An Uncertainty-aware Vision-Language Foundation Model for Auxiliary Diagnosis of Vitreoretinal lymphoma. *IEEE J Biomed Health Inform* **Pp** (2025).

20 Cui, Y., Deng, W., Chen, H. & Liu, L. Uncertainty-Aware Distillation for Semi-Supervised Few-Shot Class-Incremental Learning. *IEEE Trans Neural Netw Learn Syst* **35**, 14259-14272 (2024).

ARTICLE IN PRESS

21 Huang, A. S. *et al.* ChatGPT-Assisted Glaucoma Diagnosis: A Health-Equitable Multi-Ancestry Analysis Using Visual Field and Optical Coherence Tomography Data. *Am J Ophthalmol* **283**, 129-137 (2025).

22 Shean, R. *et al.* OpenAI o1 Large Language Model Outperforms GPT-4o, Gemini 1.5 Flash, and Human Test Takers on Ophthalmology Board-Style Questions. *Ophthalmol Sci* **5**, 100844 (2025).

23 Zhao, X. *et al.* Eyecare-cloud: an innovative electronic medical record cloud platform for pediatric research and clinical care. *EPMA Journal*, 1-10 (2024).

24 Akbari, M. *et al.* FARFUM-RoP, A dataset for computer-aided detection of Retinopathy of Prematurity. *Sci Data* **11**, 1176 (2024).

25 Timkovič, J. *et al.* Retinal Image Dataset of Infants and Retinopathy of Prematurity. *Sci Data* **11**, 814 (2024).

26 Singh, L. K., Khanna, M., Garg, H. & Singh, R. A novel soft computing based efficient feature selection approach for timely identification of COVID-19 infection using chest computed tomography images: a human centered intelligent clinical decision support system. *Multimedia Tools and Applications* **84**, 13717-13785 (2025).

27 Varghese, R. & Sambath, M. in *2024 International conference on advances in data engineering and intelligent computing systems (ADICS).* 1-6 (IEEE).

28 Zhang, B. *et al.* An Improved Microaneurysm Detection Model Based on SwinIR and YOLOv8. *Bioengineering (Basel)* **10** (2023).

29 Soh, Z. D. *et al.* An integrated language-vision foundation model for conversational diagnostics and triaging in primary eye care. *Cell Rep Med* **6**, 102476 (2025).

30 Luo, M. J. *et al.* A large language model digital patient system enhances ophthalmology history taking skills. *NPJ Digit Med* **8**, 502 (2025).

31 Luo, M. J. *et al.* Development and Evaluation of a Retrieval-Augmented Large Language Model Framework for Ophthalmology. *JAMA Ophthalmol* **142**, 798-805 (2024).

32 Chen, X., Zhou, X., Li, H., Li, J. & Jiang, H. The value of WeChat application in chronic diseases management in China. *Comput Methods Programs Biomed* **196**, 105710 (2020).

33 Zhou, Y. *et al.* WeChat usage and preservation of cognitive functions in middle-aged and older Chinese adults: indications from a nationally representative survey, 2018-2020. *BMC Public Health* **24**, 1783 (2024).

34 Fang, F., Bu, B. & Jiang, W. Combining Problem-Based Learning Methods With the WeChat Platform in Teaching Ophthalmology: Randomized Controlled Trial. *JMIR Med Educ* **12**, e65279 (2026).

35 Hui, S. *et al.* Deep learning-based mobile application for efficient eyelid tumor recognition in clinical images. *NPJ Digit Med* **8**, 185 (2025).

36 Liu, Y. *et al.* An artificial intelligence system for screening and recommending treatment modalities for retinopathy of prematurity. *Asia-Pacific journal of ophthalmology (Philadelphia, Pa.)* **12**, 468-476 (2023).

ARTICLE IN PRESS

ARTICLE IN PRESS

37 Huang, Y. P. *et al.* Automated detection of early-stage ROP using a deep convolutional neural network. *Br J Ophthalmol* **105**, 1099-1103 (2021).

38 Korot, E. *et al.* Clinician-Driven AI: Code-Free Self-Training on Public Data for Diabetic Retinopathy Referral. *JAMA Ophthalmol* **141**, 1029-1036 (2023).

