Ophthalmol Ther
https://doi.org/10.1007/s40123-026-01482-2

ORIGINAL RESEARCH

Machine Learning-Based Prediction of Cycloplegic Refraction Using the Eyerobo Vision Screener: Design—A Cross-Sectional Comparative Device Study

Mengmeng Xia · Boxue Yao · Yumei Wang · Fengwei Liang · Wuxia Zhao · Li Qin · Shoukuan Liu · Muhammad Usama · Zhihui Zhang · Wu Guo · Emmanuel Eric Pazo · Xinjun Ren

Received: April 28, 2026 / Accepted: August 6, 2026

ABSTRACT

Introduction: A cross-sectional comparative device study to compare the Eyerobo Vision Screener (VS), a portable handheld photorefractor, against a conventional auto-refractor for machine-learning-based prediction of cycloplegic spherical equivalent (SE) and power vector components (J0, J45) in a pediatric myopia cohort, to evaluate whether supplementary IOL Master biometry narrows the performance gap between devices, and to report screening-oriented classification metrics (sensitivity, specificity, positive and negative predictive values, area under the receiver operating characteristic curve [AUC]) at clinically meaningful referral thresholds that quantify the device–model combination's triage performance.

Methods: Data from 1129 eyes of 574 patients aged 6–18 years were collected using three ophthalmic devices before and after pharmacological dilation. All refractive measurements were decomposed into power vectors (SE, J0, J45) following Thibos et al. Twelve clinically motivated feature scenarios were constructed from a symmetric framework comparing the Eyerobo VS and auto-refractor (each under predilation, post-dilation, and combined conditions), with and without IOL Master biometry. TabPFN, a tabular foundation model requiring no hyperparameter tuning, was evaluated alongside seven conventional machine-learning algorithms using patient-level five-fold grouped cross-validation repeated over three random

First co-authors and equal contribution: Mengmeng Xia and Boxue Yao.

Prior Presentation: A companion screening-classification analysis of the same cohort, focused on the binary referral decision rather than on regression-MAE prediction, has been prepared for submission as a structured abstract to the 30th Congress of the Chinese Ophthalmological Society (CCOS 2026). No part of the present manuscript has been previously published.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40123-026-01482-2.

M. Xia · B. Yao · Y. Wang · F. Liang · W. Zhao · L. Qin · W. Guo
Gaobeidian Mingren Eye Hospital, Gaobeidian 074000, Hebei, China

S. Liu · Z. Zhang · E. E. Pazo · X. Ren (✉)
Tianjin Key Laboratory of Retinal Functions and Diseases, Tianjin Branch of National Clinical Research Centre for Ocular Disease, Eye Institute and School of Optometry, Tianjin Medical University Eye Hospital, Tianjin 300384, China
e-mail: xin2023ren@tmu.edu.cn

M. Usama
School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea

Published online: 25 August 2026

seeds. Agreement was assessed using Bland–Altman analysis, Pearson correlation, and clinical threshold analysis. In addition to regression metrics, a screening-classification analysis was performed at three prespecified clinical thresholds (cycloplegic SE ≤ −0.50 D, ≤ −3.00 D, and ≤ −6.00 D), reporting sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and AUC with 95% paired bootstrap confidence intervals. Formal paired statistical comparison between TabPFN and the next-best algorithm and a one-eye-per-patient sensitivity analysis are also reported.

