nature medicine

Article

https://doi.org/10.1038/s41591-022-02180-9

Early detection of visual impairment in young children using a smartphone-based deep learning system

Received: 16 June 2022

Accepted: 9 December 2022

Published online: 26 January 2023

Check for updates

A full list of authors and their affiliations appears at the end of the paper.

Early detection of visual impairment is crucial but is frequently missed in young children, who are capable of only limited cooperation with standard vision tests. Although certain features of visually impaired children, such as facial appearance and ocular movements, can assist ophthalmic practice, applying these features to real-world screening remains challenging. Here, we present a mobile health (mHealth) system, the smartphone-based Apollo Infant Sight (AIS), which identifies visually impaired children with any of 16 ophthalmic disorders by recording and analyzing their gazing behaviors and facial features under visual stimuli. Videos from 3,652 children (≤48 months in age; 54.5% boys) were prospectively collected to develop and validate this system. For detecting visual impairment, AIS achieved an area under the receiver operating curve (AUC) of 0.940 in an internal validation set and an AUC of 0.843 in an external validation set collected in multiple ophthalmology clinics across China. In a further test of AIS for at-home implementation by untrained parents or caregivers using their smartphones, the system was able to adapt to different testing conditions and achieved an AUC of 0.859. This mHealth system has the potential to be used by healthcare professionals, parents and caregivers for identifying young children with visual impairment across a wide range of ophthalmic disorders.

Visual impairment is one of the most important causes of long-term disability in children worldwide and has a detrimental impact on education and socioeconomic achievements[1,2]. Infancy and toddlerhood (early childhood) are critical periods for visual development[3], during which early detection and prompt treatment of ocular pathology can prevent irreversible visual loss[4,5]. Young children are unable to complain of visual difficulties, and since they are unwilling or find it difficult to cooperate with standard vision tests (for example, optotype tests), age-appropriate tests such as grating acuity cards are commonly used to observe their reactions to visual stimuli[6,7]. However, evaluating the vision of young children using these tests requires highly trained operators, which greatly hinders their wider adoption, especially in low-income and middle-income countries with the highest prevalence of visual impairment but poor medical resources[8]. In addition, these tests, even when performed by experienced pediatric ophthalmologists, have been shown to have low repeatability in large-scale population screening studies[9–11]. Therefore, it is imperative to develop an easy-to-use and effective detection tool to enable the timely diagnosis of visual impairment in young children and prompt intervention.

Ocular abnormalities causing visual impairment in children often manifest with typical phenotypic features, such as leukocoria (white eye) in cataract[12] and retinoblastoma[13], eyelid drooping in congenital ptosis[14], and a cloudy and enlarged cornea in congenital glaucoma[15]. In addition, previous studies have found that dynamic aberrant behavioral features such as abnormal ocular movement, fixation patterns or visual preference can also point toward an underlying ocular pathology in children[16,17]. These phenotypic manifestations are frequently seen in ocular diseases, such as amblyopia and strabismus, and they

e-mail: dingxiaowei@sjtu.edu.cn; linht5@mail.sysu.edu.cn

can provide valuable clues for diagnosing visual impairment in young children[18–20]. However, systematically recording and applying these features to real ophthalmic practice are still in their infancy due to the lack of practical and effective tools.

Given the rapid development of mobile health (mHealth) and artificial intelligence (AI) algorithms in identifying or monitoring disease states[21,22], the use of mobile devices, such as smartphones, to record and analyze phenotypic features to help identify visual impairment in young children presents great opportunities. However, developing such a system for large-scale ophthalmic application is hindered by three main challenges: (1) collecting phenotypic data that reliably reflect the visual status of the children in complex environments, (2) generalizing the system for large-scale applications and (3) providing evidence of its feasibility. The major bottleneck that impedes the widespread adoption of many medical AI systems is the limited feasibility and reliability when applied to settings with various data distributions in the real world[23,24]. A lack of cooperation is very common in pediatric ophthalmic practice, with constant head movement during examinations introducing test noise that poses several challenges to the stability of the system[25]. For the nascent technology of mHealth, rigorous evidence of clinical application is necessary but generally lacking[21]. These major difficulties explain the current lack of an effective and practical tool for detecting visual impairment in young children.

In this prospective, multicenter, observational study, we developed and validated a smartphone-based system, the Apollo Infant Sight (AIS), to identify visual impairment in young children in real-world settings. AIS was designed to induce a steady gaze in children by using cartoon-like video stimuli and collect videos that capture phenotypic features (facial appearance and ocular movements) for further analysis using deep learning (DL) models with robust quality control design against test noises. We collected more than 25,000,000 frames of videos from 3,652 children using AIS for DL model training and testing. We evaluated the system for detecting visual impairment caused by any of 16 ophthalmic disorders in five clinics at different institutions. Furthermore, we validated this system under different conditions with various test noise levels or ambient interference presented in real-world settings. We also evaluated AIS used by untrained parents or caregivers at home to test its wider applicability. This preliminary study indicates that AIS shows potential for early detection of visual impairment in young children in both clinical and community settings.

Results

Overview of the study

We conducted this prospective, multicenter and observational study (identifier: NCT04237350) in three stages from 14 January 2020 to 30 January 2022 and collected a total of 3,865 videos with 25,972,800 frames of images from 3,652 Chinese children (aged ≤48 months) to develop and validate the AIS system in clinical and at-home settings (Fig. 1). The AIS system was developed and comprehensively tested (internal validation and reliability analyses under different testing conditions) at the clinic of Zhongshan Ophthalmic Center (ZOC) in the first stage, and was further tested in four other centers (external validation) and community settings (at-home implementation) in the second and third stages, respectively.

Development of the mHealth AIS system

We developed AIS for detecting visual impairment in young children tailored to the present study (Fig. 1a and Supplementary Video 1). A child-friendly app was designed to attract children to maintain their gaze using cartoon-like stimuli (Extended Data Fig. 1). The inbuilt front camera of the smartphone recorded 3.5-min videos that captured phenotypic features of the facial appearance and ocular movements during gazing. In this process, the mHealth app interactively guided users (healthcare professionals, volunteers, parents and caregivers) to familiarize themselves with the system and complete standardized preparations, including choosing and maintaining a suitable testing setting (Extended Data Fig. 2). After data collection was completed, DL models were applied to analyze the collected features and identify visually impaired children. To ensure the system's performance in chaotic settings (environments with various interference factors or biases that can impact the system's performance), a series of algorithm-based quality checking operations, including face detection (max-margin object detection (MMOD) convolutional neural network (CNN)); facial key point localization (ensemble of regression trees); and crying, occlusion and interference factor detections (the EfficientNet-B2 backbone shown in Extended Data Fig. 3a,b), was first automatically performed by a quality control module to extract consecutive frames of high quality from the original video as short clips. Facial areas were cropped out to further eliminate environmental interference before the qualified clips were sent to a DL-based detection model for identifying visually impaired children and a diagnostic model for discriminating multiple ocular disorders (the EfficientNet-B4 backbone shown in Extended Data Fig. 3c). The final results were returned to the mHealth app to alert users to promptly refer children at high risk of visual impairment to experienced pediatric ophthalmologists for timely diagnosis and intervention.

We first developed the data quality control module. Two facial detection and key point localization models were pretrained on publicly available datasets and adopted from an open-source library[26]. Additionally, we developed three CNNs for crying, interference and occlusion detection using images sampled from raw videos collected at the ZOC clinic (Extended Data Fig. 3d and Supplementary Table 1). Then, we trained and validated the detection/diagnostic models on the development dataset collected by trained volunteers using iPhone-7/8 smartphones at the clinic of ZOC (Extended Data Fig. 3e). A total of 2,632 raw videos from 2,632 children were collected, and after automatic quality control, videos of 2,344 children (89.1%) were reserved as the development dataset (Fig. 1b), including 871 (37.2%) for children in the 'nonimpairment' group, 861 (36.7%) in the 'mild impairment' group and 612 (26.1%) in the 'severe impairment' group. Detailed information on the qualified dataset is provided in Table 1. Before model training, the development dataset was randomly split into training, tuning and validation sets stratified on sex, age and the ophthalmic condition (Supplementary Table 2). The videos utilized for quality control module development were excluded from the detection/diagnostic model validation.

Performance of the detection model in real clinical settings with trained volunteers

The detection model was trained to discriminate visually impaired children from nonimpaired children based on the high-quality clips extracted from the phenotypic videos. At the clip level, the detection model achieved an area under the receiver operating curve (AUC) of 0.925 (95% confidence interval (95% CI), 0.914–0.936) in the internal validation (Extended Data Fig. 4a). Furthermore, we evaluated the performance of the detection model via an independent external validation performed by trained volunteers using iPhone-7/iPhone-8 smartphones at the routine clinics of four other centers. In this stage, quality checking was embedded in the data acquisition process, and the quality control module automatically reminded volunteers to recollect data when the videos were of low quality (Fig. 1b). Qualified videos for 298 children undergoing ophthalmic examinations were utilized for final validation, including 188 (63.1%) nonimpaired children, 67 (22.5%) mildly impaired children and 43 (14.4%) severely impaired children (Table 1). At the clip level, the detection model achieved an AUC of 0.814 (95% CI, 0.790–0.838) in the external validation (Extended Data Fig. 4b).

The performance of the detection model to identify visually impaired children was evaluated by averaging the clip-level predictions. Figure 2a shows distinguished clip predicted probability patterns for children with various visual conditions. At the child level, the

https://doi.org/10.1038/s41591-022-02180-9

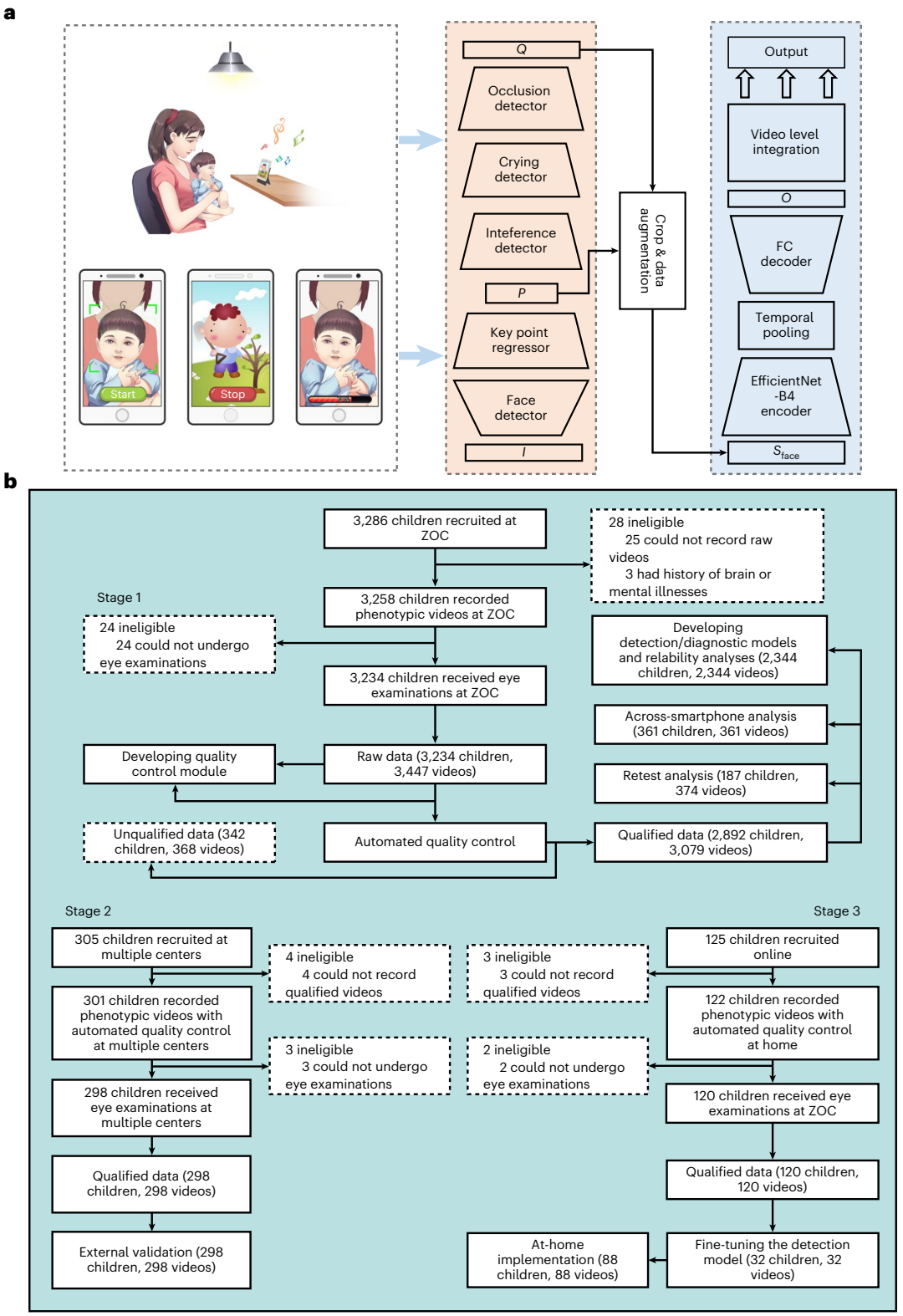


Fig. 1 | Overall study design and participant flow diagram. a, Workflow of the system. The smartphone-based AIS system consists of two key components: an app for user education, testing preparation and data collection and a DL-based back end for data analysis. Parents or other users utilize the app to induce children to gaze at the smartphone, allowing the app to record their phenotypic states as video data. Then, the phenotypic videos are sent to a quality control module to discard low-quality frames. After automatic quality checking, multiple sets of consecutive qualified frames are extracted from the original video as clips, and the child's facial regions are cropped from the clips to serve as candidate inputs to the detection/diagnostic models. A small rectangle indicates input or output data, a large rectangle indicates mathematical operation, and a trapezoid indicates DL or machine learning algorithm. **b**, Participant flow diagram. Children were recruited at multiple clinics to develop and comprehensively test the AIS system in stage 1 and stage 2. Children were recruited online to perform an at-home validation by untrained parents or caregivers in stage 3. *I*, input video; *O*, clip-level model outputs; *P*, key point coordinates; *Q*, qualified clips; S_{face}, facial regions of the clips; FC, fully connected.

https://doi.org/10.1038/s41591-022-02180-9

Table 1 | Summary of the qualified datasets used in this study

	Dataset A (*n*=2,344 children, 2,344 videos)	Dataset B (*n*=187 children, 374 videos)	Dataset C (*n*=361 children, 361 videos)	Dataset D (*n*=298 children, 298 videos)	Dataset E (*n*=32 children, 32 videos)	Dataset F (*n*=88 children, 88 videos)
Sources	ZOC clinic	ZOC clinic	ZOC clinic	Clinics of multiple hospitals	At-home environment	At-home environment
Usage of dataset	Model development and reliability analyses	Retest analysis	Across-smartphone analysis	External validation	Model fine-tuning	At-home validation
Images, *n*	15,751,680	2,513,280	2,425,920	2,002,560	215,040	591,360
Visual conditions, *n* (%)						
Nonimpairment	871 (37.2%)	102 (54.5%)	87 (24.1%)	188 (63.1%)	10 (31.3%)	31 (35.2%)
Mild impairment	861 (36.7%)	52 (27.8%)	169 (46.8%)	67 (22.5%)	14 (43.8%)	31 (35.2%)
Severe impairment	612 (26.1%)	33 (17.6%)	105 (29.1%)	43 (14.4%)	8 (25.0%)	26 (29.5%)
Age of months (mean±s.d.)	25.2±11.7	28.7±10.9	25.7±10.8	28.0±13.0	29.5±10.1	30.0±10.9
Sex, *n* (%)						
Boys	1,265 (54.0%)	107 (57.2%)	202 (56.0%)	169 (56.7%)	16 (50.0%)	50 (56.8%)
Girls	1,079 (46.0%)	80 (42.8%)	159 (44.0%)	129 (43.3%)	16 (50.0%)	38 (43.2%)
Room illumination (mean±s.d.)	289.5±130.1	280.5±122.5*	334.0±117.5	N/A	N/A	N/A
Testing distance, *n* (%)						
Short	195 (8.3%)	45 (12.0)*	34 (9.4%)	61 (20.5%)	6 (18.8%)	19 (21.6%)
Medium	1,738 (74.2%)	291 (77.8%)*	279 (77.3%)	125 (42.0%)	9 (28.1%)	51 (58.0%)
Long	411 (17.5%)	38 (10.2%)*	48 (13.3%)	112 (37.6%)	17 (53.1%)	18 (20.5%)
Laterality of the eye disorder, *n* (%)						
Bilateral	995 (67.6%)	49 (57.7%)	209 (76.3%)	77 (70.0%)	8 (36.4%)	29 (50.9%)
Unilateral	478 (32.5%)	36 (42.4%)	65 (23.7%)	33 (30.0%)	14 (63.6%)	28 (49.1%)
Smartphones used	iPhone-7/iPhone-8	iPhone-7/iPhone-8	Huawei Honor-6 Plus/ Redmi Note-7	iPhone-7/iPhone-8	Parents' own smartphones (no restriction)	

*Metrics calculated in the unit of video. Except for the asterisk-marked metrics in dataset B, metrics were calculated in the unit of child. ZOC, Zhongshan Ophthalmic Center; N/A, not applicable.

detection model achieved an AUC of 0.940 (95% CI, 0.920–0.959), an accuracy of 86.5% (95% CI, 83.4%–89.0%), a sensitivity of 84.1% (95% CI, 80.2%–87.4%) and a specificity of 91.9% (95% CI, 86.9%–95.1%) in the internal validation (Fig. 2b and Supplementary Table 3). It achieved a child-level AUC of 0.843 (95% CI, 0.794–0.893), an accuracy of 82.6% (95% CI, 77.8%–86.4%), a sensitivity of 80.9% (95% CI, 72.6%–87.2%) and a specificity of 83.5% (95% CI, 77.6%–88.1%) in the external validation (Fig. 2c and Supplementary Table 3).