39 Li, L. *et al.* DeepQuality improves infant retinopathy screening. *NPJ Digit Med* **6**, 192 (2023).

40 Chen, S. *et al.* SABPI-Net: A Structure-aware Bidirectional Proxy Interaction Network for Infantile Retinal Disease Diagnosis. *IEEE Trans Med Imaging* **Pp** (2026).

41 Wu, Z., Xiao, M., Fang, C. & Lin, Z. Designing Universally-Approximating Deep Neural Networks: A First-Order Optimization Approach. *IEEE Trans Pattern Anal Mach Intell* **46**, 6231-6246 (2024).

42 Chen, H. *et al.* A human-LLM collaborative annotation approach for screening articles on precision oncology randomized controlled trials. *BMC Med Res Methodol* **25**, 219 (2025).

43 Zhao, X. *et al.* An artificial intelligence system for the whole process from diagnosis to treatment suggestion of ischemic retinal diseases. *Cell Rep Med* **4**, 101197 (2023).

ARTICLE IN PRESS

Figure legends:

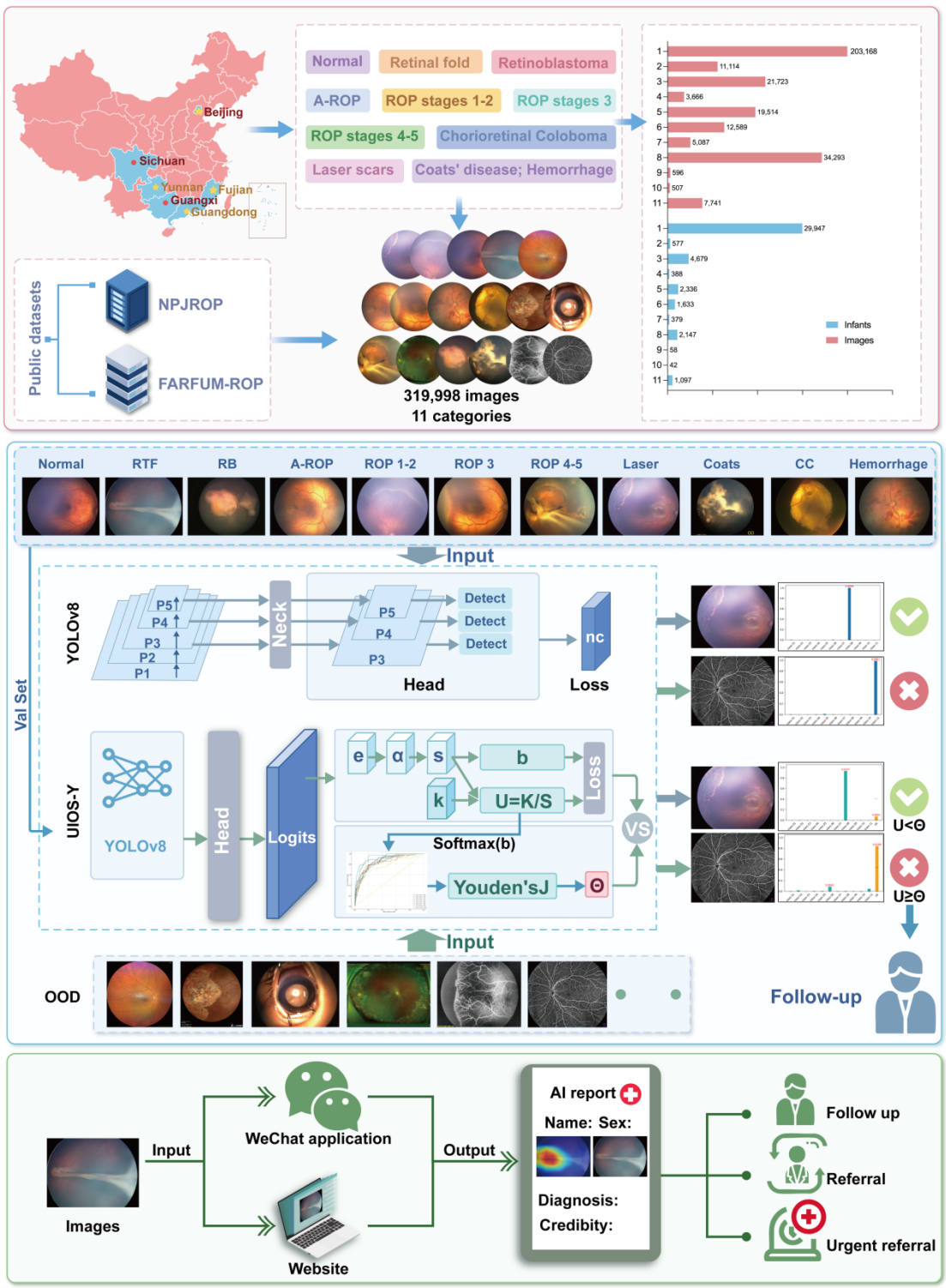
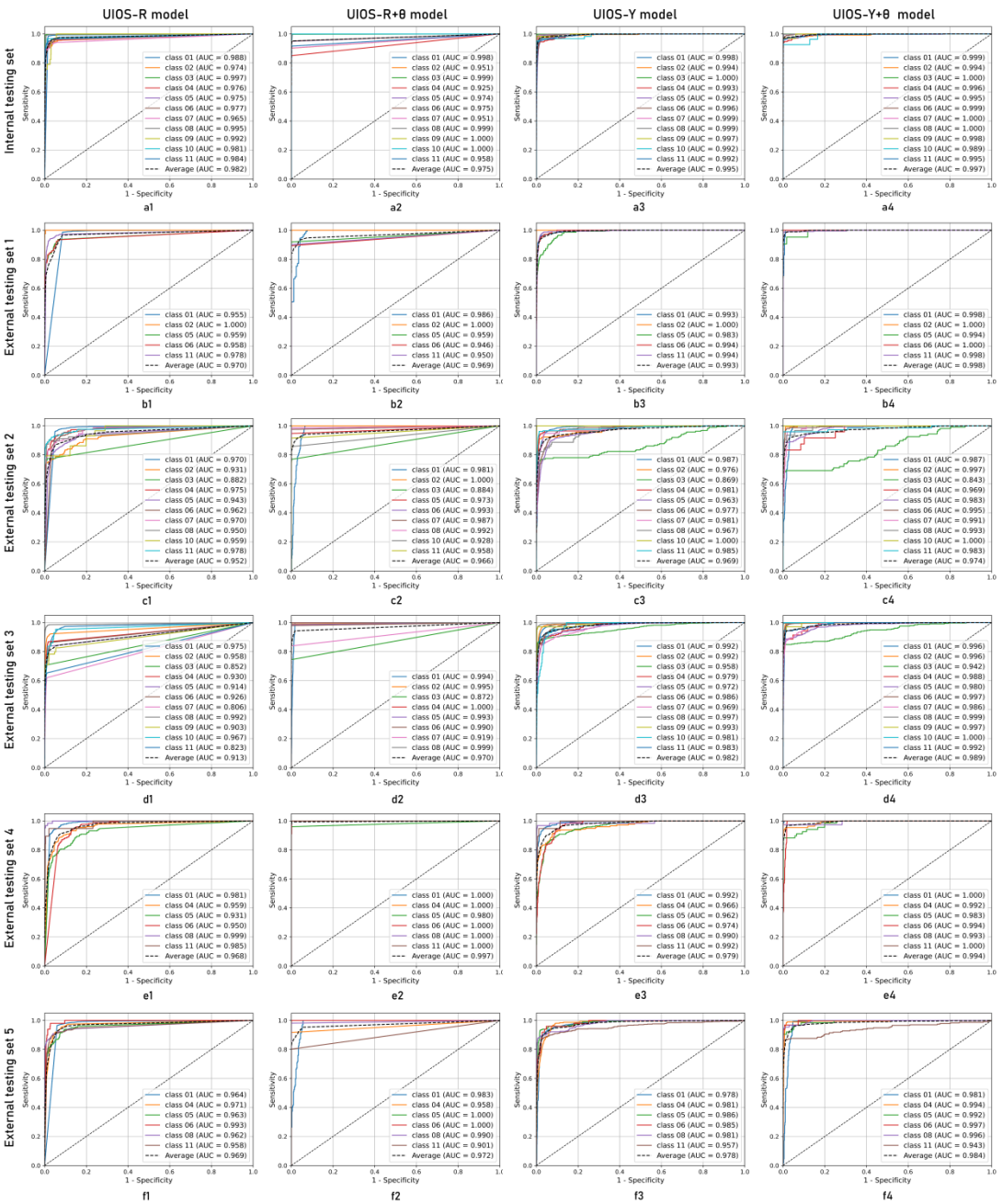