***Results*:** TabPFN achieved the lowest mean absolute error (MAE) on all Eyerobo scenarios and was competitive with Ridge on auto-refractor scenarios; on the headline Eyerobo Pre + IOL scenario, the per-eye paired difference in MAE between TabPFN (0.425 D) and the next-best gradient-boosting learner (XGBoost, 0.542 D) was 0.117 D in favour of TabPFN, with paired Wilcoxon signed-rank $p < 0.001$. For predilation SE prediction, the auto-refractor achieved an MAE of 0.295 D (87.4% within ± 0.50 D), compared with 0.657 D (61.6%) for the Eyerobo VS. Adding IOL Master biometry to the Eyerobo reduced this gap by 64%, yielding an MAE of 0.425 D (72.0%). For astigmatic power vector components, the gap narrowed further: J0 MAE was 0.106 D (auto-refractor) versus 0.143 D (Eyerobo + IOL), and J45 MAE was 0.060 D versus 0.088 D. Auto-refractor post-dilation predictions approached cycloplegic accuracy (MAE = 0.203 D, 96.0% within ± 0.50 D). Subgroup analysis showed the Eyerobo performed best for mild myopia (MAE = 0.398 D, 75.6% within ± 0.50 D) and degraded for high myopia and the small nonmyopic stratum. In the screening-classification analysis, the Eyerobo and the auto-refractor were operationally equivalent at the any-myopia threshold (Eyerobo AUC 0.964 [95% CI 0.939–0.982], sensitivity 0.946, specificity 0.894; auto-refractor AUC 0.969 [0.948–0.986], sensitivity 0.960, specificity 0.886; paired AUC difference 0.005 [95% CI −0.022 to +0.034], formally noninferior under a prespecified $\delta = 0.05$ margin).

***Conclusions*:** Within this single-center pediatric referral cohort, the Eyerobo VS combined with machine learning produced cycloplegic SE estimates of a precision consistent with use as a triage adjunct to identify children who should proceed to a full cycloplegic examination, with near-equivalent performance to the auto-refractor for astigmatic components. The addition of IOL Master biometry narrowed the device gap by 64%. TabPFN achieved the best performance among the evaluated algorithms within this cohort and requires no hyperparameter tuning, but the modest margin over linear and gradient-boosting baselines and the absence of external validation preclude any recommendation as the definitive algorithm of choice. The proposed approach is not a substitute for cycloplegic refraction; prospective external validation in a general pediatric screening cohort is required before clinical deployment.

Video abstract. See video abstract at https://youtu.be/wy7_Fw5Dnw4 or in the online or HTML version of the manuscript. (MP4 4546 KB)

Keywords: Cycloplegic refraction; Eyerobo vision screener; Auto-refractor; Machine learning; TabPFN; Myopia; Pediatric; Power vectors

DIGITAL FEATURES

This article is published with digital features, including a video abstract, to facilitate understanding of the article. To view digital features for this article, go to https://doi.org/10.6084/m9.figshare.33179303.

Key Summary Points

Why carry out this study?

Cycloplegic refraction is the gold standard for pediatric refractive assessment but requires pharmacological eye drops that limit throughput and patient compliance, motivating research on machine-learning (ML) correction of noncycloplegic measurements.

All previously published ML-based cycloplegic-prediction studies used desk-mounted laboratory-grade auto-refractors as the input device; portable handheld photorefractors, which are the only refractive instruments available in many school and community screening settings, had not been evaluated.

We asked how the portable Eyerobo Vision Screener compares with a conventional desk-mounted auto-refractor for ML-based cycloplegic prediction, whether IOL Master biometry narrows the device gap, and how the device–model combination performs as a triage adjunct at clinically meaningful referral thresholds.

What was learned from the study?

Within a single-center cohort of 1129 eyes from 574 pediatric patients, the desk-mounted auto-refractor produced more accurate predilation cycloplegic SE estimates (TabPFN MAE 0.295 D, 87.4% within ±0.50 D) than the Eyerobo (0.657 D, 61.6%); adding IOL Master biometry to the Eyerobo reduced this gap by 64% (Eyerobo+IOL MAE 0.425 D).

For the screening-classification endpoint at the any-myopia threshold (≤ -0.50 D), the two devices were operationally equivalent: the paired difference in area under the receiver operating characteristic curve was 0.005 (95% CI −0.022 to +0.034), formally noninferior under a prespecified $\delta = 0.05$ margin, with Eyerobo sensitivity 0.946 and specificity 0.894.