Furthermore, we investigated whether our system could identify visual impairment with any of 16 common ophthalmic disorders at the child level (Table 2 and Supplementary Table 4). For different ophthalmic disorders, the predicted probabilities of the detection model were all significantly higher than those for nonimpairment (Fig. 2d). AIS achieved AUCs of over 0.800 in 15 of 16 binary classification tasks to distinguish visual impairment with various causes from nonimpairment (Fig. 2e,f and Supplementary Table 5), except for limbal dermoid with an AUC of 0.747 (95% CI, 0.646–0.849). Even for diseases not present in the training set, our system showed effective discriminative capabilities, revealing wider extendibility and generalizability to other conditions (Fig. 2f). In addition, we initially recruited children with aphakia (including iatrogenic aphakia cases with common features of visual impairment, accounting for 10.2% of the visually impaired participants enrolled) to increase diversity of training samples for the robustness of the system. Therefore, to evaluate the performance of AIS in the natural population without iatrogenic cases or cases with medical interventions, the children with aphakia were removed from the validation datasets for further analysis and AIS remained reliable (Supplementary Table 6). These results indicate the advanced classification of AIS to detect common causes of visual impairment in young children.

Additionally, the performance of AIS in discriminating mild or severe impairment from nonimpairment was assessed at the child level (Fig. 2g–j and Supplementary Table 3). Significantly lower predicted probabilities of AIS were obtained for the nonimpaired group than for the mild or severe impairment groups. For discriminating mild impairment from nonimpairment, an AUC of 0.936 (95% CI, 0.912–0.960) and an AUC of 0.833 (95% CI, 0.774–0.892) were obtained for the internal validation and the external validation, respectively. For discriminating severe impairment from nonimpairment, an AUC of 0.944 (95% CI, 0.919–0.969) and an AUC of 0.859 (95% CI, 0.779–0.939) were obtained for the internal validation and the external validation, respectively.

To further evaluate the performance of AIS when applied to a population with a rare-case prevalence of visual impairment, we conducted a 'finding a needle in a haystack' test based on the internal validation dataset, with the simulated prevalences ranging from 0.1% to 9%. AIS successfully identified visually impaired children at different simulated prevalences, with AUCs stabilized around 0.940 (Supplementary Table 7).

Performance of the detection model in at-home settings with untrained parents or caregivers

After validation in real clinical settings, we further implemented a more challenging application in at-home settings by parents or caregivers using their smartphones according to the system's instructions (Fig. 1b). Of the 125 children recruited online from the Guangdong area, 122 children (97.6%) successfully completed qualified video

https://doi.org/10.1038/s41591-022-02180-9

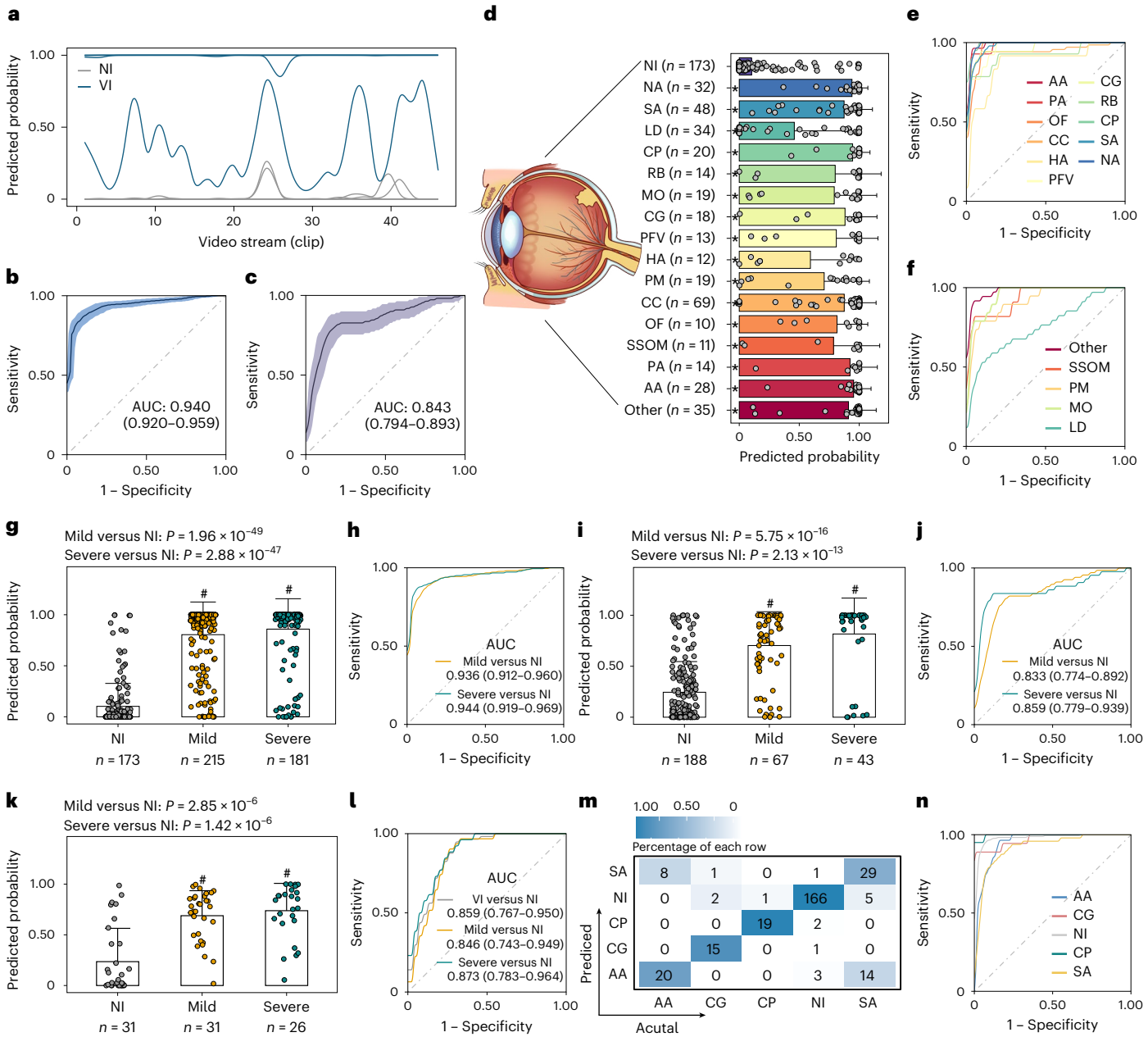


Fig. 2 | Performance of the AIS system in clinical and at-home settings. **a**, Typical predicted probability patterns of the detection model. **b**,**c**, Receiver operating characteristic (ROC) curves of the detection model for distinguishing visually impaired children from nonimpaired children in the internal validation set (**b**) and in the external validation set (**c**). Center lines show ROC curves and shaded areas show 95% CIs. **d**, The predicted probabilities of children with the indicated ophthalmic disorders and nonimpaired children in the internal validation set. Results are expressed as mean ± s.d. $^{*}P < 0.001$ (ranging from 4.83×10^{-27} for congenital cataract (CC) to 2.40×10^{-5} for high ametropia (HA) compared with nonimpairment (NI), two-tailed Mann–Whitney *U*-tests). **e**,**f**, ROC curves of the detection model for distinguishing nonimpaired children from children with the indicated ophthalmic disorders that overlap (**e**) or did not overlap (**f**) with those in the training set (AUCs range from 0.747 for limbal dermoid (LD) to 0.989 for congenital ptosis (CP)). **g**,**i**,**k**, The predicted probabilities of the detection model for the nonimpaired, mildly impaired and severely impaired groups in the internal validation set (**g**), in the external validation set (**i**) and in the at-home implementation (**k**). Results are expressed as mean ± s.d. $^{\#}P < 0.001$, two-tailed Mann–Whitney *U*-tests. **h**,**j**, ROC curves of the detection model for distinguishing mildly or severely impaired children from nonimpaired children in the internal validation set (**h**) and in the external validation set (**j**). **l**, ROC curves of the detection model for distinguishing impaired, mildly impaired or severely impaired children from nonimpaired children in the at-home implementation. **m**, The confusion matrix of the diagnostic model. **n**, ROC curves of the diagnostic model for discriminating each category of ophthalmic disorder from the other categories (aphakia (AA), AUC = 0.947 (0.918–0.976); congenital glaucoma (CG), AUC = 0.968 (0.923–1.000); NI, AUC = 0.976 (0.959–0.993); CP, AUC = 0.996 (0.989–1.000); strabismus (SA), AUC = 0.918 (0.875–0.961)). 95% DeLong CIs are shown for AUC values. MO, microphthalmia; NA, nystagmus; OF, other fundus diseases; PA, Peters' anomaly; PFV, persistent fetal vasculature; PM, pupillary membrane; RB, retinoblastoma; SSOM, systemic syndromes with ocular manifestations; VI, visual impairment.

collection, among whom 120 children undergoing ophthalmic examinations were enrolled. Other detailed information on the qualified data is summarized in Table 1. Given the great difference in data distributions for the home environments compared with the clinics, we fine-tuned the detection model using qualified videos from 32 children and then tested it by the subsequently collected validation set from another

https://doi.org/10.1038/s41591-022-02180-9

Table 2 | Summary of the ophthalmic conditions of participants in this study

	Dataset A (*n*=2,344 children)	Dataset B (*n*=187 children)	Dataset C (*n*=361 children)	Dataset D (*n*=298 children)	Dataset E (*n*=32 children)	Dataset F (*n*=88 children)
Ophthalmic conditions, *n* (%)						
Nonimpairment	871 (37.2%)	102 (54.5%)	87 (24.1%)	188 (63.1%)	10 (31.3%)	31 (35.2%)
Aphakia	153 (6.5%)	7 (3.7%)	44 (12.2%)	4 (1.3%)	6 (18.8%)	6 (10.5%)
Congenital cataract	348 (14.8%)	7 (3.7%)	133 (36.8%)	28 (9.4%)	10 (31.3%)	30 (52.6%)
Congenital glaucoma	95 (4.1%)	5 (2.7%)	0	2 (0.7%)	0	1 (1.8%)
High ametropia	69 (2.9%)	7 (3.7%)	2 (0.6%)	10 (3.4%)	2 (6.3%)	3 (5.3%)
Peters' anomaly	39 (1.7%)	4 (2.1%)	0	0	0	1 (1.8%)
Nystagmus	174 (7.4%)	7 (3.7%)	39 (10.8%)	21 (7.0%)	1 (3.1%)	4 (7.0%)
PFV	36 (1.5%)	2 (1.1%)	6 (1.7%)	0	1 (3.1%)	3 (5.3%)
Other fundus diseases	54 (2.3%)	4 (2.1%)	2 (0.6%)	7 (2.3%)	0	0
Congenital ptosis	101 (4.3%)	10 (5.3%)	0	2 (0.7%)	0	0
Retinoblastoma	41 (1.7%)	6 (3.2%)	2 (0.6%)	3 (1.0%)	0	2 (3.5%)
Strabismus	245 (10.5%)	15 (8.0%)	35 (9.7%)	28 (9.4%)	1 (3.1%)	6 (10.5%)
Limbal dermoid	34 (1.5%)	3 (1.6%)	0	0	0	0
Microphthalmia	19 (0.8%)	2 (1.1%)	1 (0.3%)	3 (1.0%)	0	0
Pupillary membranes	19 (0.8%)	2 (1.1%)	7 (1.9%)	1 (0.3%)	1 (3.1%)	1 (1.8%)
SSOM	11 (0.5%)	0	0	1 (0.3%)	0	0
Other	35 (1.5%)	4 (2.1%)	3 (0.8%)	0	0	0

PFV, persistent fetal vasculature; SSOM, systemic syndromes with ocular manifestations.

88 children. On the validation set, 31 (35.2%) children were classified as nonimpaired and 57 (64.8%) children were classified as visually impaired. AIS achieved effective performance in the at-home implementation, with an AUC of 0.817 (95% CI, 0.756–0.881) for discriminating clips of visually impaired children from those of nonimpaired children (Extended Data Fig. 4c). At the child level, significantly lower predicted probability patterns were obtained for the nonimpaired children compared with mildly or severely impaired children (Fig. 2k). An AUC of 0.859 (95% CI, 0.767–0.950), an accuracy of 77.3% (95% CI, 67.5%–84.8%), a sensitivity of 77.2% (95% CI, 64.8%–86.2%) and a specificity of 77.4% (95% CI, 60.2%–88.6%) were attained for discriminating visual impairment from nonimpairment (Fig. 2l and Supplementary Table 3).

Model visualization and explanation

We improved the interpretability of the detection model outputs by visualizing the model results in the internal validation set. After being projected into a two-dimensional space, the feature information extracted by the detection model exhibited distinct patterns between the visually impaired and nonimpaired clips (Fig. 3a). The attention patterns of the detection model presented by the average heat maps varied with the children's visual functions and underlying ophthalmic disorders (Fig. 3b,c). Among the visually impaired children, the detection model focused more on the eyes and areas around the neck (Fig. 3c). In particular, for the clips extracted from visually impaired samples, those classified by human experts as having abnormal patterns were more likely to be predicted by our system as 'visual impairment' than those that were randomly extracted (Fig. 3d,e and Supplementary Table 8), indicating that the detection model might pay more attention to the morphological appearance or behavioral patterns of the eye and head regions, as we previously reported[16].

Additionally, the clips misidentified by the system exhibited different clustering characteristics from the correctly recognized clips (true visually impaired or true nonimpaired clips), and more of the misidentified clips fell in the intermediate zone of the two clusters for the correctly recognized clips (Extended Data Fig. 5). Moreover, for the 20% of samples with the lowest predicted confidence values, the false identification rate was significantly higher than that of other groups and the system was equivocal. We aimed to find a solution when the system was unreliable by filtering out equivocal samples for manual review by ophthalmologists. The results show that the system performance was substantially improved with the increasing ratios for manual review. For instance, when selecting cases with confidence values less than 0.071 for manual review, accounting for 3% of the total cases, the sensitivity improved from 84.1% to 85.1% and the specificity improved from 91.9% to 93.1%; when selecting cases with confidence values less than 0.193 for manual review, accounting for 7% of the total cases, the sensitivity and specificity improved to 85.4% and 94.2%, respectively (Extended Data Fig. 6).

Multiple-category classification of ophthalmic disorders

Considering that our system exhibited different attention patterns for visual impairment caused by specific ophthalmic disorders (Fig. 3c), we further developed a DL-based diagnostic model to differentiate ophthalmic disorders with characteristic attention patterns by the detection model (aphakia, congenital glaucoma, congenital ptosis and strabismus) and nonimpairment at the child level. In the diagnostic validation, our system effectively discriminated multiple ophthalmic disorders, achieving AUCs ranging from 0.918 for strabismus to 0.996 for congenital ptosis (Fig. 2m,n).