Fig. 1. Overview of the AI framework for fundus abnormality identification. AI, artificial intelligence; retinopathy of prematurity, ROP. Class 1, Normal; Class 2, Retinal fold; Class

3, Retinoblastoma; Class 4, A-ROP; Class 5, ROP stages 1-2; Class 6, ROP stage 3; Class 7, ROP stages 4-5; Class 8, Laser scars; Class 9, Coats' disease; Class 10, Chorioretinal coloboma; and Class 11, Hemorrhage.



ARTICLE IN PRESS

Fig. 2. The ROC curves of UIOS-Y and UIOS-R models on internal and external datasets. AUC, area under the receiver operating characteristic curve; ROC, receiver operating characteristic; UIOS, uncertainty-inspired open-set.

ARTICLE IN PRESS

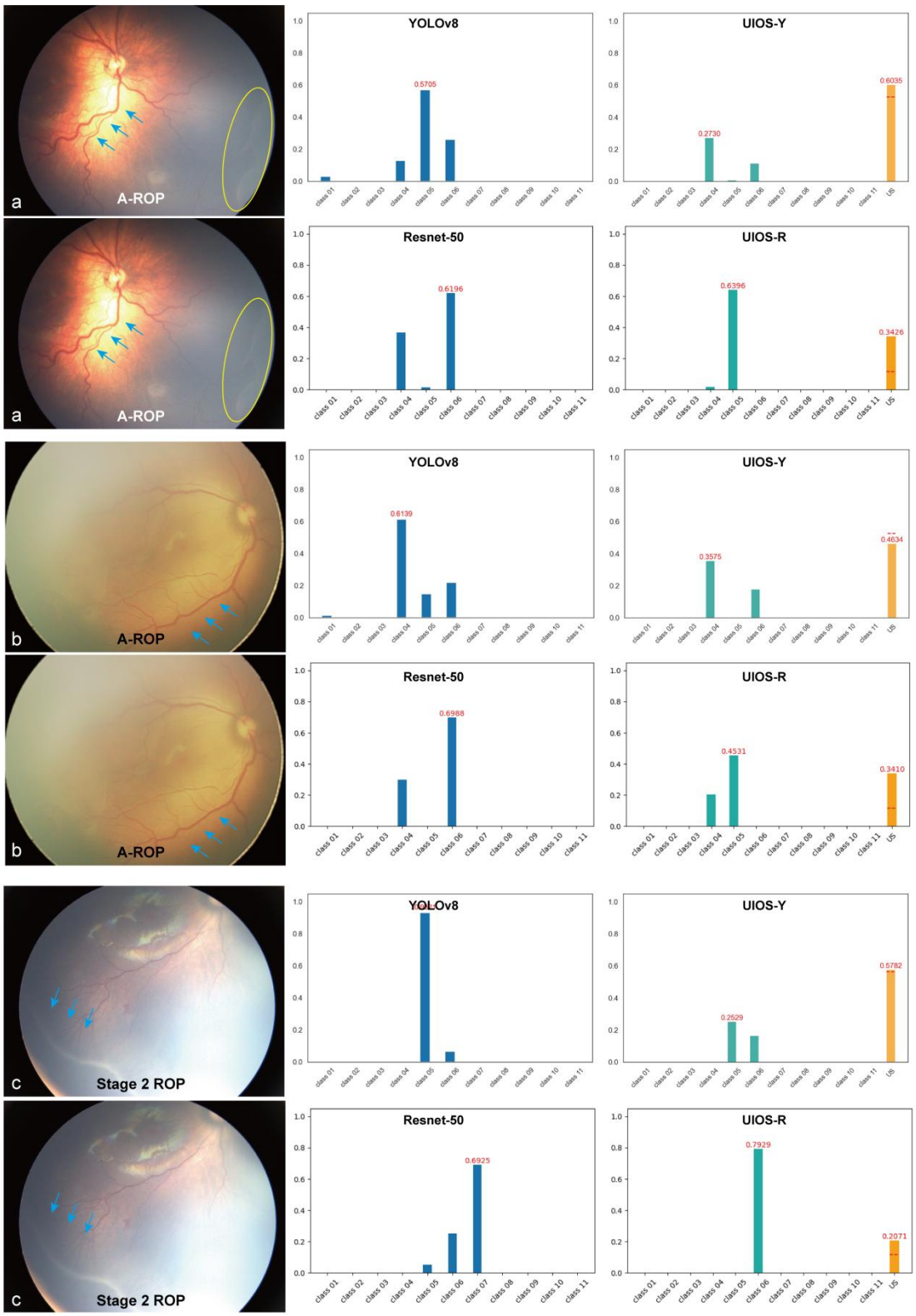


Fig. 3. Representative cases of image classification by UIOS-Y and UIOS-R. Panels (a-c) show three cases where UIOS-Y provided correct diagnoses, while UIOS-R

misclassified the images. These inconsistencies arise from multiple coexisting features within the same images. In panel a, the blue arrow indicates A-ROP characteristics, while the yellow circle shows stage 2-3 ROP