Performance was degraded for hyperopic and high-myopic strata and for emmetropic eyes, reflecting both the cohort's myopia-dominant spectrum bias and the device's measurement range; the proposed triage rule should therefore be paired with age-appropriate visual-acuity testing and is not a substitute for cycloplegic refraction.

Findings are based on internal cross-validation only; prospective external validation in a general-school cohort is required before clinical deployment.

INTRODUCTION

Myopia is the most prevalent ocular condition worldwide, with projections estimating that 4.76 billion individuals will be affected by 2050 [1]. In children, early onset and rapid progression raise the risk of high myopia and associated sight-threatening complications including myopic macular degeneration, retinal detachment, and glaucoma [2, 3]. Early detection is critical because effective interventions exist to slow myopia progression when initiated before axial elongation becomes irreversible [4]. Accurate measurement of refractive error is fundamental to myopia management, yet the clinical gold standard, cycloplegic refraction, requires instillation of pharmacological agents to paralyze the ciliary muscle before measurement [5]. The drops cause pupil dilation, blurred near vision, photophobia, and occasional systemic reactions, reducing patient compliance and clinical throughput [6, 7]. Noncycloplegic autorefraction avoids these problems but systematically overestimates myopia in children, with discrepancies of 0.3 to 0.9 diopters (D) depending on age and instrument [8–11]. Because this accommodative bias varies nonlinearly with age, refractive state, and device characteristics, simple correction factors are insufficient [12].

Several groups have demonstrated that machine learning (ML) can partially correct the discrepancy between noncycloplegic and cycloplegic measurements. Du et al. [13] were among the first to predict the spherical equivalent (SE) difference using ML, reporting results on 5920 eyes. Liu et al. [14] reported a mean absolute error (MAE) of 0.360 D using stacked ensembles with biometric features. Su et al. [15] achieved 0.307 D with XGBoost and lens curvature data. Ying et al. [16] obtained R^2 of 0.913–0.935 using XGBoost and random forest. Chen et al. [17] and Hao et al. [18] reported similar results on other parameter sets, and Kim et al. [19] showed ensemble methods generalize to partial interferometry data. Chandra et al. [20] provided the first external validation of these models. However, all existing studies use laboratory-grade auto-refractors. No study has evaluated portable handheld photorefractors for ML-based cycloplegic prediction, and the potential of artificial intelligence for myopia screening and management remains incompletely explored [21].

The Eyerobo Vision Screener (VS) is a handheld binocular photorefractor that provides spherical power (SPH), cylindrical power (CYL), and cylinder axis (AX) at approximately one meter working distance. Its predecessor, the 2WIN-S, was validated by Liu et al. [22] against cycloplegic retinoscopy in 194 eyes of 97 children, achieving Pearson $r = 0.875$ and intraclass correlation coefficient (ICC) of 0.900 for SE agreement. Other portable screeners have been similarly validated: the Plusoptix A09 showed moderate agreement with cycloplegic refraction in preschoolers [23, 24], and the Spot Vision Screener demonstrated effective amblyopia risk detection [25]. The Eyerobo VS features updated software and sensors, and its portability and low cost make it attractive for screening in schools and community settings where auto-refractors are unavailable. The broader Eyerobo platform has recently been extended to fundus imaging, with the Eyerobo FC portable fundus camera validated for AI-assisted diabetic retinopathy screening using deep learning [26], demonstrating the device family's expanding role in portable ophthalmic diagnostics. Whether ML can close the precision gap between this affordable device and the more expensive auto-refractor for refractive screening is an open question.

TabPFN (Tabular Prior-data Fitted Network) is a transformer-based model pretrained on millions of synthetic tabular datasets to approximate Bayesian inference [27, 28]. It requires zero hyperparameter tuning, handles missing values natively, and has demonstrated strong performance on datasets of up to 10,000 samples, making it well-suited for clinical datasets. TabPFN has not previously been applied to ophthalmology cycloplegic prediction.