Reliability and adjusted analyses

Stable performance is critical for real-world applications of mHealth and medical AI systems. Thus, we investigated the reliability of AIS at the clinic of ZOC. We first evaluated the influences of patient-related factors, including sex, age, laterality of the eye disorder and the apparency of the phenotypic features, on the performance of AIS. For the reliability stratified by sex, AIS achieved an AUC of 0.948 (95% CI, 0.921–0.971) in the boys group and an AUC of 0.931 (95% CI, 0.899–0.961) in the girls group (Fig. 4a). The predicted probability pattern of AIS remained

https://doi.org/10.1038/s41591-022-02180-9

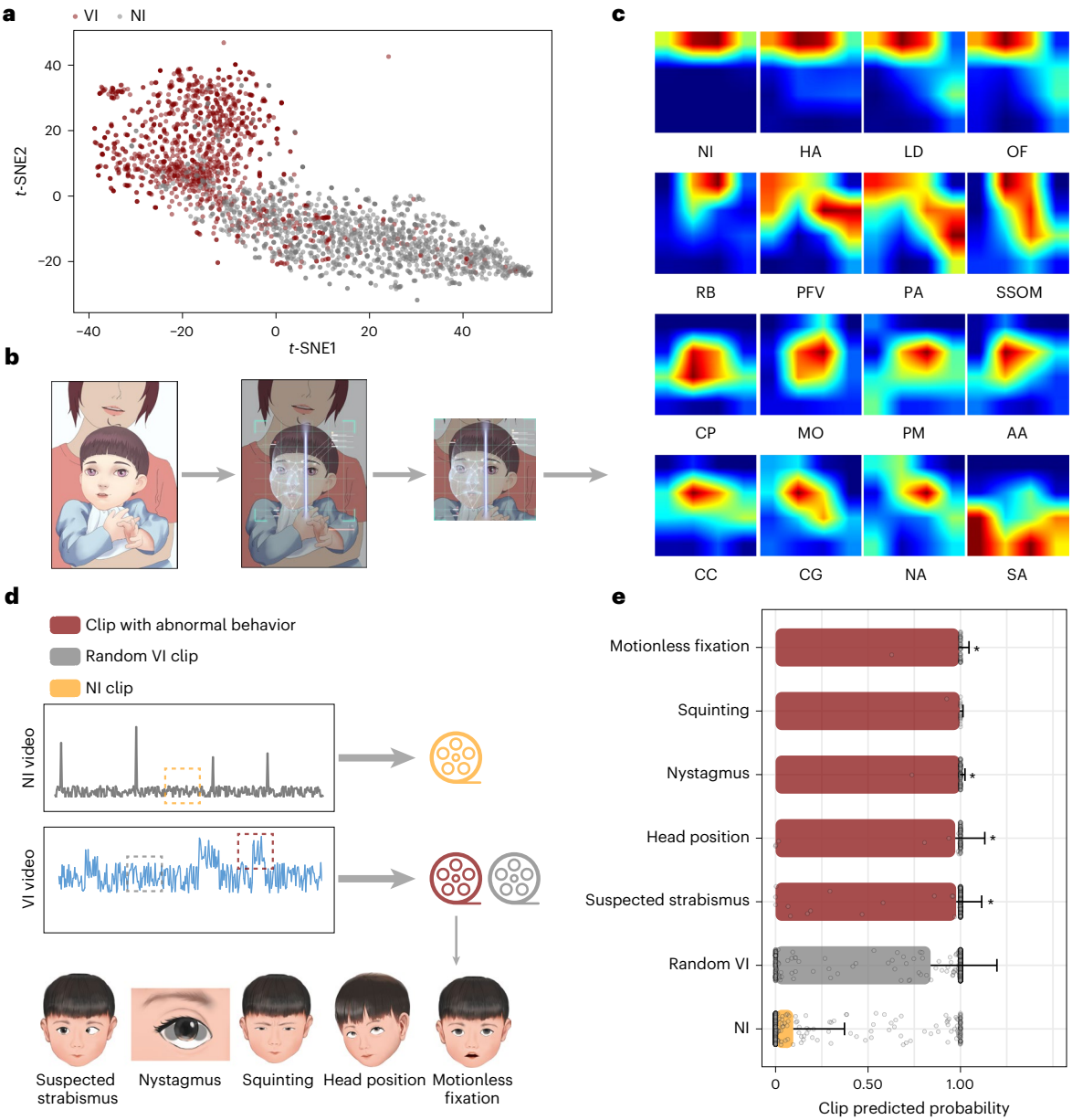


Fig. 3 | Interpretability and visualization of the detection model. a, The *t*-SNE algorithm was applied to visualize the detection model at the clip level. **b**, Facial detection and facial landmark localization algorithms were applied to detect and crop the facial regions of the children before data served as inputs to the AIS system. **c**, Average heat maps obtained from the detection model based on the inputs of facial regions in (**b**) for nonimpaired children and for children with the indicated ophthalmic disorders. **d**, The predicted probabilities for various types of clips were compared: clips randomly extracted from the videos of nonimpaired children, clips randomly extracted from the videos of visually impaired children and clips labeled by experienced ophthalmologists as having abnormal behavioral patterns extracted from videos of visually impaired children. **e**, Predicted probabilities of the detection model for various types of clips in (**d**) were compared: motionless fixation, $n = 48$; squinting, $n = 18$; nystagmus (NA), $n = 95$; head position, $n = 115$; suspected strabismus (SA), $n = 360$; random visual impairment (VI), $n = 1{,}000$; nonimpairment (NI), $n = 1{,}000$. Results are expressed as mean ± s.d. $^*P < 0.01$ for comparisons with random VI (motionless fixation, $P = 4.60 \times 10^{-8}$; suspected SA, $P = 1.09 \times 10^{-15}$; NA, $P = 1.52 \times 10^{-7}$; head position, $P = 0.005$; two-tailed Mann–Whitney *U*-tests). AA, aphakia; CC, congenital cataract; CG, congenital glaucoma; CP, congenital ptosis; HA, high ametropia; LD, limbal dermoid; MO, microphthalmia; OF, other fundus diseases; PA, Peters' anomaly; PFV, persistent fetal vasculature; PM, pupillary membrane; RB, retinoblastoma; SSOM, systemic syndromes with ocular manifestations.

stable under various age conditions (Fig. 4b), and the system achieved AUCs ranging from 0.909 for age group 4 to 0.954 for age group 3 (Fig. 4c). Additionally, AIS effectively identified visually impaired children with bilateral or unilateral eye disorders, with an AUC of 0.921 (95% CI, 0.891–0.952) in the unilateral group and an AUC of 0.952 (95% CI, 0.932–0.973) in the bilateral group (Fig. 4d). In addition, AIS achieved satisfactory performance with an AUC of 0.939 (95% CI, 0.918–0.960) in identifying hard-to-spot visually impaired children, who could have insidious phenotypic features and were easily neglected by community ophthalmologists (Supplementary Table 9).

Furthermore, we investigated the reliability of AIS under different data capture conditions, including testing distance, room illuminance, repeated testing and duration of the video recording. Similarly, AIS obtained stable detection performance among groups of different testing distances, with the lowest AUC of 0.935 (95% CI, 0.912–0.958) in the medium-distance group (Fig. 4e). Additionally, the AIS predicted

https://doi.org/10.1038/s41591-022-02180-9

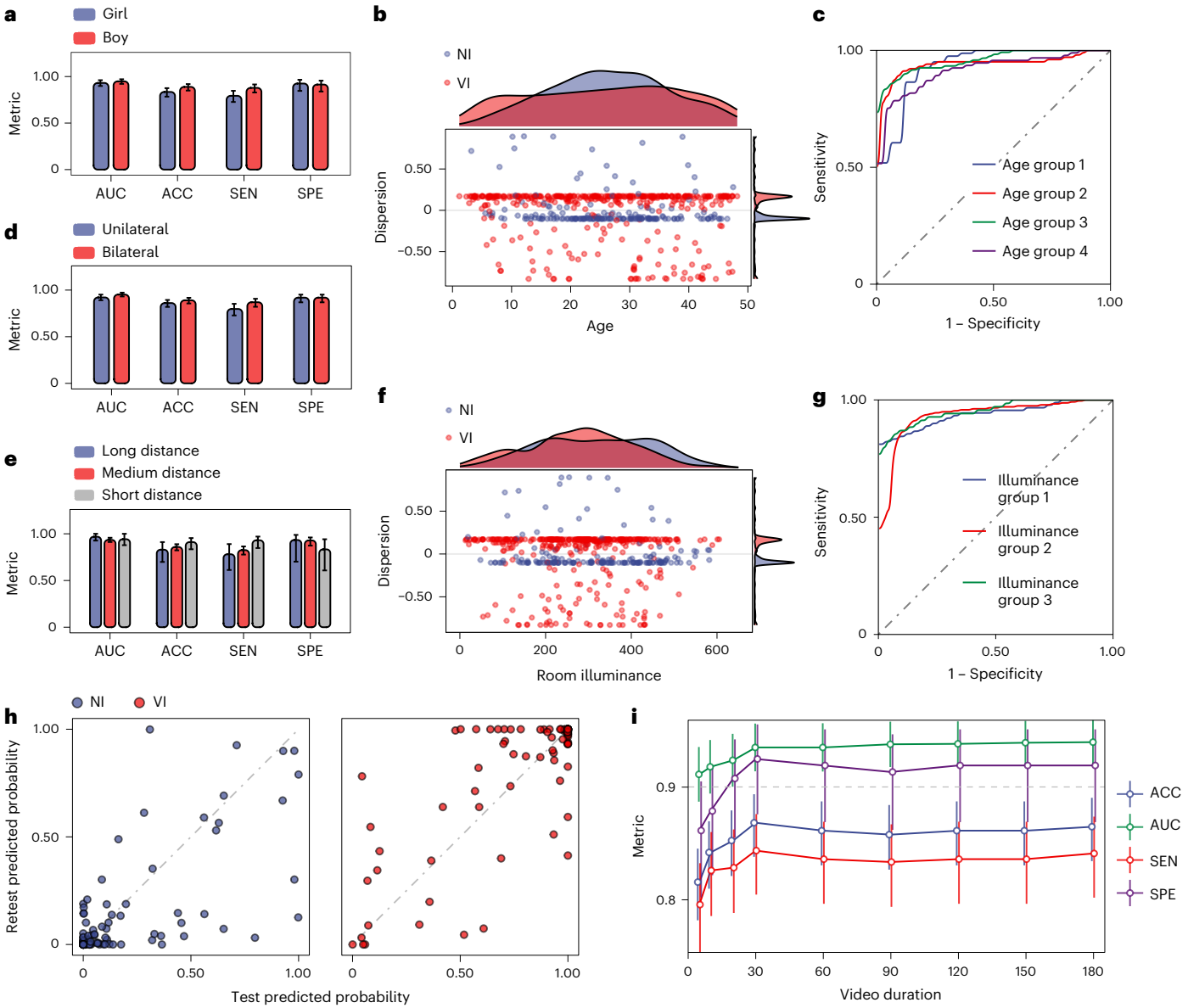


Fig. 4 | Performance of the AIS system in reliability analyses. a, Performance of AIS in detecting children with visual impairment (VI) based on sex: girls, $n = 254$; boys, $n = 315$. **b**, Scatterplot of dispersion of the AIS predicted probability changes by age (months). **c**, Receiver operating characteristic (ROC) curves of AIS for detecting children with VI by age groups: age group 1, age ≤ 12 months, $n = 98$, AUC = 0.925 (0.847–1.000); age group 2, 12 months < age ≤ 24 months, $n = 160$, AUC = 0.936 (0.895–0.977); age group 3, 24 months < age ≤ 36 months, $n = 189$, AUC = 0.954 (0.928–0.980); age group 4, 36 months < age ≤ 48 months, $n = 122$, AUC = 0.909 (0.855–0.964). **d**, Performance of AIS for identifying children with unilateral or bilateral VI: unilateral, $n = 158$; bilateral, $n = 238$; nonimpairment (NI), $n = 173$. **e**, Performance of AIS for detecting children with VI under various testing distance conditions: long distance, $n = 47$; medium distance, $n = 432$; short distance, $n = 90$. **f**, Scatterplot of dispersion of the AIS predicted probability changes by room illuminance (in lux (lx)). **g**, ROC curves of AIS for distinguishing children with VI under various room illuminance conditions: illuminance group 1, room illuminance ≤ 200 lx, $n = 125$, AUC = 0.936 (0.895–0.976); illuminance group 2, 200 lx < room illuminance ≤ 400 lx, $n = 317$, AUC = 0.932 (0.901–0.963); illuminance group 3, room illuminance > 400 lx, $n = 127$, AUC = 0.950 (0.915–0.985). **h**, Predicted probabilities of the detection model for repeated detection tests (NI, $n = 102$; VI, $n = 85$). **i**, Performance curves of AIS by video duration. In **a**,**d**,**e**, results are expressed as means and 95% CIs with DeLong CIs for AUC values and 95% Wilson CIs for other metrics. ACC, accuracy; SEN, sensitivity; SPE, specificity.

probability pattern remained stable under different room illuminance conditions (Fig. 4f). Our system achieved the lowest AUC of 0.932 (95% CI, 0.901–0.963) in the medium illuminance group (Fig. 4g). In the retest analysis, the system remained robust with an intraclass correlation coefficient for predicted probabilities of 0.880 (95% CI, 0.843–0.908) and a Cohen's κ for predicted categories of 0.837 (95% CI, 0.758–0.916) in another independent validation population recruited at ZOC (Fig. 4h and Table 1). In addition, as the duration of the video recording increased, AIS remained stable and achieved a maximal AUC of 0.931 (95% CI, 0.914–0.956) with a video duration longer than 30 s (Fig. 4i).

To further verify that the detecting results of our system were reliable and not solely mediated by baseline characteristics as confounders, we examined the odds ratios (ORs) of the AIS predictions adjusted for baseline characteristics at the child level. Even after controlling for potential baseline confounders, the AIS predictions had statistically significant adjusted ORs for detecting visual impairment in the internal and external validations and the at-home implementation ($P < 0.001$).

The adjusted ORs ranged from 3.034 to 3.248 for tasks in the internal validation (Supplementary Table 10) and from 2.307 to 2.761 for tasks in the external validation (Supplementary Table 11). For the at-home implementation, the AIS predictions had a statistically significant adjusted OR of 2.496 (95% CI, 1.748–3.565, $P = 4.815 \times 10^{-7}$) for detecting visual impairment (Supplementary Table 12).

Performance of the AIS across different smartphone platforms

To test the stability of our system in more complex settings, we performed adjustments to a dataset randomly sampled from the ZOC validation set with various blurring, brightness, color or Gaussian noise adjustment gradients to simulate the diversity of data quality collected by different smartphone cameras. Our system remained reliable and achieved AUCs of over 0.800 with blurring factors no more than 25 or brightness factors no more than 0.7, and it achieved AUCs of over 0.930 under different color adjustments and over 0.820 under various Gaussian noise adjustments (Extended Data Fig. 7).

Furthermore, an independent validation set from 389 children was collected at ZOC using the Huawei Honor-6 Plus and Redmi Note-7 smartphones with the Android operation system to evaluate the performance of AIS (Fig. 1b and Supplementary Table 13). After data quality checking, videos of 361 children were reserved (92.8%), including 87 (24.1%) children without visual impairment, 169 (46.8%) children with mild visual impairment and 105 (29.1%) children with severe visual impairment (Table 1). AIS showed significantly higher predicted probabilities for mild or severe impairment than for nonimpairment and achieved an AUC of 0.932 (95% CI, 0.902–0.963) for identifying visual impairment for the Android system at the child level (Extended Data Fig. 8).

Discussion

With the high incidence of visual problems during the first few years of life, timely intervention to counter pathological visual deprivation mechanisms during this critical development period can prevent or minimize long-term visual loss[3]. However, early detection of visual impairment in young children is challenging due to the lack of accurate and easy-to-use tools applicable to both clinical and community environments. To overcome these challenges, we developed and validated a smartphone-based system (AIS) that provides a holistic and quantitative technique to identify visual impairment in young children in real-world settings. We comprehensively evaluated this system for 16 important causes of childhood vision loss. Our system achieved an AUC of 0.940 in the internal validation and an AUC of 0.843 in the external validation at the clinics of four different hospitals. Furthermore, our system proved reliable when used by parents or caregivers at home, achieving an AUC of 0.859 under these specific testing conditions.