characteristics. In panel b, the blue arrow marks A-ROP features alongside stage 2 ROP changes in the temporal retina. In panel c, the blue arrow points to severe plus disease, which may lead to a stage 3 ROP classification.

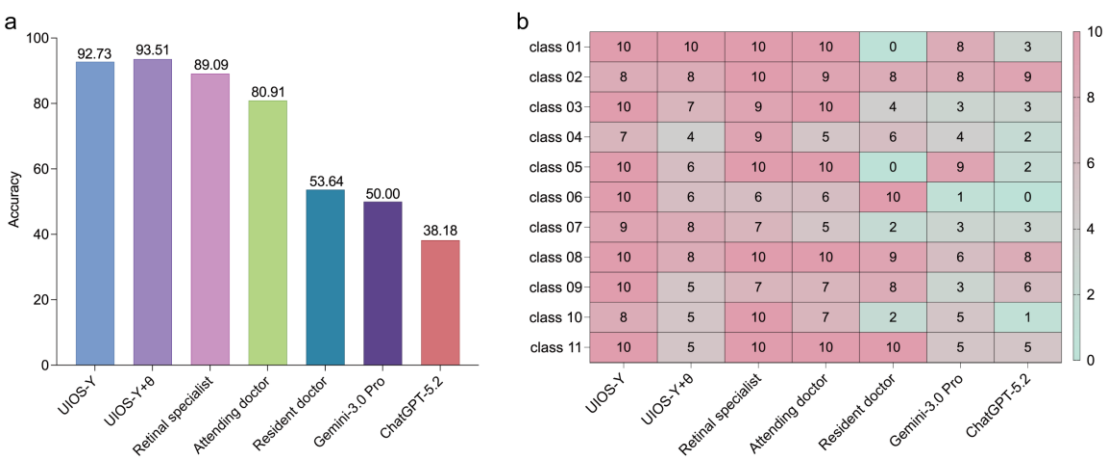


Fig. 4. Comparisons of classification accuracy among UIOS-Y based models, AI chatbots, and clinicians. (a) Overall classification accuracy (%) across different diagnostic agents, including UIOS-Y, UIOS-Y+θ, retinal specialist, attending doctor, resident doctor, Gemini-3.0 Pro, and ChatGPT-5.2. (b) Correct prediction counts for each diagnostic agent across 11 disease categories. Each cell represents the number of correctly classified cases within a given class (maximum=10 per class), illustrating the consistency and robustness of the UIOS-Y models. Class 01, Normal; Class 02, Retinal fold; Class 03, Retinoblastoma; Class 04, A-ROP; Class 05, ROP stages 1-2; Class 06, ROP stage 3; Class 07, ROP stages 4-5; Class 08, Laser scars; Class 09, Coats' disease; Class 10, Chorioretinal coloboma; and Class 11, Hemorrhage.