The present study addresses three questions. First, how does the Eyerobo VS compare to a conventional auto-refractor for ML-based cycloplegic prediction across spherical equivalent and power vector components (J0, J45)? Second, does supplementary IOL Master biometry narrow the performance gap between devices? Third, does TabPFN outperform conventional ML algorithms for this task? We evaluate these questions through a symmetric 12-scenario framework using 1129 eyes from 574 pediatric patients. Figure 1 provides an overview of the complete study pipeline.

RELATED WORK

Cycloplegic Versus Noncycloplegic Agreement

The discrepancy between noncycloplegic and cycloplegic refraction in children has been extensively documented. Wilson et al. [8] conducted a systematic review and meta-analysis of 44 studies, reporting that noncycloplegic autorefraction overestimates myopia by 0.3 to 0.9 D relative to the cycloplegic reference, with the magnitude varying by age, instrument, and refractive state. Guo et al. [9] confirmed this pattern in a large Chicago school-aged sample, demonstrating that the discrepancy was greatest among younger children and those with hyperopia, where accommodative tone is highest. Lin et al. [10] reported similar findings in Beijing urban children, further establishing that the cycloplegic–noncycloplegic difference is associated with subsequent myopia progression,

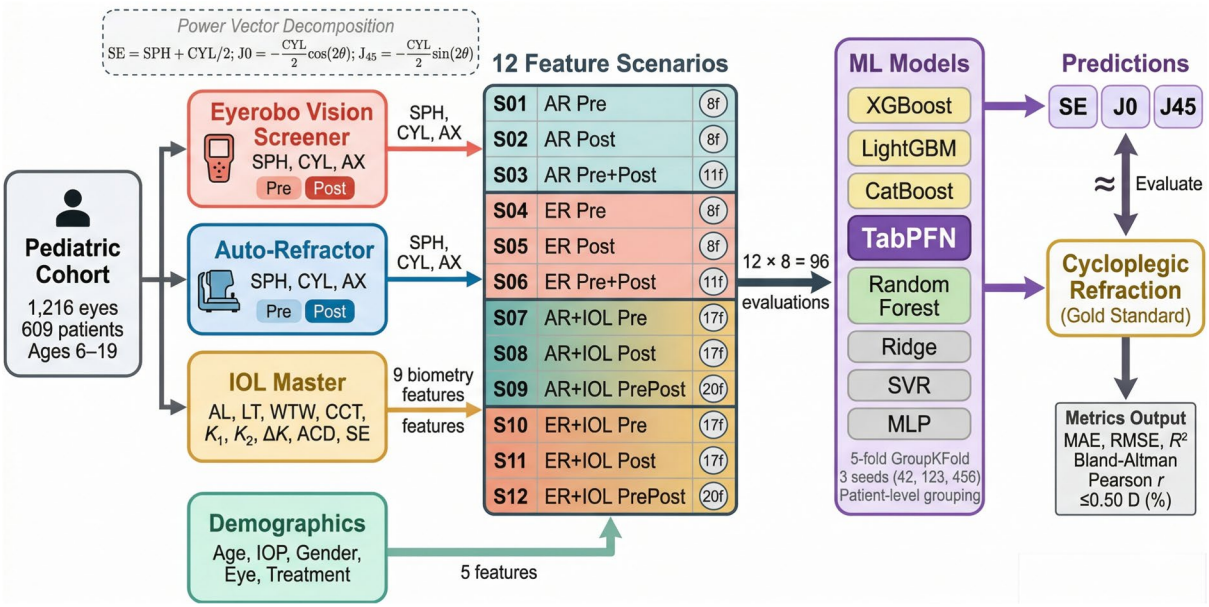

Fig. 1 Study pipeline for cycloplegic refraction prediction. Data from 1129 eyes of 574 pediatric patients (ages 6–18) were collected using three ophthalmic devices: the Eyerobo Vision Screener (portable handheld photorefractor), a conventional desk-mounted auto-refractor, and the IOL Master (ocular biometry). Each refraction device recorded spherical power (SPH), cylindrical power (CYL), and cylinder axis (AX) before and after pharmacological dilation. All refractive measurements were decomposed into power vector components ($SE = SPH + CYL/2$; $J0 = -(CYL/2)\cos(2\theta)$; $J45 = -(CYL/2)\sin(2\theta)$) following Thibos et al. Twelve feature scenarios (S01–S12) were constructed from a symmetric framework: each device was evaluated under three dilation conditions (pre, post, pre + post), with and without IOL Master biometry (9 features), plus five demographic features in all scenarios. The circled numbers beside each scenario (e.g., 8f, 11f, 17f, 20f) indicate the total number of input features: "f" denotes features, ranging from 8 (3 device features + 5 demographics) to 20 (6 device features from both dilation conditions + 9 IOL biometry + 5 demographics). Eight machine learning models, led by TabPFN (a zero-tuning tabular foundation model), were evaluated via patient-level five-fold grouped cross-validation across three random seeds, producing $12 \times 8 = 96$ model–scenario evaluations. Predictions of cycloplegic SE, J0, and J45 were compared against the cycloplegic refraction gold standard using MAE, RMSE, R^2, Bland–Altman agreement, Pearson correlation, and clinical threshold analysis (≤ 0.50 D)