One of the merits of AIS is in its applicability to different ocular diseases. Previous studies have utilized photographs to detect ocular and visual abnormalities in childhood[27,28]. These technologies, which focus on a single static image, are not suitable for large-scale applications due to their limited effectiveness and inability to handle multiple abnormalities with variable patterns. Given the complexity of ocular pathologies in children, the concept of accurately assessing a broad range of ocular conditions is attractive. In our prospective multicenter study, we analyzed more than 25,000,000 frames of information-rich phenotypic videos and accurately identified visual impairment caused by a wide range of sight-threatening eye diseases. Strikingly, AIS was able to detect most of the common causes of visual impairment in childhood, including anterior and posterior segment disorders, strabismus, ocular neoplasms, developmental abnormalities and ocular manifestations of systemic and genetic diseases[29]. Although cases like congenital cataracts tend to be easily diagnosed in specialist settings by experienced doctors, they are still frequently missed in the community, especially in areas with pediatric ophthalmic resource shortfall[28]. To apply AIS to various scenarios, we recruited cases of a broad range of eye disorders with variable severity in terms of their impact on vision. Our system was reasonably accurate in identifying mildly impaired children who could have subtle phenotypic features, making them easy to miss. Furthermore, our results indicate that AIS can be extended to diseases that have not been previously encountered in the training process, demonstrating its broader applicability.

The use of smartphones to detect visual impairment caused by extraocular diseases or systematic diseases is an important application in the future, but the feasibility remained to be further verified. Some systemic diseases, such as cardiovascular, hepatobiliary and renal diseases, can exhibit ocular manifestations that are recognizable by algorithms, which is also indicated by our findings in small samples[30–32]. Furthermore, disorders of neurological system can impact vision and cause cerebral visual impairment with pathology outside the eye, which is a common type of visual impairment in developed countries but lacking in this study[33,34]. Therefore, future work is needed to evaluate the merit of AIS in detecting visual impairment caused by a broad range of diseases, such as cerebral visual impairment, and in reducing the extraocular morbidity associated with systemic diseases in a larger population: for example, the cardiovascular complications linked with Marfan syndrome.

A major strength of AIS is its reliability in real-world practice. Although a large number of medical AI systems have been evaluated with high performance in the laboratory setting, only a few systems have demonstrated real-world medical feasibility[23,25]. Bias from training data and low stability of the model design greatly limit the generalizability of these AI systems. Previously, we evaluated the feasibility of identifying visual impairment in children by analyzing their phenotypic characteristics using DL algorithms[16]. For that study, the evaluation was conducted by experienced experts under a tightly controlled, standardized laboratory setting to strictly control for interference factors, which is not possible in routine ophthalmic practice. In this study, we prospectively collected a large amount of phenotypic data (facial features and ocular movements) to develop a DL system with a highly reliable design. Our results show that AIS exhibited high stability and prediction effectiveness under various testing conditions. Importantly, AIS remained effective in multicenter external validation and crucially, when rolled out in the community and used by parents or caregivers at home. When transferred to at-home settings, factors such as environmental interference, blurring, brightness, pixels of different cameras and the influence of untrained operators may impact the system's performance. Therefore, we used a pilot dataset to fine-tune our system for its generalizability to various home environments and broader applications. AIS achieved an acceptable AUC of 0.859 in the subsequent implementation, which indicates that it can benefit from further model updating on larger-scale datasets for broader applications. Importantly, AIS kept stable in 88 different types of home environments after one round of fine-tuning, demonstrating its potential to be used generally in a variety of complex environments with no requirement of regular adaptations or fine-tuning in the future application.

Our findings demonstrate that sensory states, especially vision, can be derived from phenotypic video data recorded using consumer-grade smartphones. Two types of underlying features seemed to be captured by smartphones. First, changes in facial appearance caused by ocular pathologies can be directly recorded by mobile devices, especially those of the ocular surface or adnexa: for example, eyelid drooping in congenital ptosis. Second and more importantly, individuals may display aberrant behaviors to adapt to changes in their sensory modality, a process conserved from arthropods to mammals[35,36] and confirmed in human children[16]. Our results show that the model can focus on behavioral features replicated in various eye diseases, such as abnormal ocular movement or alignment/fixation patterns. These common behavioral patterns may broaden the

applicability of AIS to multiple ocular diseases, including posterior segment abnormalities that are more challenging to diagnose based on phenotypic video data.

A smartphone-based system to detect ocular pathology in children has obvious clinical implications. Early identification by parents or caregivers of ocular abnormalities facilitates timely referral to pediatric ophthalmologists and prompt intervention. AIS does not require professional medical equipment; smartphones and simple stabilization are sufficient. This low-barrier system is a promising tool for the timely testing of children in the community, which is a major advantage given the rapidly changing nature of the ocular pathology encountered in children. This could have a major impact by improving vision-related outcomes and even survival rates in cases such as retinoblastoma[37,38]. Furthermore, AIS is a promising tool to screen young children for ocular abnormalities remotely, which can reduce ophthalmologists' exposure risk to infectious agents, as exemplified by the impact of the coronavirus disease 2019 (COVID-19) pandemic, in the so-called 'new normal' period[39].

This study has several limitations. First, although we may miss the recruitment of some patients with conditions causing slight visual impairment in specialist clinical settings, our system was satisfactorily accurate in identifying mildly impaired children with subtle phenotypic features. Importantly, the versatile AIS system kept reliable performance to detect visually impaired children who were hard to spot even for community ophthalmologists, which sheds light on its significant application prospect of expanding our future work to the general population and groups of children with mild or early-stage ocular pathology. Second, to develop the quality control module and analyze the influencing factors, only a single video was collected for each child at ZOC, accounting for the relatively high rate of unsuccessful cases in this stage. However, our system allowed users to repeat video recordings until the qualified videos were acquired. As a result, the successful rate of identification greatly improved. Although a proportion of uncooperative children may not be appropriate for our tool, our AIS system has greatly lowered the minimal operating threshold for untrained users, indicating the potential for the general applications. Third, our cohorts recruited in clinical settings may not represent the real-world population. Although AIS effectively identified visually impaired children in the finding a needle in a haystack test with a prevalence simulated to a general population, a large-scale screening trial is needed in the future to validate the utility of the AIS system in the real-world applications. Fourth, AIS requires collecting facial information from children, which may pose a risk of privacy exposure. To avoid potential privacy risks, future techniques such as lightweight model backbones[40] and model pruning[41] could be applied to deploy the DL system in individual smartphones with no requirement for additional computing resources. In addition, digital fingerprint technology, such as blockchain[42], can also be applied to monitor data usage and mitigate abuse effectively. Additionally, we developed a real-time three-dimensional facial reconstruction technology to irreversibly erase biometric attributes while retaining gaze patterns and eye movements[43], which can be used in the future to safeguard children's privacy when using AIS.

In conclusion, we developed and validated an innovative smartphone-based technique to detect visual impairment in young children affected with a broad range of eye diseases. Given the ubiquity of smartphones, AIS is a promising tool that can be applied in real-world settings for secondary prevention of visual loss in this particularly vulnerable age group.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-022-02180-9.

References

1. Kliner, M., Fell, G., Pilling, R. & Bradbury, J. Visual impairment in children. *Eye* **25**, 1097–1097 (2011).
2. Mariotti, A. & Pascolini, D. Global estimates of visual impairment. *Br. J. Ophthalmol.* **96**, 614–618 (2012).
3. Bremond-Gignac, D., Copin, H., Lapillonne, A. & Milazzo, S. Visual development in infants: physiological and pathological mechanisms. *Curr. Opin. Ophthalmol.* **22**, S1–S8 (2011).
4. Teoh, L., Solebo, A. & Rahi, J. Temporal trends in the epidemiology of childhood severe visual impairment and blindness in the UK. *Br. J. Ophthalmol.* https://doi.org/10.1136/bjophthalmol-2021-320119 (2021).
5. Gothwal, V. K., Lovie-Kitchin, J. E. & Nutheti, R. The development of the LV Prasad-Functional Vision Questionnaire: a measure of functional vision performance of visually impaired children. *Investigative Ophthalmol. Vis. Sci.* **44**, 4131–4139 (2003).
6. Brown, A. M. & Yamamoto, M. Visual acuity in newborn and preterm infants measured with grating acuity cards. *Am. J. Ophthalmol.* **102**, 245–253 (1986).
7. Dutton, G. N. & Blaikie, A. J. How to assess eyes and vision in infants and preschool children. *BMJ Br. Med. J.* **350**, h1716 (2015).
8. Blindness and Vision Impairment (World Health Organization, 2021); https://www.who.int/en/news-room/fact-sheets/detail/blindness-and-visual-impairment
9. Mayer, D. L. & Dobson, V. in *Developing Brain Behaviour* (ed. Dobbing, J.) 253–292 (Academic, 1997).
10. Quinn, G. E., Berlin, J. A. & James, M. The Teller acuity card procedure: three testers in a clinical setting. *Ophthalmology* **100**, 488–494 (1993).
11. Johnson, A., Stayte, M. & Wortham, C. Vision screening at 8 and 18 months. Steering Committee of Oxford Region Child Development Project. *Br. Med. J.* **299**, 545–549 (1989).
12. Long, E. et al. Monitoring and morphologic classification of pediatric cataract using slit-lamp-adapted photography. *Transl. Vis. Sci. Technol.* **6**, 2 (2017).
13. Balmer, A. & Munier, F. Differential diagnosis of leukocoria and strabismus, first presenting signs of retinoblastoma. *Clin. Ophthalmol.* **1**, 431 (2007).
14. SooHoo, J. R., Davies, B. W., Allard, F. D. & Durairaj, V. D. Congenital ptosis. *Surv. Ophthalmol.* **59**, 483–492 (2014).
15. Mandal, A. K. & Chakrabarti, D. Update on congenital glaucoma. *Indian J. Ophthalmol.* **59**, S148 (2011).
16. Long, E. et al. Discrimination of the behavioural dynamics of visually impaired infants via deep learning. *Nat. Biomed. Eng.* **3**, 860–869 (2019).
17. Brown, A. M. & Lindsey, D. T. Infant color vision and color preferences: a tribute to Davida Teller. *Vis. Neurosci.* **30**, 243–250 (2013).
18. Holmes, J. M. & Clarke, M. P. Amblyopia. *Lancet* **367**, 1343–1351 (2006).
19. Abadi, R. & Bjerre, A. Motor and sensory characteristics of infantile nystagmus. *Br. J. Ophthalmol.* **86**, 1152–1160 (2002).
20. Wright, K. W., Spiegel, P. H. & Hengst, T. *Pediatric Ophthalmology and Strabismus* (Springer, 2013).
21. Sim, I. Mobile devices and health. *N. Engl. J. Med.* **381**, 956–968 (2019).
22. Grady, C. et al. Informed consent. *N. Engl. J. Med.* **376**, 856–867 (2017).
23. Beede, E. et al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proc. 2020 CHI Conference on Human Factors in Computing Systems* 1–12 (Association for Computing Machinery, 2020)..
24. Davenport, T. H. & Ronanki, R. Artificial intelligence for the real world. *Harvard Bus. Rev.* **96**, 108–116 (2018).

25. Lin, H. et al. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. *eClinicalMedicine* **9**, 52–59 (2019).
26. King, D. E. Dlib-ml: a machine learning toolkit. *J. Mach. Learn. Res.* **10**, 1755–1758 (2009).
27. Munson, M. C. et al. Autonomous early detection of eye disease in childhood photographs. *Sci. Adv.* **5**, eaax6363 (2019).
28. Long, E. et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. *Nat. Biomed. Eng.* **1**, 0024 (2017).
29. Gogate, P., Gilbert, C. & Zin, A. Severe visual impairment and blindness in infants: causes and opportunities for control. *Middle East Afr. J. Ophthalmol* **18**, 109–114 (2011).
30. Cheung, C. Y. et al. A deep-learning system for the assessment of cardiovascular disease risk via the measurement of retinal-vessel calibre. *Nat. Biomed. Eng.* **5**, 498–508 (2021).
31. Sabanayagam, C. et al. A deep learning algorithm to detect chronic kidney disease from retinal photographs in community-based populations. *Lancet Digital Health* **2**, e295–e302 (2020).
32. Xiao, W. et al. Screening and identifying hepatobiliary diseases through deep learning using ocular images: a prospective, multicentre study. *Lancet Digital Health* **3**, e88–e97 (2021).
33. Pehere, N., Chougule, P. & Dutton, G. N. Cerebral visual impairment in children: causes and associated ophthalmological problems. *Indian J. Ophthalmol.* **66**, 812–815 (2018).
34. Gilbert, C. & Foster, A. Childhood blindness in the context of VISION 2020—the right to sight. *Bull. World Health Organ* **79**, 227–232 (2001).
35. Dey, S. et al. Cyclic regulation of sensory perception by a female hormone alters behavior. *Cell* **161**, 1334–1344 (2015).
36. Klein, M. et al. Sensory determinants of behavioral dynamics in *Drosophila* thermotaxis. *Proc. Natl Acad. Sci. USA* **112**, E220–E229 (2015).
37. Finger, P. T. & Tomar, A. S. Retinoblastoma outcomes: a global perspective. *Lancet Glob. Health* **10**, e307–e308 (2022).
38. Wong, E. S. et al. Global retinoblastoma survival and globe preservation: a systematic review and meta-analysis of associations with socioeconomic and health-care factors. *Lancet Glob. Health* **10**, E380–E389 (2022).
39. Romano, M. R. et al. Facing COVID-19 in ophthalmology department. *Curr. Eye Res.* **45**, 653–658 (2020).
40. Howard, A. et al. Searching for mobilenetv3. In *Proc. IEEE/CVF International Conference on Computer Vision* 1314–1324 (IEEE, 2019).
41. Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N. & Peste, A. Sparsity in deep learning: pruning and growth for efficient inference and training in neural networks. *J. Mach. Learn. Res.* **22**, 1–124 (2021).
42. Leeming, G., Ainsworth, J. & Clifton, D. A. Blockchain in health care: hype, trust, and digital health. *Lancet* **393**, 2476–2477 (2019).
43. Yang, Y. et al. A digital mask to safeguard patient privacy. *Nat. Med.* **28**, 1883–1892 (2022).

Wenben Chen[1,24], Ruiyang Li[1,24], Qinji Yu[2,24], Andi Xu[1,24], Yile Feng[3,24], Ruixin Wang[1], Lanqin Zhao[1], Zhenzhe Lin[1], Yahan Yang[1], Duoru Lin[1], Xiaohang Wu[1], Jingjing Chen[1], Zhenzhen Liu[1], Yuxuan Wu[1], Kang Dang[3], Kexin Qiu[3], Zilong Wang[3], Ziheng Zhou[3], Dong Liu[1], Qianni Wu[1], Mingyuan Li[1], Yifan Xiang[1], Xiaoyan Li[1], Zhuoling Lin[1], Danqi Zeng[1], Yunjian Huang[1], Silang Mo[4], Xiucheng Huang[4], Shulin Sun[5], Jianmin Hu[6], Jun Zhao[7], Meirong Wei[8], Shoulong Hu[9,10], Liang Chen[11], Bingfa Dai[6], Huasheng Yang[1], Danping Huang[1], Xiaoming Lin[1], Lingyi Liang[1], Xiaoyan Ding[1], Yangfan Yang[1], Pengsen Wu[1], Feihui Zheng[12], Nick Stanojcic[13], Ji-Peng Olivia Li[14], Carol Y. Cheung[15], Erping Long[1], Chuan Chen[16], Yi Zhu[17], Patrick Yu-Wai-Man[14,18,19,20], Ruixuan Wang[21], Wei-shi Zheng[21], Xiaowei Ding[2,3] ✉ & Haotian Lin[1,22,23] ✉

[1]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Vision Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou, China. [2]Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China. [3]VoxelCloud, Shanghai, China. [4]School of Medicine, Sun Yat-sen University, Shenzhen, China. [5]Department of Urology, Peking University Third Hospital, Peking University Health Science Center, Beijing, China. [6]Department of Ophthalmology, The Second Affiliated Hospital of Fujian Medical University, Quanzhou, China. [7]Shenzhen People's Hospital (The Second Clinical Medical College, Jinan University; The First Affiliated Hospital, Southern University of Science and Technology), Shenzhen, China. [8]Liuzhou Maternity and Child Healthcare Hospital, Affiliated Women and Children's Hospital of Guangxi University of Science and Technology, Liuzhou, China. [9]National Center for Children's Health, Department of Ophthalmology, Beijing Children's Hospital, Capital Medical University, Beijing, China. [10]Department of Ophthalmology, Zhengzhou Children's Hospital, Zhengzhou, China. [11]Shenzhen Eye Hospital, Jinan University, Shenzhen Eye Institute, Shenzhen, China. [12]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore. [13]Department of Ophthalmology, St. Thomas' Hospital, London, UK. [14]Moorfields Eye Hospital, London, UK. [15]Department of Ophthalmology & Visual Sciences, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China. [16]Sylvester Comprehensive Cancer Center, University of Miami Miller School of Medicine, Miami, FL, USA. [17]Department of Molecular and Cellular Pharmacology, University of Miami Miller School of Medicine, Miami, FL, USA. [18]University College London Institute of Ophthalmology, University College London, London, UK. [19]Cambridge Eye Unit, Addenbrooke's Hospital, Cambridge University Hospitals, Cambridge, UK. [20]Cambridge Center for Brain Repair and Medical Research Council (MRC) Mitochondrial Biology Unit, Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK. [21]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China. [22]Hainan Eye Hospital and Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Haikou, China. [23]Center for Precision Medicine and Department of Genetics and Biomedical Informatics, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China. [24]These authors contributed equally: Wenben Chen, Ruiyang Li, Qinji Yu, Andi Xu, Yile Feng. ✉e-mail: dingxiaowei@sjtu.edu.cn; linht5@mail.sysu.edu.cn

Methods

Ethics approval

The predefined protocol of the clinical study was approved by the Institutional Review Board/Ethics Committee of ZOC and prospectively registered at ClinicalTrials.gov (identifier: NCT04237350), and it is shown in Supplementary Note. Consent was obtained from all individuals whose eyes or faces are shown in the figures or video for publication. Before data collection, informed written consent was obtained from at least one parent or guardian of each child. The investigators followed the requirements of the Declaration of Helsinki throughout the study.