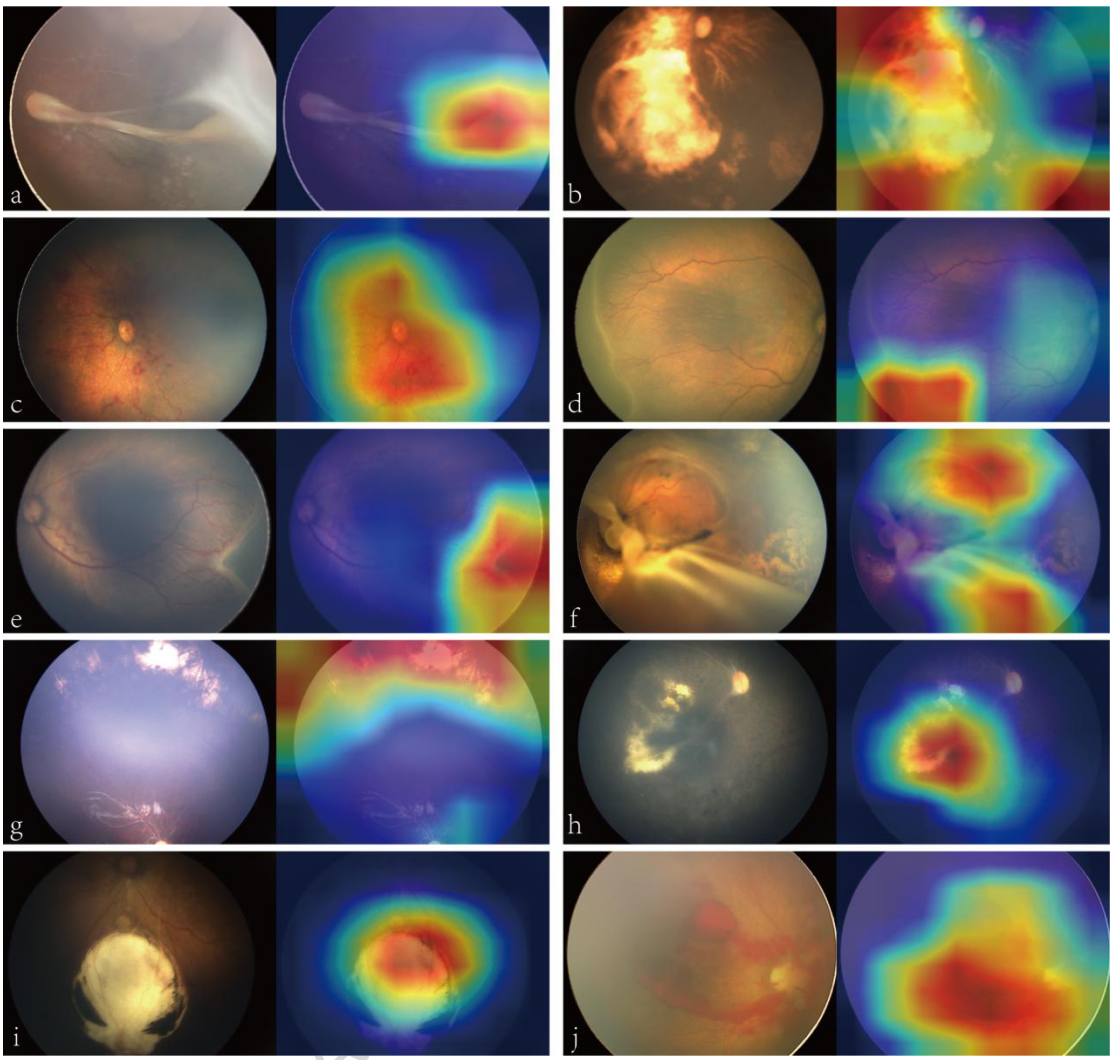


Fig. 5. Heatmaps for demonstration of typical features for each category. a. Retinal fold; b. Retinoblastoma; c. A-ROP; d. ROP stages 1-2; e. ROP stage 3; f. ROP stages 5; g. Laser scars; h. Coats' disease; i. Chorioretinal coloboma; j. Hemorrhage.