making it clinically relevant beyond mere measurement accuracy.

The Tehran Eye Study by Hashemi et al. [11] provided population-level evidence that even subjective refraction without cycloplegia introduces systematic hyperopic underestimation and myopic overestimation, though the magnitude differs from that observed with autorefraction. Gopalakrishnan et al. [12] addressed the practical consequence of this discrepancy for prevalence estimation, defining thresholds for noncycloplegic myopia screening in children and demonstrating that noncycloplegic measurements systematically underestimate hyperopia while overestimating myopia. Collectively, these studies establish that the noncycloplegic to cycloplegic correction varies nonlinearly with age and refractive state, making simple linear correction factors or fixed offsets insufficient and motivating the use of flexible ML approaches.

Machine Learning for Cycloplegic Refraction Prediction

The application of ML to predict cycloplegic refraction from noncycloplegic data has

accelerated since 2023. Du et al. [13] were the first to frame the problem as predicting the SE difference before and after cycloplegia using ML, training XGBoost on 5920 eyes with refraction and biometric features and achieving an MAE of 0.31 D. Their work established the paradigm adopted by subsequent studies: combining noncycloplegic refractive measurements with ocular biometry as input features to gradient boosting models.

Ying et al. [16] expanded on this approach with 3414 eyes, comparing XGBoost and random forest models and reporting R^2 values of 0.913–0.935 with MAE of 0.393–0.480 D depending on feature configuration. Liu et al. [14] scaled the methodology to 7366 eyes and employed stacked ensemble methods, achieving an MAE of 0.360 D. Su et al. [15] introduced lens-related features (lens thickness, lens power) as predictors, achieving 0.307 D on 1694 eyes and demonstrating that crystalline lens parameters carry independent predictive information beyond standard refractive and biometric variables.

More recent contributions include Chen et al. [17], who trained random forest and XGBoost models on 1440 eyes with an MAE of approximately 0.35 D; Hao et al. [18], who demonstrated that random forest algorithms can predict post-cycloplegic myopic corrections from noncycloplegic clinical data with an MAE of approximately 0.40 D; and Kim et al. [19], who showed that ensemble methods applied to partial interferometry measurements from childhood eyes can predict clinical refraction with an MAE of 0.44 D on 750 eyes. Chandra et al. [20] provided the first external validation of ML-based cycloplegic prediction, applying the models of Ying et al. [16] to an independent cohort of 3414 eyes and reporting an MAE of 0.48 D, a modest degradation that highlights the challenge of cross-site generalization.