Study design and study population

This prospective, multicenter and observational study was conducted between 14 January 2020 and 30 January 2022 to recruit children for the development and validation of the mHealth system in three stages (Fig. 1b). Major eligibility criteria included an age of 48 months or younger and informed written consent obtained from at least one parent or guardian of each child. We did not include children having central nervous system diseases, mental illnesses or other known illnesses that could affect their behavioral patterns, in the absence of ocular manifestations. Children who could not cooperate to complete the ophthalmic examinations or the detection test using AIS were excluded. We also excluded children who had received ocular interventions and treatments in the month immediately preceding data collection.

In the first stage completed from 14 January 2020 to 15 September 2021, children were enrolled at the clinic of ZOC (Guangdong Province) to develop and comprehensively validate (internal validation and reliability analyses) the system. In the second stage, which occurred from 22 September 2021 to 19 November 2021, children were enrolled at the clinics of the Second Affiliated Hospital of Fujian Medical University (Fujian Province), Shenzhen Eye Hospital (Guangdong Province), Liuzhou Maternity and Child Healthcare Hospital (Guangxi Province) and Beijing Children's Hospital of Capital Medical University (Beijing) to additionally evaluate the system (external validation). We selected these sites from three provinces across northern and southern China, representing the variations in clinical settings. In the first two stages, recruited children underwent ophthalmic examinations by clinical staff, and phenotypic videos were collected by trained volunteers using mHealth apps installed on iPhone-7 or iPhone-8 smartphones at each center. In the third stage conducted from 24 November 2021 to 30 January 2022, we advertised our study through the online platform of the Pediatric Department of ZOC and the social media of WeChat. We recruited children and their parents or caregivers online from the Guangdong area for at-home implementation. The investigators recruited the children following the same eligibility criteria as the previous two stages by collecting their basic information and medical history online. In addition, children who could not come to ZOC for an ophthalmic assessment or who had been included in other stages of this study were excluded. Untrained parents or caregivers recorded the phenotypic videos with their smartphones according to the instructions of the AIS app at home (Extended Data Figs. 1 and 2). The quality control module automatically reminded parents or caregivers to repeat data collection when the video recordings were unqualified. In this stage, all the children who completed successful video recordings underwent ophthalmic examinations at ZOC. A total of 3,652 children were finally enrolled, recording more than 25,000,000 frames of videos for development and validation of the system.

Definition of visual impairment

Comprehensive functional and structural examinations were performed to stratify children's visual conditions for developing and validating the DL-based AIS. For unified examination, a teller vision card (Stereo Optical Company) was utilized to measure children's monocular visual acuity[44]. In addition, high-resolution slit lamp examinations, fundoscopy examinations and cycloplegic refraction were used to detect abnormalities in the eyes. Additional examinations, such as intraocular pressure, ultrasound, computerized tomography scans and genetic tests, were determined by experienced pediatric ophthalmologists when necessary.

According to the results of the abovementioned examinations and a referenced distribution of monocular visual acuity[45], experienced pediatric ophthalmologists comprehensively stratified children's visual conditions into three groups. Children with the best-corrected visual acuity (BCVA) of both eyes in the 95% referenced range with no abnormalities of structure or other examination results were assigned to the nonimpaired group. Children with the BCVA in the 99% referenced range in both eyes with abnormalities of structure or other examination results were assigned to the mildly impaired group. Children with the BCVA of at least one eye outside the 99% referenced range or worse than light perception with structural abnormalities or other examination results were assigned to the severely impaired group[16]. We recruited visually impaired children with primary diagnoses of the following 16 ocular disorders: aphakia, congenital cataract, congenital glaucoma, high ametropia, Peters' anomaly, nystagmus, congenital ptosis, strabismus, persistent fetal vasculature, retinoblastoma, other fundus diseases, limbal dermoid, microphthalmia, pupillary membranes, systemic syndromes with ocular manifestations and other ocular conditions (Table 2 and Supplementary Table 4). A tiered panel consisting of two groups of experts assigned and confirmed the primary diagnosis as the most significant diagnostic label for each child. The first group of experts consisted of two pediatric ophthalmologists with over 10 years of experience in each recruiting ophthalmic center who separately provided the preliminary labeling information. If a consensus was not reached at this stage, a second group of more senior pediatric ophthalmologists with over 20 years of experience at ZOC verified the diagnostic labels as the ground truth. The diagnoses of children recruited online for the at-home implementation were made by experts at ZOC following the same criteria.

Concept of the AIS system

The AIS system consisted of a smartphone app (available for iPhone and Android operating systems) for data collection and a DL back end for data analysis (Fig. 1a and Extended Data Fig. 1). To ensure the quality of data collected in real-world settings, AIS interactively instructed users to follow a standardized preparation sequence for data collection (Extended Data Fig. 2). Before data collection, a short demo video was displayed to instruct users on the standard operation and how to choose an appropriate environment to minimize testing biases (for example, room illuminance, background, testing distance and interference). Once the smartphone was firmly in place, a face-positioning frame was shown on the screen to help adjust the distance and position of the child in relation to the smartphone. After all preparations were completed properly, AIS played a cartoon-like video stimulus lasting approximately 3.5 min to attract children's attention, and the inbuilt front camera recorded the children's phenotypic features (ocular movements and facial appearance) in video format.

Then, the collected data were transferred to the DL-based back end, where the quality control module automatically performed quality checking on each frame first. To eliminate background interference, the children's facial regions were then cropped out of consecutive frames of sufficient quality to form short video clips as inputs of the subsequent DL models for final decision-making (a detection model to distinguish visually impaired children from nonimpaired individuals and a diagnostic model to discriminate multiple ocular disorders). The DL models produced classification probabilities for short video clips, which were eventually merged into the video-level classification probability as the final outcome by averaging. The final results were returned to the mHealth app to alert users to promptly refer children at high risk of visual impairment to experienced pediatric ophthalmologists for further diagnosis and intervention.

Deep quality control module

To ensure prediction reliability, we adopted a strict data quality control strategy to ensure that the input clips of the detection/diagnostic models satisfied certain quality criteria (Fig. 1a). First, for each frame, the child's facial area was detected, and frames without successful face detection were rejected. If two or more faces were detected in a given frame, it suggested that the child's parents or other persons were inside the scene, and such a frame was also rejected. The facial region detection algorithm was based on MMOD CNN[46], which consisted of a series of convolutional layers for feature learning and max-margin operation during model training. In this study, the MMOD CNN face detector pretrained on publicly available datasets was adopted from the Dlib Python Library, which has been proven to be effective and robust in facial detection tasks[26].

Second, a facial key point localization algorithm was applied to the detected facial area to extract the landmarks of facial regions, including the left eye, right eye, nose tip, chin and mouth corners, which served as the reference coordinates for the cropping of facial regions. The facial key point localization algorithm was realized based on a pretrained ensemble of regression trees, which was also provided by the Dlib Python Library[47,48]. We adopted a cascade of regressors to take the facial region of the frame as the input. The network was able to learn coarse-to-fine feature representations of the child's face, especially details of the facial patterns. The output of this model was then fitted to the coordinates representing facial structures to generate 68 key target points. The coordinates of the key points then served as the reference for facial region cropping. All video data and image data processing were performed using the FFmpeg toolkit and OpenCV Python Library[49].

Then, a combination of crying, interference and occlusion classification models based on EfficientNet-B2 networks (Extended Data Fig. 3a,b) was applied to each frame, which was trained based on the data collected at ZOC (Extended Data Fig. 3d and Supplementary Table 1)[50]. During model training and inference, the input frame was first rescaled to 384 × 384 resolution and then sent into the models for deep feature representation learning (Supplementary Table 14). Positive outputs by the models indicated that the child was crying, was interfered with or had its facial region blocked by objects such as toys or other persons' hands, and the corresponding frames were also discarded. In practice, we fine-tuned the models pretrained on the ImageNet public dataset[51].

Eventually, the remaining frames were considered high-quality candidates, and consecutive high-quality frames were selected to form short video clips. Each clip lasted at least 1.5 s and at most 5 s. The child's facial region within each clip was then cropped out to serve as the final input of the subsequent detection/diagnostic models based on the facial key point coordinates to eliminate the interference of the background region. A qualified video should contain more than ten clips; otherwise, the video was treated as a low-quality sample and discarded.

DL framework of the detection/diagnostic models

Two models with various clinical purposes were developed in this study: a detection model to detect visually impaired children from nonimpaired children and a five-category diagnostic model to discriminate specific ophthalmic disorders (aphakia, congenital glaucoma, congenital ptosis and strabismus) and nonimpairment. The backbone of each DL model was built on a deep convolutional network known as EfficientNet-B4 (Extended Data Fig. 3c and Supplementary Table 14)[50]. The models made predictions on the children's cropped facial regions. Specifically, spatial cues of the input clips were learned by cascaded convolutional layers, while temporal cues were integrated by temporal average pooling layers, which was inspired by successful applications in gait recognition[52]. The temporal average pooling operator was given by $\frac{1}{n}\sum_{i=1}^{n}\bar{x_i}$, where n was the number of frames in the input clip and $\bar{x_i}$ was the feature map of each frame output by the last convolutional layer of the network. Before training, all convolutional blocks were initialized by the parameters of the models pretrained on the ImageNet dataset[51]. At the inference stage, class scores given by the models were treated as the final clip-level probability outcomes. For the detection model, the output of the last classification layer, indicated by x_i, was normalized to the range between 0.00 and 1.00 for each clip using the sigmoid function $p_i = \frac{1}{1+\exp(-x_i)}$, representing the final probability of the ith clip being classified as a visually impaired candidate. To train the detection model, the cost function was given by the classic binary cross-entropy loss $L = -\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y_i}\log(p_i) + (1-\hat{y_i})\log(1-p_i)\right)$, where N was the number of clips within each batch, $\hat{y_i}$ was the ground truth label of the ith clip and p_i was the output classification probability of the model.

The diagnostic model was developed based on the same EfficientNet-B4 backbone as the detection model. The only difference was that the output of the diagnostic model was activated by a five-category softmax function that indicated the probability of each class: $p_k = \frac{e^{x_k}}{\sum_{j=1}^{5} e^{x_j}}$, where x_k was the output of the last classification layer for the kth class. The cost function of the network was given by the stochastic cross-entropy loss $L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{5}\left(y_k^i \log\left(p_k^i\right)\right)$ where N was the batch size and $y_k^i \in \{0,1\}$ was the binary Boolean variable of the ith input clip within each batch, indicating whether the kth class matched the ground truth label of the ith clip.

Child-level classification was based on clip-level predictions. A sliding window integrated with the quality control module was applied along the temporal dimension of the whole video to extract high-quality clips. Such clips then served as the candidate inputs for detection/diagnostic models. For the detection model, if the average score of the clips exceeded 0.50 within each video, the child was eventually classified as a visually impaired individual. For the diagnostic model, the category with the highest average probability was treated as the final prediction outcome.

Model training and internal validation

We first developed the data quality control module using both publicly available datasets and the ZOC dataset (Supplementary Table 1). Then, we trained and validated the detection/diagnostic models with the ground truth of the visual conditions using the development dataset at ZOC. In this stage, data collection preceded the development of the quality control module, so raw videos without quality checking were collected. In total, raw videos from 2,632 children undergoing ophthalmic examinations were collected by trained volunteers using the mHealth apps installed on iPhone-7 or iPhone-8 smartphones. After initial quality checking by the quality control module, qualified videos of 2,344 (89.1%) children were reserved as the development dataset, which was randomly split into training, tuning and validation (internal validation) sets using a stratified sampling strategy according to sex, age and the category of ophthalmic disorder to train and internally validate the detection/diagnostic models (Fig. 1b, Extended Data Fig. 3e and Supplementary Table 2). The age distribution and the proportions of children with unilateral and bilateral severe visual impairment for different datasets are shown in Supplementary Tables 15 and 16, respectively. Internal validation refers to the assessment of the performance of the selected optimized model, after training and hyperparameter selection and tuning, on the independent datasets from the same settings as training datasets. The top-performing checkpoint was selected on the basis of accuracy on the tuning set. In particular, the videos utilized for quality control module development did not overlap with those in the detection/diagnostic model validation.

Finding a needle in a haystack test

To estimate the performance of the AIS system in the general population with a rare-case prevalence of visual impairment, we simulated a gradient of prevalences ranging from 0.1% to 9% to conduct a finding a

needle in a haystack test. For each simulated prevalence, we resampled 10,000 children based on the internal validation dataset in a bootstrap manner to test whether the AIS system could pick up the 'needle' (visually impaired children at the simulated prevalence) in the 'haystack' (10,000 resampled children) and repeated this process 100 times to estimate the 95% CIs.

Data augmentation

To ensure better model capacity and reliability in complex environments, data augmentation was performed during model training using brightness and contrast adjustments, together with blurring techniques. Specifically, the brightness of the input frames was randomly adjusted by a factor of 0.40, and the contrast was randomly adjusted by a factor of 0.20. Blurring techniques included Gaussian blur, median blur and motion blur. The factor of all blurring techniques was set to five. Each input frame had a probability of 0.50 to perform data augmentation (Supplementary Table 17). All data augmentation processes were based on a publicly available Python library known as Albumentations[53].

Multicenter external validation

External validation refers to the assessment of the performance of the AI system using independent datasets, captured from different clinical settings. This is to ensure the generalizability of the system to different settings. Trained volunteers used mHealth apps installed on iPhone-7 or iPhone-8 smartphones to perform external validation in the ophthalmology clinics of the Second Affiliated Hospital of Fujian Medical University, Shenzhen Eye Hospital, Liuzhou Maternity and Child Healthcare Hospital and Beijing Children's Hospital of Capital Medical University. In this stage, the quality control module automatically reminded volunteers to repeat data collection when the videos were of low quality. In total, 305 children were recruited and qualified videos for 301 children (98.7%) were successfully collected. Qualified videos for 298 children undergoing ophthalmic examinations were reserved for final validation of the detection model (see Fig. 1b and Table 1 for details of the participants and the dataset used for external validation).

Implementation by untrained parents or caregivers at home

We further challenged our system in an application administered by untrained parents or caregivers with their smartphones in daily routines (Fig. 1b). Children (independent from the development and external validation participants) were recruited online, and their parents or caregivers autonomously used AIS at home according to the system's instructions to collect qualified videos and perform tests without pretraining or controlling any biases before testing, such as brands and models of smartphones and the home environment. This process generated data with huge variations of distributions that had an extremely high requirement of generalizability and extensibility for the DL-based system. Thus, before final implementation, we performed a pilot study to collect a dataset for fine-tuning our system to chaotic home environments. To efficiently evaluate the performance of AIS for identifying visual impairment in at-home settings, a sufficient proportion of visually impaired children with various ocular diseases were recruited. Of the 125 children recruited, 122 children (97.6%) successfully completed the detection tests and collected qualified videos, among whom 120 children undergoing ophthalmic examinations were enrolled to fine-tune and evaluate the detection model. We fine-tuned the detection model using qualified videos from 32 children collected first and then tested it by the subsequently collected validation set from another 88 children. See Fig. 1b and Table 1 for more information on the fine-tuning and implementation.

Reliability analyses and adjusted analyses

To test the stability and generalizability of AIS under various conditions, investigators conducted a batch of reliability analyses and adjusted analyses (Fig. 1b and Table 1).