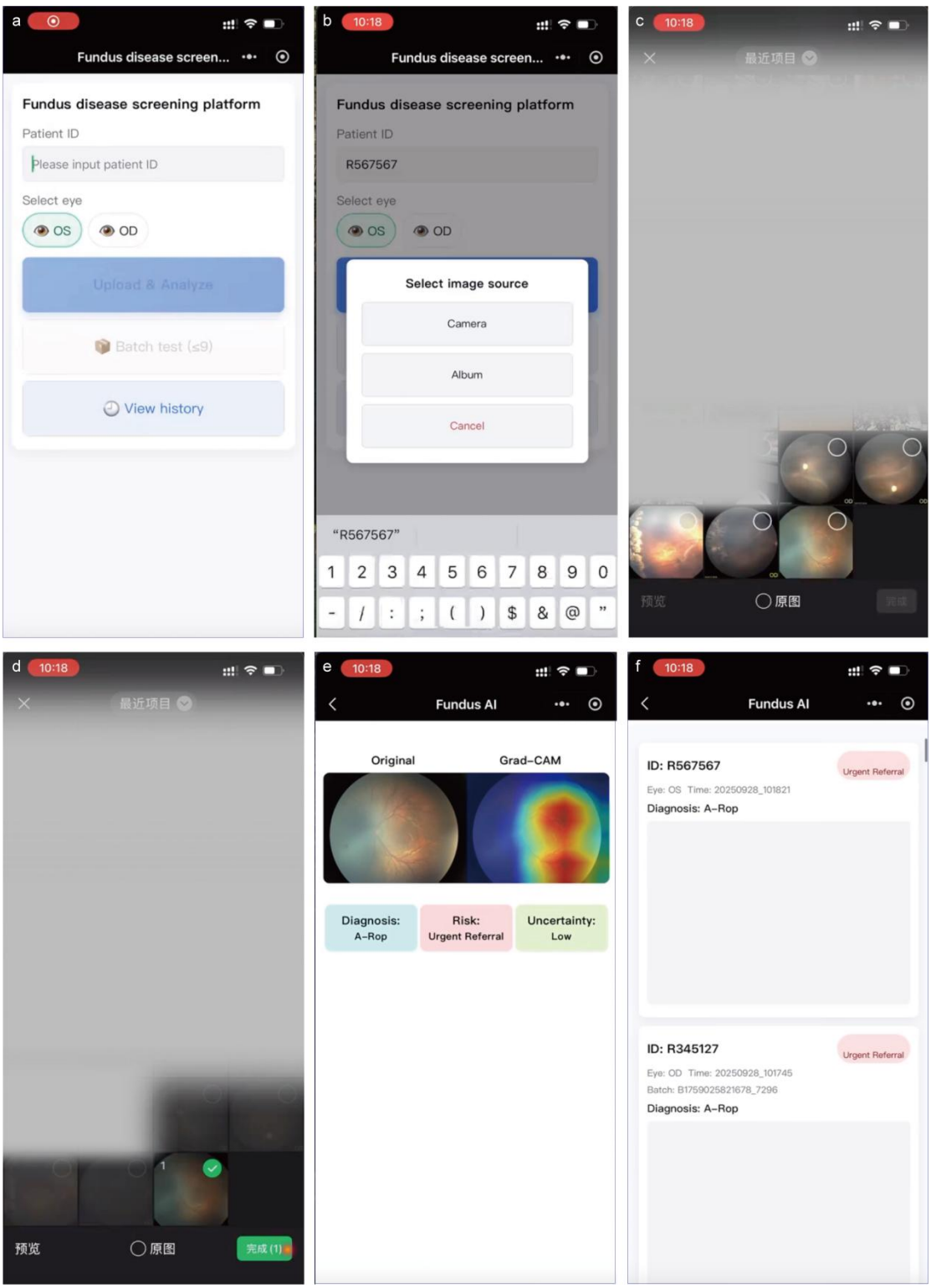


Fig. 6. Overview of the "Fundus Disease Screening Platform" WeChat Mini Program. The program features a straightforward screening workflow and a user-friendly interface, and

ARTICLE IN PRESS

provides outputs including diagnostic results, disease severity for referral decision-making, and uncertainty estimates. a: Login interface, b-d: Select images, e: Automatic output of diagnostic results, f: The program automatically records it in the history to facilitate follow-up.

ARTICLE IN PRESS