Despite this rapid progress, several gaps remain in the literature. All published studies use desk-mounted laboratory-grade auto-refractors as the noncycloplegic input device. No study has evaluated portable handheld photorefractors, which are the only refractive instruments available in many screening settings. No study has applied TabPFN or any tabular foundation model to this task. Finally, no study has directly compared two different input devices in a head-to-head framework to quantify the effect of measurement precision on ML prediction accuracy.

Portable Vision Screeners and TabPFN

Portable vision screeners based on photorefraction have been validated for amblyopia risk detection and refractive screening in children. Liu et al. [22] evaluated the 2WIN-S portable refractor, the predecessor of the Eyerobo VS used in the present study, against cycloplegic retinoscopy in 194 eyes of 97 school-aged children, reporting Pearson $r = 0.875$ for SE agreement. Thomas et al. [23] compared the Plusoptix A09 photorefractor against cycloplegic refraction in preschool children, finding moderate agreement for SE but significant discrepancies for cylinder measurements. Peterseim et al. [25] demonstrated that the Spot Vision Screener effectively detects amblyopia risk factors in pediatric populations, establishing portable screeners as viable tools for community-based screening. Dahlmann-Noor et al. [24] provided earlier evidence for the Plusoptix platform, comparing it against orthoptic assessment and finding acceptable sensitivity and specificity for detecting significant refractive errors.

These validation studies establish that portable screeners provide clinically useful but noisier measurements than desk-mounted auto-refractors, raising the question of whether ML can compensate for this reduced precision. The present study is the first to combine a portable photorefractor with ML-based cycloplegic prediction.

Tabular Prior-data Fitted Network (TabPFN) represents a fundamentally different approach to tabular prediction. Originally introduced by Hollmann et al. [28] for classification, the model was extended to regression and scaled to larger datasets by Hollmann et al. [27], who demonstrated in a Nature publication that TabPFN achieves competitive or superior performance to tuned gradient boosting ensembles on datasets with up to 10,000

samples, with zero hyperparameter tuning. TabPFN approximates Bayesian inference by pretraining a transformer on millions of synthetic datasets drawn from a prior distribution over data-generating processes. At inference time, the training data is passed as context and the model produces predictions in a single forward pass, requiring no gradient updates. This zero-tuning property, combined with native handling of missing values and deterministic predictions, makes TabPFN particularly attractive for clinical deployment where ML expertise may be limited. To our knowledge, TabPFN has not previously been applied to any ophthalmology prediction task.

DATASET

The study was approved by the ethics board of Gaobeidian Mingren Eye Hospital (2026RN-001) and conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from each participating child's parent or legal guardian, with assent from children of an appropriate age. All data were de-identified before analysis.

This cross-sectional study enrolled consecutive pediatric patients presenting for routine ophthalmic examination at a single center. Inclusion criteria were age 6–18 years, availability of cycloplegic refraction as the reference standard, and measurements from at least one of the three study devices. Patients with ocular pathology other than refractive error were excluded. Data cleaning included removal of two exact duplicate entries, correction of nine data entry errors (decimal point misplacements in keratometry and white-to-white measurements, physiologically impossible intraocular pressure (IOP) values > 45 mmHg set to missing), retention of first-visit data only for 17 patients with repeat measurements (39 rows removed), and exclusion of 87 eyes with a placeholder cycloplegic SE of exactly 0.00 D, which were judged to represent missing or unmeasured cycloplegic refractions rather than true emmetropia. The final dataset comprised 1129 eyes from 574 patients.