Reliability across different smartphone platforms

We performed adjustments at different blur, brightness, color or Gaussian noise adjustment gradients to a dataset (n = 200 children and n = 200 qualified videos) randomly sampled from the ZOC validation set to simulate the characteristics of data collected by various cameras and evaluate the reliability of AIS. Furthermore, we collected another dataset in an independent population of children at ZOC to assess the reliability of the AIS system across different operating systems. In total, raw videos from 389 children undergoing ophthalmic examinations were collected by trained volunteers using two Android smartphones, Redmi Note 7 and Huawei Hornor-6 plus. After initial quality checking, qualified videos of 361 (92.8%) children were reserved for testing. The technical specifications of the smartphones used in this study are summarized in Supplementary Table 13.

Retest reliability analysis

We performed detection tests for each child twice by two volunteers at least 1 day apart on another independent population recruited at ZOC to evaluate the retest reliability. Raw videos from 213 children undergoing ophthalmic examinations were collected using iPhone-7 or iPhone-8 smartphones. Qualified videos of 187 (87.8%) children were reserved for retest analysis after initial quality checking (Fig. 1b and Table 1). An intraclass correlation coefficient was calculated for repeated predicted probabilities of the detection model, and a Cohen's κ was calculated for repeated predicted categories to evaluate retest reliability.

Hard-to-spot test

To investigate the influence of the apparency of the phenotypic features on the AIS system, a panel of 14 community ophthalmologists with 3–5 years of clinical experience identified 'likely impaired' children based on the phenotypic videos in the ZOC validation dataset. The true impaired and nonimpaired children were mixed at a ratio of 1:1 during identification. Each case was independently reviewed by three ophthalmologists. When no more than one ophthalmologist provided 'likely impaired' labels for one true impaired child, this child was classified as a hard-to-spot case with insidious phenotypic features rather than a relatively evident case. The performance of the AIS system for relatively evident/hard-to-spot cases was assessed.

Other reliability analyses

We tested AIS under different room illuminance conditions. Photometers (TESTES-1330A; TES Electrical Electronic Corp.) were used to measure the mean room illuminance intensity before and after data collection. The following criteria were applied to estimate the distances between the children and the smartphones to assess the reliability of AIS in different testing distance groups. When most of the vertical lengths of a child's head regions were less than one-third of the height of the smartphone screen at the frame level, the video was determined to be taken from a long distance. When most of the lengths were between one-third and one-half of the height of the screen, the video was judged to be taken from a medium distance, and when most of the lengths were larger than one-half of the height of the screen, the video was judged to be taken at a close distance. For each full-length video, subvideos with various durations were generated to serve as inputs to evaluate the influence of the duration of the video recording on the performance of AIS. We also evaluated the performances of AIS grouped by patient-related factors including sex, age and laterality of the eye disorder.

Adjusted analyses

To further verify that the predictions of this system were not solely mediated by sample characteristics as confounders, we performed adjusted analyses to examine the ORs of the predictions of the system adjusted for sample characteristics leveraging logistic regression models.

Detection model visualization and explanation

Two strategies were used to interpret and visualize the detection model: *t*-distributed stochastic neighbor embedding (*t*-SNE) and gradient-weighted class activation mapping (Grad-CAM)[54–56]. The former was used to visualize the high-dimensional activation status of the deep CNN at the clip level by projecting its feature vector into a two-dimensional space, and the latter was adopted to create a heat map showing the area within each frame of the clip that contributed most to the output class of the network. In practice, the feature vectors output by the temporal average pooling layer and flatten operation and the feature maps output by the last convolutional layer before the temporal average pooling operation were chosen to visualize the results generated by *t*-SNE and Grad-CAM, respectively. Specifically, 1,200 visually impaired clips and 1,200 nonimpaired clips were randomly selected from the ZOC validation set to perform *t*-SNE analysis. To generate average heat maps, we randomly sampled ten videos for each ophthalmic disorder from the internal validation dataset. Since each video had multiple clips, we ranked these clips according to the model predicted probabilities and selected the two clips with the highest probabilities. For each selected clip, we took 30 frames at equal intervals to generate the corresponding average heat map. In summary, we had a total of 600 heat maps for each type of disorder, and we summed and averaged these heat maps to obtain the typical heat map for a certain disease. A public machine learning Python library named Scikit-learn was used to generate two-dimensional coordinates of *t*-SNE results, and Grad-CAM analysis was performed based on an open-source GitHub code set[57].

Additionally, we compared the model-predicted probabilities of three groups of clips (clips randomly sampled from videos of nonimpaired children, clips randomly sampled from videos of visually impaired children, and clips annotated by experts as having abnormal behavioral patterns from videos of visually impaired children) to investigate whether the detection model focused on specific behavioral patterns in children (Fig. 3d and Supplementary Table 8).

Triage-driven approach to select equivocal cases for manual review

We assessed a triage strategy to find a solution when the system was likely unreliable by choosing equivocal cases for manual review in the internal validation set. An equivocal case referred to a child predicted by the AIS system with a low confidence value, given by $|p - 0.50|$, where p was the predicted probability for the child. Three ophthalmologists from ZOC with over 10 years of clinical experience vetted the phenotypic videos of the equivocal cases and the AIS predictions in a voting manner. Additional information, including baseline information and medical histories, was provided when necessary. An increasing ratio from 0 to 19% of equivocal cases with the lowest confidence values was chosen for manual review to evaluate this triage strategy.

Statistical analysis

The primary outcomes were the AUCs of the detection/diagnostic models. The secondary outcomes included the accuracy, sensitivity and specificity of the models and the reliability of the detection model under various settings. The 95% CIs of the AUC, accuracy, sensitivity and specificity of the models were estimated. Specifically, the DeLong CIs of AUCs were calculated at the child level. To eliminate bias due to the association of multiple clips for the same child, the bootstrap CIs of the AUCs of the detection model were calculated at the clip level. One clip for each child was randomly taken to form a bootstrap sample, and this process was repeated 1,000 times. Wilson CIs were reported for other proportional metrics. Descriptive statistics, including means, s.d., numbers and percentages, were used. Mann–Whitney *U*-tests were used to compare means on continuous variables, and Fisher exact tests were used to compare distributions on categorical variables. A two-sided *P* value of <0.05 indicates statistical significance. All statistical analyses were performed in R Statistics (v.4.1.2) or Python Programs (v.3.9.7), and plots were created with the ggplot2 package (v.3.3.5) in R Statistics.

Computational hardware

Hardware information for this study is shown as follows: graphics processing unit (GPU), Nvidia Titan RTX 24 GB memory × 4, Driver v.440.82, Cuda v.10.2; central processing unit (CPU), Intel(R) Xeon(R) CPU E5-2678 v.3 @ 2.50 GHz × 2, 48 threads; random access memory (RAM), Samsung 64 GB RAM × 8, configured speed 2,133 MHz.

Use of human data

The ethical review of this study was approved by the Institutional Review Board/Ethics Committee of ZOC. The test was prospectively registered at ClinicalTrials.gov (identifier: NCT04237350).

Reporting summary

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

Data availability

The data that support the findings of this study are divided into two groups: published data and restricted data. The authors declare that the published data supporting the main results of this study can be obtained within the paper and its Supplementary Information. For research purposes, a representative video deidentified using digital masks on children's faces for each disorder or behavior in this study is available. In the case of noncommercial use, researchers can sign the license, complete a data access form provided at https://github.com/RYL-gif/Data-Availability-for-AIS and contact H.L. Submitted license and data access forms will be evaluated by the data manager. For requests from verified academic researchers, access will be granted within 1 month. Due to portrait rights and patient privacy restrictions, restricted data, including raw videos, are not provided to the public.

Code availability

Since we made use of proprietary libraries in our study, our codes for system development and validation release to the public are therefore not feasible. We detail the methods and experimental protocol in this paper and its Supplementary Information to provide enough information to reproduce the experiment. Several major components of our work are available in open-source repositories: PyTorch (v.1.7.1): https://pytorch.org; Dlib Python Library (v.19.22.1): https://github.com/davisking/dlib (frameworks for facial region detection and facial key point localization); EfficientNet-PyTorch: https://github.com/lukemelas/EfficientNet-PyTorch (frameworks for models in the quality control module and the detection/diagnostic models); Albumentations (v.0.5.2): https://github.com/albumentations-team/albumentations (data augmentation); and OpenCV Python Library (v.4.5.3.56): https://github.com/opencv/opencv-python (video data and image data processing).

References

44. Drover, J. R., Wyatt, L. M., Stager, D. R. & Birch, E. E. The teller acuity cards are effective in detecting amblyopia. *Optom. Vis. Sci.* **86**, 755 (2009).
45. Mayer, D. L. et al. Monocular acuity norms for the Teller Acuity Cards between ages one month and four years. *Investigative Ophthalmol. Vis. Sci.* **36**, 671–685 (1995).
46. King, D. E. Max-margin object detection. Preprint at https://ui.adsabs.harvard.edu/abs/2015arXiv150200046K (2015)..
47. Zhou, E., Fan, H., Cao, Z., Jiang, Y. & Yin, Q. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *2013 IEEE International Conference on Computer Vision Workshops* 386–391 (IEEE, 2013).

48. Kazemi, V. & Sullivan, J. One millisecond face alignment with an ensemble of regression trees. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* 1867–1874 (IEEE, 2014).
49. Bradski, G. The openCV library. *Dr. Dobb's J. Softw. Tools* **25**, 120–123 (2000).
50. Tan, M. & Le, Q. EfficientNet: rethinking model scaling for convolutional neural networks. In *Proc. 36th International Conference on Machine Learning* 6105–6114 (PMLR, 2019).
51. Deng, J. et al. Imagenet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
52. Chao, H., He, Y., Zhang, J. & Feng, J. GaitSet: regarding gait as a set for cross-view gait recognition. In *Proc. Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence* Article 996 (AAAI Press, 2019).
53. Buslaev, A. et al. Albumentations: fast and flexible image augmentations. *Information* **11**, 125 (2020).
54. Hinton, G. E. & Roweis, S. Stochastic Neighbor Embedding. In *Advances in Neural Information Processing Systems 15* (Eds. Becker, S., Thrun, S. and Obermayer, K.) 833–840 (NIPS, 2002).
55. Belkina, A. et al. Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nat. Commun.* **10**, 5415 (2019).
56. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)* 618–626 (IEEE, 2017).
57. Zuppichini, F. S. FrancescoSaverioZuppichini/cnn-visualisations. *GitHub* https://github.com/FrancescoSaverioZuppichini/cnn-visualisations (2018).

Acknowledgements

We thank all the participants and the institutions for supporting this study. We thank H. Sun, T. Wang, T. Li, W. Lai, X. Wang, L. Liu, T. Cui, S. Zhang, Y. Gong, W. Hu, Y. Huang, Y. Pan and C. Lin for supporting the data collection; M. Yang for the help with statistical suggestions and Y. Mu for the help with our demo video. This study was funded by the National Natural Science Foundation of China (grant nos. 82171035 and 91846109 to H.L.), the Science and Technology Planning Projects of Guangdong Province (grant no. 2021B1111610006 to H.L.), the Key-Area Research and Development of Guangdong Province (grant no. 2020B1111190001 to H.L.), the Guangzhou Basic and Applied Basic Research Project (grant no. 2022020328 to H.L.), the China Postdoctoral Science Foundation (grant no. 2022M713589 to W.C.), the Fundamental Research Funds of the State Key Laboratory of Ophthalmology (grant no. 2022QN10 to W.C.) and Hainan Province Clinical Medical Center (H.L.). P.Y.-W.-M. is supported by an Advanced Fellowship Award (NIHR301696) from the UK National Institute of Health Research (NIHR). P.Y.-W.-M. also receives funding from Fight for Sight (UK), the Isaac Newton Trust (UK), Moorfields Eye Charity (GR001376), the Addenbrooke's Charitable Trust, the National Eye Research Centre (UK), the International Foundation for Optic Nerve Disease, the NIHR as part of the Rare Diseases Translational Research Collaboration, the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014) and the NIHR Biomedical Research Centre based at Moorfields Eye Hospital National Health Service Foundation Trust and University College London Institute of Ophthalmology. The views expressed are those of the author(s) and not necessarily those of the National Health Service, the NIHR or the Department of Health. The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

W.C., R.L. and H.L. contributed to the concept of the study and designed the research. W.C., R.L., A.X., Ruixin Wang, Yahan Yang, D. Lin, X.W., J.C., Z. Liu, Y.W., K.Q., Z.Z., D. Liu, Q.W., Y.X., X.L., Zhuoling Lin, D.Z., Y.H., S.M., X.H., S.S., J.H., J.Z., M.W., S.H., L.C., B.D., H.Y., D.H., X.L., L.L., Xiaoyan Ding, Yangfan Yang and P.W. collected the data. W.C., R.L., Q.Y., Y.F., Zhenzhe Lin, K.D., Z.W., M.L. and Xiaowei Ding conducted the study. W.C., R.L. and L.Z. analyzed the data. W.C., R.L., Q.Y., Y.F. and H.L. cowrote the manuscript. D. Lin, X.W., F.Z., N.S., J.-P.O.L., C.Y.C., E.L., C.C., Y.Z., P.Y.-W.-M., Ruixuan Wang and W.-s.Z. critically revised the manuscript. Zhenzhe Lin, Ruixuan Wang, W.-s.Z, Xiaowei Ding and H.L. performed the technical review. All authors discussed the results and provided comments regarding the manuscript.

Competing interests

Zhongshan Ophthalmic Center and VoxelCloud have filed for patent protection for W.C., R.L., A.X., Y.F., Zhenzhe Lin, K.D., K.Q., Xiaowei Ding and H.L. for work related to the methods of detection of visual impairment in young children. All other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-022-02180-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-022-02180-9.

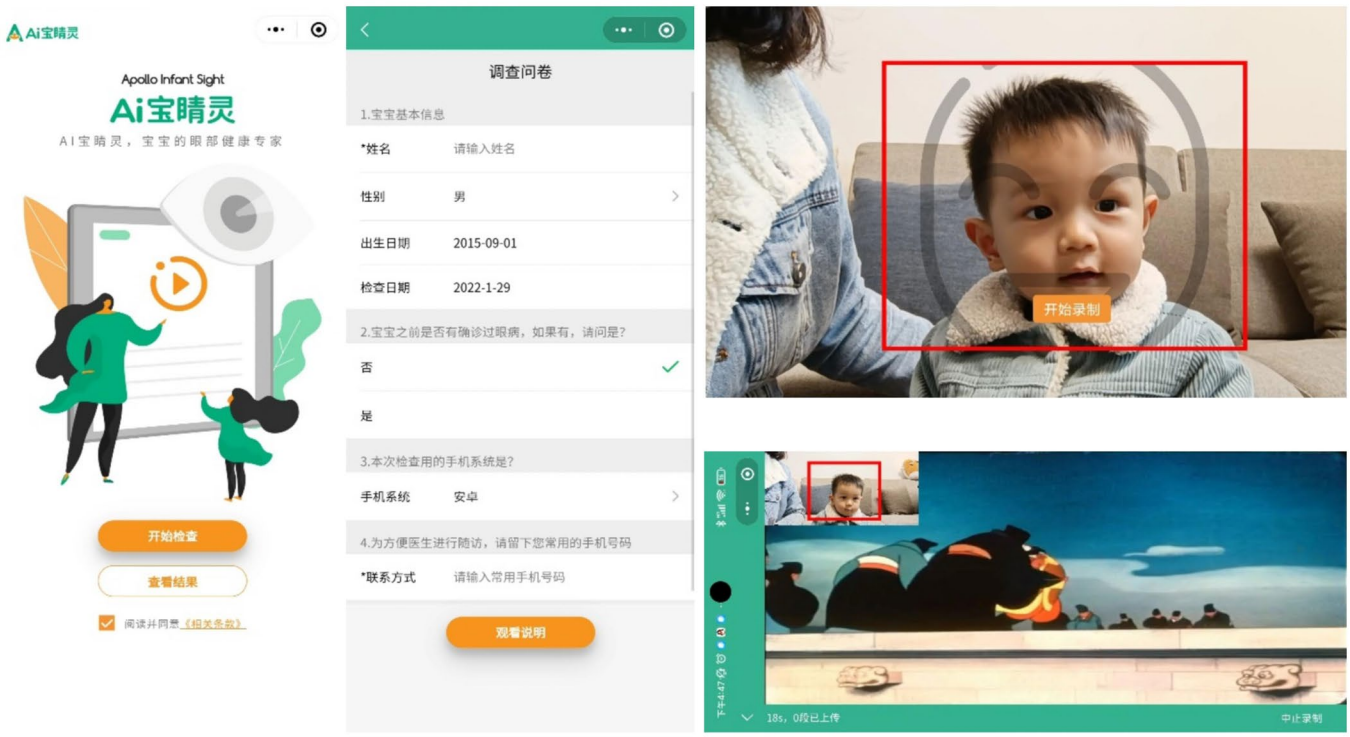
Correspondence and requests for materials should be addressed to Xiaowei Ding or Haotian Lin.

Peer review information *Nature Medicine* thanks Pete Jones, Ameenat Lola Solebo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Michael Basson, in collaboration with the *Nature Medicine* team.

Reprints and permissions information is available at www.nature.com/reprints.

https://doi.org/10.1038/s41591-022-02180-9

a



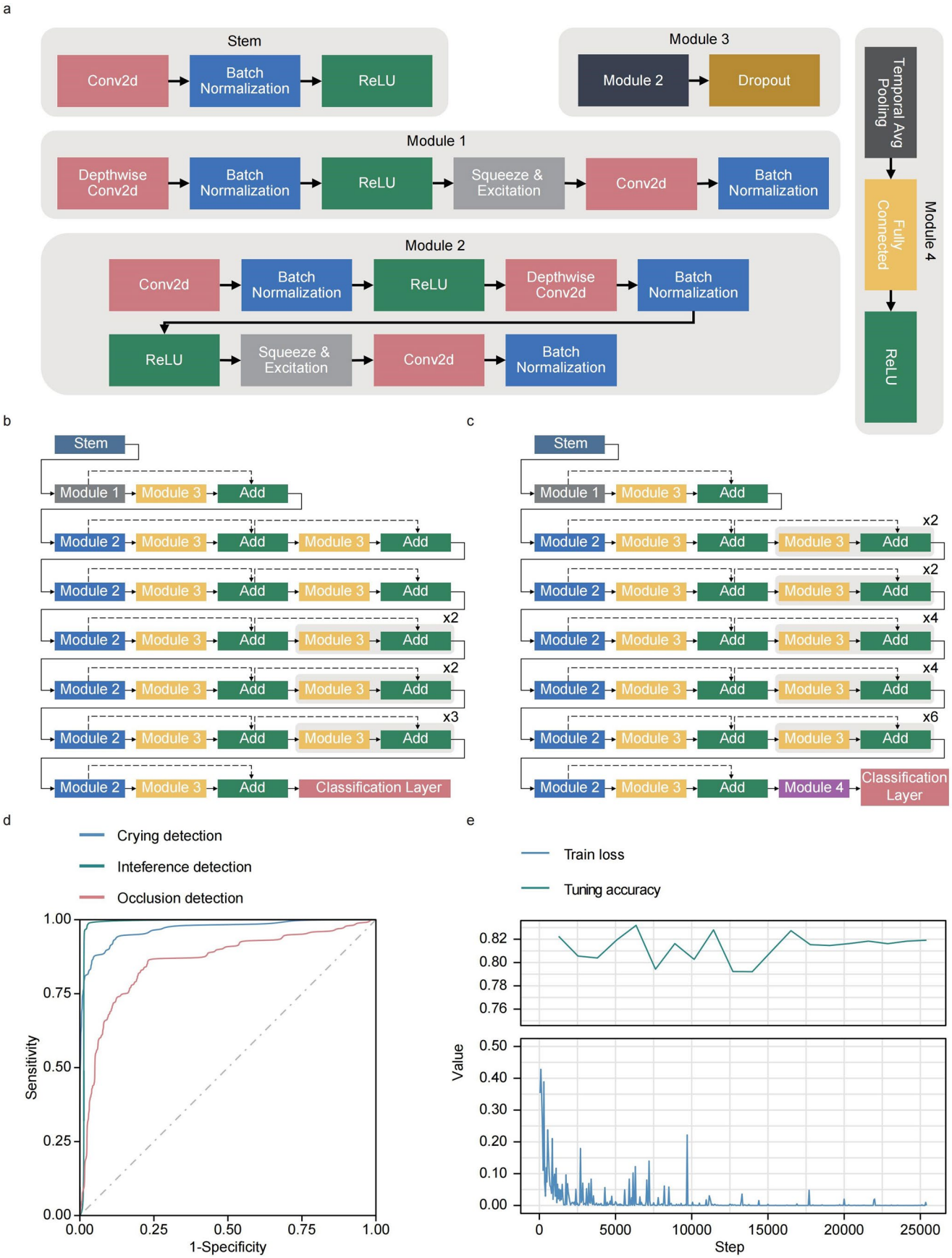
b

Extended Data Fig. 1 | The app for data collection. a, The operation interface of the app. **b**, Utilize the smartphone for data collection in real-world settings.

https://doi.org/10.1038/s41591-022-02180-9

Extended Data Fig. 2 | The standard preparation sequence guided by the app for data collection.

https://doi.org/10.1038/s41591-022-02180-9

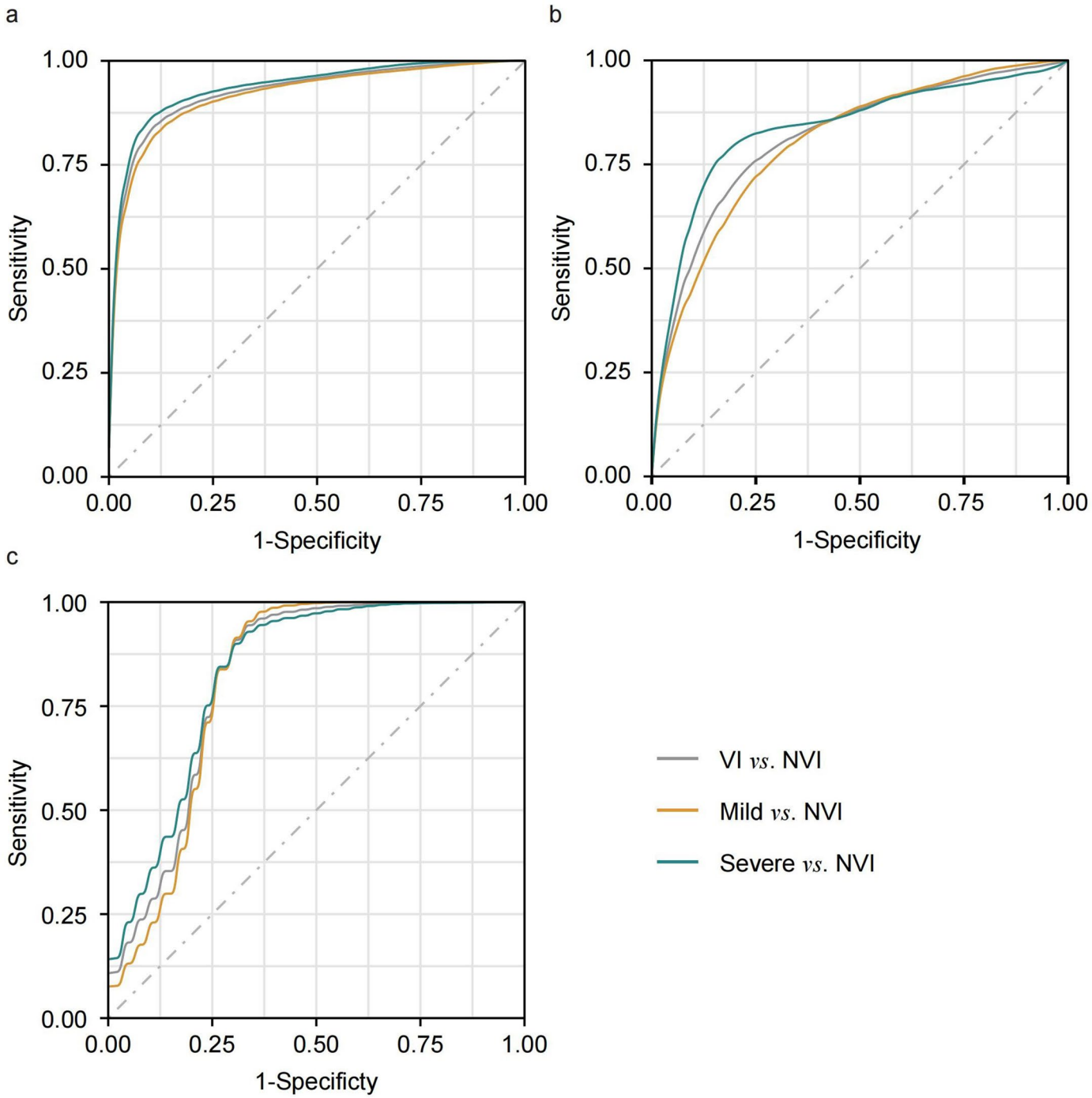


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41591-022-02180-9

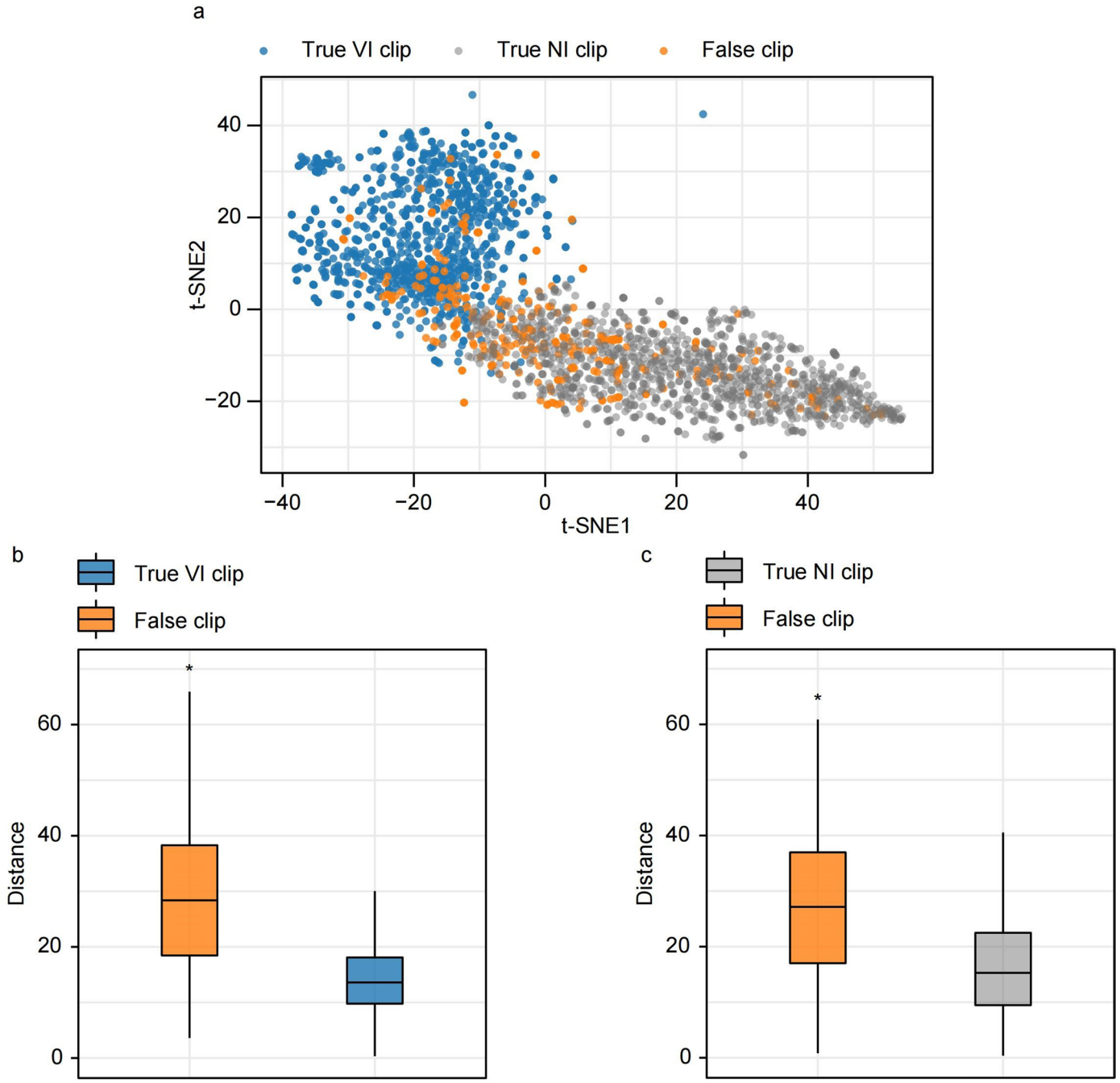
Extended Data Fig. 3 | Development of deep learning models of the AIS system. a, Basic building blocks and architecture of EfficientNet. Two model architectures, EfficientNet-B2 and EfficientNet-B4, were used in data quality control for detection/diagnostic tasks, respectively. **b**, Architecture of the EfficieNet-B2 model. **c**, Architecture of the EfficientNet-B4 model. **d**, ROC curves of the models trained for the quality control module. **e**, The training and tuning curves of the detection model at the clip level. Conv 2d, 2-dimensional convolutional layer; ReLU, rectified linear unit; Temporal Avg Pooling, average pooling along the temporal dimension; ROC curve, receiver operating characteristic curve; AIS, Apollo Infant Sight.

https://doi.org/10.1038/s41591-022-02180-9



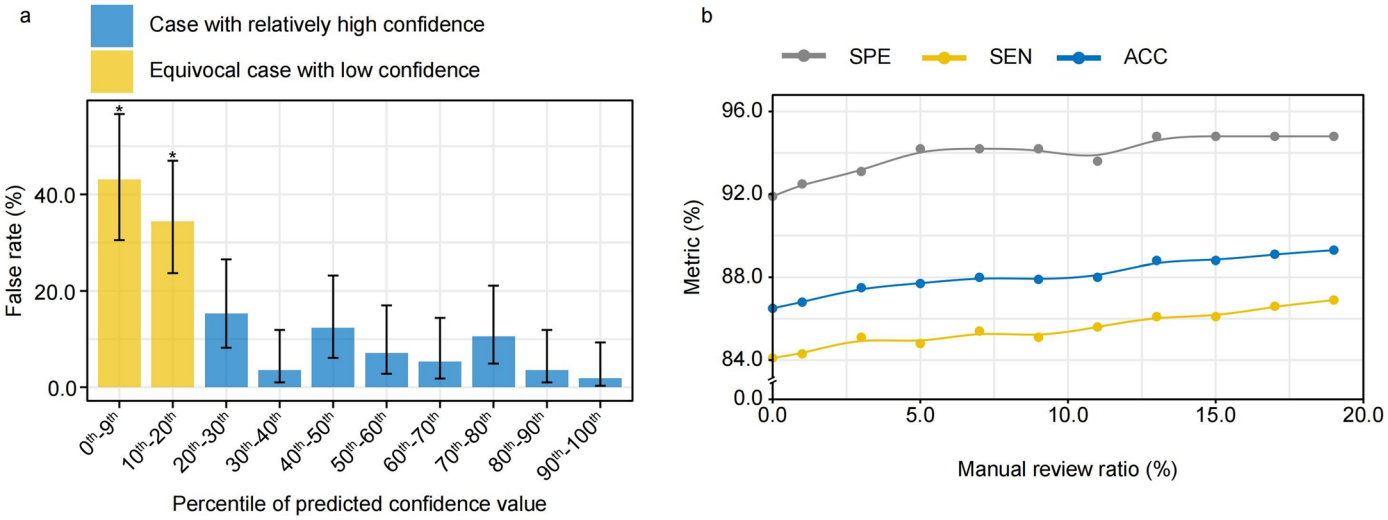
Extended Data Fig. 4 | Performance of the detection model at the clip level. **a**, ROC curves of the detection model in the internal validation (NI, n = 6,735; mild, n = 8,310; severe, n = 6,685; VI versus NI, AUC = 0.925 (0.914–0.936); mild versus NI, AUC = 0.916 (0.904–0.928); severe versus NI, AUC = 0.935 (0.924–0.946)). **b**, ROC curves of the detection model in the external validation (NI, n = 7,392; mild, n = 2,580; severe, n = 1,569; VI versus NI, AUC = 0.814 (0.790–0.838); mild versus NI, AUC = 0.802 (0.770–0.831); severe versus NI, AUC = 0.834 (0.807–0.863)). **c**, ROC curves of the detection model in the at-home implementation by parents or caregivers (NI, n = 947; mild, n = 943; severe, n = 809; VI versus NI, AUC = 0.817 (0.756–0.881); mild versus NI, AUC = 0.809 (0.735–0.884); severe versus NI, AUC = 0.825 (0.764–0.886)). Parentheses show 95% bootstrap CIs. A cluster-bootstrap biased-corrected 95% CI was computed, with individual children as the bootstrap sampling clusters. NI, nonimpairment; VI, visual impairment; ROC curve, receiver operating characteristic curve; AUC, area under the curve; CI, confidence interval.

https://doi.org/10.1038/s41591-022-02180-9



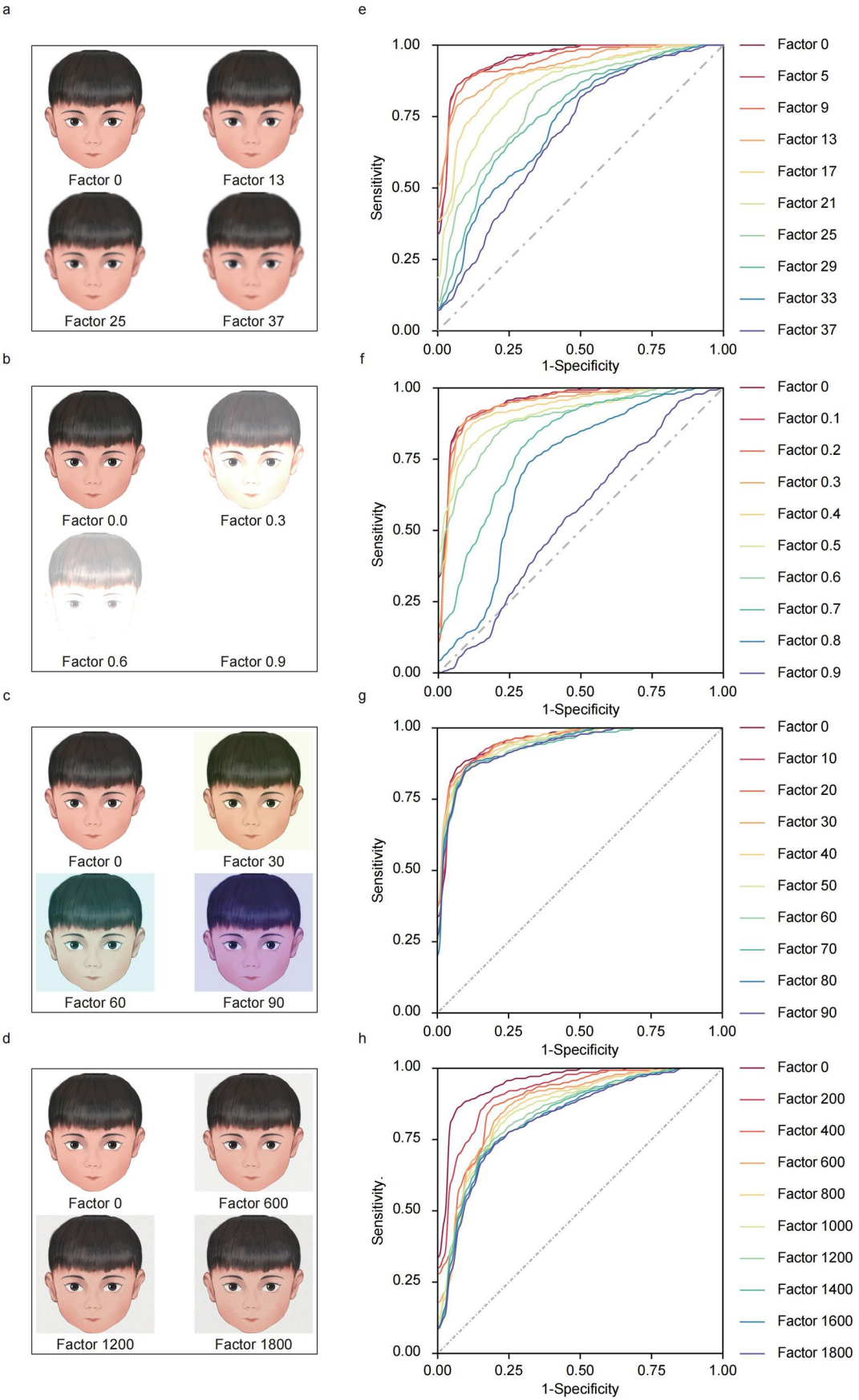
Extended Data Fig. 5 | Visualization of the clips correctly classified or misclassified by the detection model. a, The t-distributed stochastic neighbor embedding (t-SNE) algorithm was applied to visualize the clustering patterns of clips correctly classified or misclassified by the detection model. **b**, Distances from true VI and false clips to the center of true VI clips in the t-SNE scatter plot were compared. *$P < 0.001$ (true VI clip, n = 999; false clip, n = 317; $P < 1.00 \times 10^{-36}$, two-tailed Mann-Whitney U test) **c**, Distances from true NI and false clips to the center of true NI clips in the t-SNE scatter plot were compared. *$P < 0.001$ (true NI clip, n = 1084; false clip, n = 317; $P < 1.00 \times 10^{-36}$, two-tailed Mann-Whitney U test). The thick central lines denote the medians, the lower and upper box limits denote the first and third quartiles, and the whiskers extend from the box to the outermost extreme value but no further than 1.5 times the interquartile range (IQR). VI, visual impairment; NI, nonimpairment.

https://doi.org/10.1038/s41591-022-02180-9



Extended Data Fig. 6 | The triage-driven approach to select the equivocal cases with the lowest predicted confidence values for manual review. a, The false predicted rate (both false positive and false negative) in different percentile intervals of predicted confidence values. *$P < 0.001$ (0^{th}–9^{th}, n = 51; 10^{th}–20^{th}, n = 61; 20^{th}–30^{th}, n = 59; 30^{th}–40^{th}, n = 57; 40^{th}–50^{th}, n = 57; 50^{th}–60^{th}, n = 56; 60^{th}–70^{th}, n = 57; 70^{th}–80^{th}, n = 57; 80^{th}–90^{th}, n = 57; 90^{th}–100^{th}, n = 57; 0^{th}–9^{th} percentile versus other percentile intervals, P ranging from 7.92×10^{-8} for 90^{th}–100^{th} to 1.45×10^{-3} for 20^{th}–30^{th}; 10^{th}–20^{th} percentile versus other percentile intervals, P ranging from 2.02×10^{-6} for 90^{th}–100^{th} to 2.02×10^{-2} for 20^{th}–30^{th}; two-tailed Fisher's exact tests). Results are expressed as means and the 95% Wilson confidence intervals (CIs). **b**, The performance of the triage-driven system with increasing manual review ratios for the equivocal cases. SPE, specificity; SEN, sensitivity; ACC, accuracy.

https://doi.org/10.1038/s41591-022-02180-9



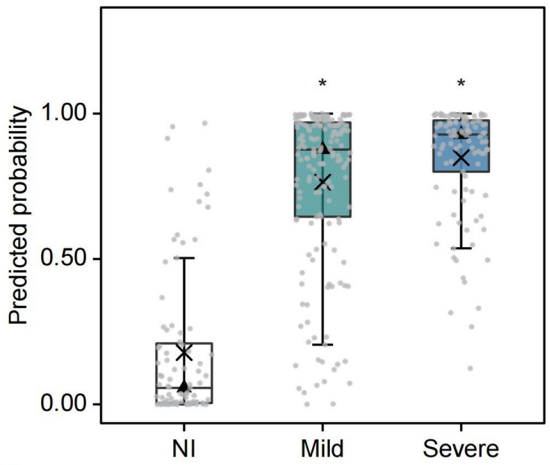
Extended Data Fig. 7 | See next page for caption.

https://doi.org/10.1038/s41591-022-02180-9

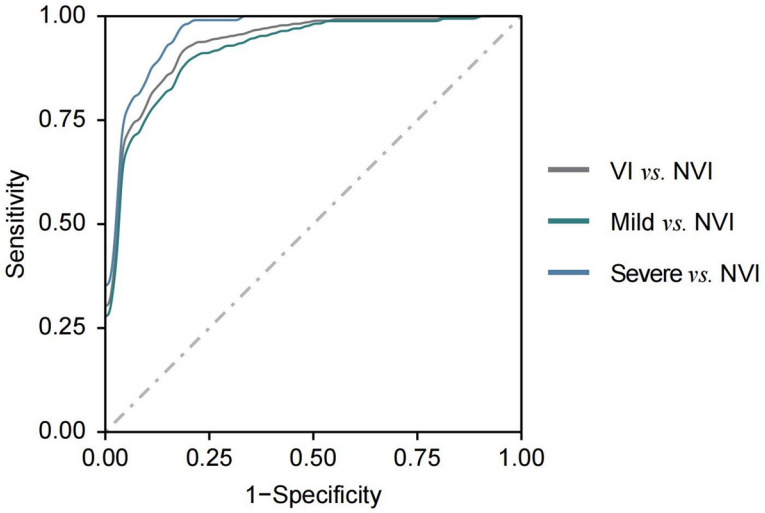
Extended Data Fig. 7 | Performance of the detection model under blurring, brightness, color, and noise adjustment gradients. a, Cartoon diagram showing adjusting effect on the input data by blurring factors. **b**, Cartoon diagram showing adjusting effect on the input data by brightness factors. **c**, Cartoon diagram showing adjusting effect on the input data by color factors. **d**, Cartoon diagram showing adjusting effect on the input data by noise factors. **e**, ROC curves of the detection model for identifying visual impairment change by blurring factors (AUCs range from 0.683 for factor 37 to 0.951 for factor 0). **f**, ROC curves of the detection model for identifying visual impairment change by brightness factors (AUCs range from 0.551 for factor 0.9 to 0.951 for factor 0). **g**, ROC curves of the detection model for identifying visual impairment change by color factors (AUCs range from 0.930 for factor 70 to 0.952 for factor 20). **h**, ROC curves of the detection model for identifying visual impairment change by noise factors (AUCs range from 0.820 for factor 1800 to 0.951 for factor 0). NI, n = 60; VI, n = 140; ROC curve, receiver operating characteristic curve; VI, visual impairment; NI, nonimpairment.

https://doi.org/10.1038/s41591-022-02180-9

a



b



c

	No. of children (n)		Metric (95%CI)			
	Total	Positive	AUC	Accuracy (%)	Sensitivity (%)	Specificity (%)
VI *vs.* NI	361	274	0.932 (0.902-0.963)	85.3 (81.3-88.6)	85.8 (81.1-89.4)	83.9 (74.8-90.2)
Mild *vs.* NI	256	169	0.917 (0.881-0.953)	82.4 (77.3-86.6)	81.7 (75.1-86.8)	83.9 (74.8-90.2)
Severe *vs.* NI	192	105	0.956 (0.929-0.984)	88.5 (83.3-92.3)	92.4 (85.7-96.1)	83.9 (74.8-90.2)

Extended Data Fig. 8 | Performance of the AIS system using Huawei Honor-6 Plus/Redmi Note-7 smartphones. a, Comparisons of the predicted probabilities for the AIS system between the nonimpairment, mild impairment, and severe impairment groups. $^{*}P < 0.001$ (NI versus mild, $P = 8.10 \times 10^{-28}$; NI versus severe, $P = 1.51 \times 10^{-27}$; two-tailed Mann-Whitney U tests). The cross symbols denote the means, the thick central lines and triangle symbols denote the medians, the lower and upper box limits denote the first and third quartiles, and the whiskers extend from the box to the outermost extreme value but no further than 1.5 times the interquartile range (IQR). **b**, ROC curves of the AIS system with Android smartphones. **c**, Performance of the AIS system in the across-smartphone analysis. VI, visual impairment; NI, nonimpairment; ROC curve, receiver operating characteristic curve; AIS, Apollo Infant Sight.

nature portfolio

Corresponding author(s): HL and XD

Last updated by author(s): Nov 14, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data were collected through the AIS app (version 1.0.1) tailored to the present study.
Data analysis	Data were analyzed using R Statistics (version 4.1.2) or Python Programs (version 3.9.7), and plots were created with the ggplot2 package (version 3.3.5) in R Statistics. Several major components of the codes for system development in our work are available in open source repositories: PyTorch (version 1.7.1); Dlib Python Library (version 19.22.1); EfficientNet-PyTorch; Albumentations (version 0.5.2); OpenCV Python Library (version 4.5.3.56).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data that support the findings of this study are divided into two groups: published data and restricted data. The authors declare that the published data supporting the main results of this study can be obtained within the paper and its Supplementary Information. For research purposes, a representative video de-identified using digital masks on children's faces for each disorder or behavior in this study is available. In the case of non-commercial use, researchers can sign the license and complete a data access form provided at https://github.com/RYL-gif/Data-Availability-for-AIS, and contact H.L.. Submitted license and data access forms will be evaluated by the data manager. For requests from verified academic researchers, access will be granted within one month. Due to portrait rights and patient privacy restrictions, restricted data, including raw videos, were not provided to the public.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	The children were recruited without sex restriction. The baseline information of sex was obtained from Chinese-government-issued official Resident Identity Card. A total of 3,652 children were included in this study (1991 males, 1661 females). We evaluated the performance of the AIS system stratified by sex for reliability analysis.
Population characteristics	All the children were no more than 48 months. Children having central nervous system diseases, mental illnesses or other known illnesses that could affect their behavioral patterns, in the absence of ocular manifestations, were not included. Children who could not cooperate to complete the ophthalmic examinations or the detection test using AIS were excluded. Children who had received ocular interventions and treatments in the month immediately preceding data collection were also excluded.
Recruitment	All 3,652 children were recruited from January 14, 2020 to January 30, 2022. Major eligibility criteria included an age of 48 months or younger and informed written consent obtained from at least one parent or guardian of each child. We did not include children having central nervous system diseases, mental illnesses or other known illnesses that could affect their behavioral patterns, in the absence of ocular manifestations. Children who could not cooperate to complete the ophthalmic examinations or the detection test using AIS were excluded. We also excluded children who had received ocular interventions and treatments in the month immediately preceding data collection. In the first stage completed from January 14, 2020, to September 15, 2021, children were enrolled at the clinic of Zhongshan Ophthalmic Center. In the second stage, which occurred from September 22, 2021, to November 19, 2021, children were enrolled at the clinics of the Second Affiliated Hospital of Fujian Medical University, Shenzhen Eye Hospital, Liuzhou Maternity and Child Healthcare Hospital, and Beijing Children's Hospital of Capital Medical University. In the third stage conducted from November 24, 2021, to January 30, 2022, we advertised our study through the online platform of Pediatric Department of Zhongshan Ophthalmic Center and social media of WeChat. We recruited children and their parents or caregivers online from the Guangdong area for at-home implementation. The investigators recruited the children following the same eligibility criteria as the previous two stages by collecting their basic information and medical history online. In addition, children who could not come to Zhongshan Ophthalmic Center for an ophthalmic assessment or who had been included in other stages of this study were excluded.
Ethics oversight	The predefined protocol of the clinical study was approved by the Institutional Review Board/Ethics Committee of the ZOC and prospectively registered at ClinicalTrials.gov (identifier: NCT04237350) and is shown in the Supplementary Note. Consent was obtained from all individuals whose eyes or faces are shown in the figures or video for publication. Before data collection, informed written consent was obtained from at least one parent or guardian of each child. The investigators followed the requirements of the Declaration of Helsinki throughout the study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We used frames of phenotypic videos to discriminate visual impairment in young children. Referring to the previous study by Long et al. (doi:

Sample size	10.1038/s41551-019-0461-9) and considering abundant cases for deep learning model development and validation, an appropriate sample size was obtained.
Data exclusions	Children having central nervous system diseases, mental illnesses or other known illnesses that could affect their behavioral patterns, in the absence of ocular manifestations, were not included. Children who could not cooperate to complete the ophthalmic examinations or the detection test using AIS were excluded. Children who had received ocular interventions and treatments in the month immediately preceding data collection were also excluded. In addition, unqualified data from 342 children (368 videos) were excluded after quality assessment.
Replication	Our findings of performance of our AIS system were derived from the artificial intelligence algorithms and statistical analysis.
Randomization	For development of the detection/diagnostic models, we split the development dataset collected in Zhongshan Ophthalmic Center into training, tuning, and validation sets using a stratified random sampling strategy according to sex, age and the category of ophthalmic condition.
Blinding	The clinical staff who were responsible for eye examinations were blinded to the group assignment.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	This study was registered at ClinicalTrials.gov (identifier: NCT04237350).
Study protocol	The study protocol was available at ClinicalTrials.gov (identifier: NCT04237350) and was provided as Supplementary Note.
Data collection	The phenotypic videos of 3,652 children were colleted from January 14, 2020 to January 30, 2022, in the ophthalmology clinics of Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Province, the Second Affiliated Hospital of Fujian Medical University, Fujian Province, Shenzhen Eye Hospital, Guangdong Province, Liuzhou Maternity and Child Healthcare Hospital, Guangxi Province and Beijing Children's Hospital, Capital Medical University, Beijing City, and at-home environment from Guangdong area.
Outcomes	The primary outcomes were the AUCs of the detection/diagnostic models. The secondary outcomes included the accuracy, sensitivity and specificity of the models and the reliability of the detection model under various settings.