Study Population

Table 1 summarizes the demographic and clinical characteristics. The cohort was balanced by gender (571 male, 558 female; 50.6% male, 49.4% female) and eye laterality (566 right eyes [OD], 563 left eyes [OS]). The mean age was 11.2 ± 2.6 years (range 6–18), with peak enrolment at ages 11–13 (Fig. 2a). The mean cycloplegic SE was − 2.61 ± 2.07 D (range −11.88 to +7.12 D). The cohort was overwhelmingly myopic, reflecting its origin as a tertiary-referral population rather than a general-school cohort: 700 eyes (62.0%) had mild myopia (SE 0 to −3 D, excluding 0), 321 (28.4%) had moderate myopia (−3 to −6 D, excluding −3), and 59 (5.2%) had high myopia (SE < −6 D). In total, 49 eyes (4.3%) were not myopic; within this stratum, 14 eyes (1.2%) met the conventional emmetropic definition ($0 \leq SE \leq +0.5$ D) and 35 eyes (3.1%) were hyperopic (SE > +0.5 D) (Fig. 2b). The implications of this myopia-dominant spectrum for model generalisability are addressed in the Limitations section.

Instruments and Device Measurements

Three ophthalmic devices were used: the Eyerobo Vision Screener (VS), a handheld binocular photorefractor operating at approximately one meter working distance; a desk-mounted autorefractor (AR-1, Nidek, Japan); and the IOL (intraocular lens) Master (Carl Zeiss Meditec). Each refraction device (Eyerobo VS and autorefractor) recorded SPH, CYL, and AX. The IOL Master provided ocular biometry including axial length (AL), lens thickness (LT), white-to-white corneal diameter (WTW), central corneal thickness (CCT), flat and steep keratometry (K_1, K_2), keratometric astigmatism (ΔK), anterior chamber depth (ACD), and keratometric SE. All devices collected measurements before and after pharmacological dilation.

Figure 3a compares the distribution of SE measurements across devices and dilation conditions against the cycloplegic reference. Both devices showed a myopic shift relative to cycloplegic refraction in the predilation condition,

Table 1 Demographic and clinical characteristics of the study cohort

Characteristic	Value
Eyes (patients)	1129 (574)
Age, mean ± SD (range), years	11.2 ± 2.6 (6–18)
Male, *n* (%)	571 (50.6%)
Female, *n* (%)	558 (49.4%)
OD / OS	566 / 563
IOP, mean ± SD, mmHg	17.7 ± 2.9
Cycloplegic SE, mean ± SD (range), D	− 2.61 ± 2.07 (− 11.88 to + 7.12)
Refractive error distribution	
Hyperopic (SE > +0.5 D)	35 (3.1%)
Emmetropic (0 ≤ SE ≤ +0.5 D)	14 (1.2%)
Mild myopia (0 to −3 D, excludes 0)	700 (62.0%)
Moderate myopia (−3 to −6 D, excludes −3)	321 (28.4%)
High myopia (< −6 D)	59 (5.2%)
Treatment status at presentation	
Glasses	552 (48.9%)
No glasses	543 (48.1%)
Myopia management (atropine / orthokeratology)	6 (0.5%)
Not recorded	28 (2.5%)

Percentages are over the analytic cohort of 1129 eyes

SD standard deviation, *SE* spherical equivalent, *D* diopters, *IOP* intraocular pressure, *OD* right eye, *OS* left eye

consistent with accommodative artifact. Missing data rates (Fig. 3b) were low for most feature groups (≤ 6.3%), with auto-refractor post-dilation data showing the highest rate at 6.3%.

Cycloplegic refraction was performed by instilling cyclopentolate 1% eye drops (two drops administered five minutes apart), with autorefraction measurement taken 30 min after the first instillation and confirmed by the examining optometrist. The cycloplegic autorefraction SE, computed as $SE = SPH + CYL/2$, served as the prediction target. We note that while cycloplegic retinoscopy is considered the absolute gold standard in some clinical traditions, cycloplegic autorefraction is widely accepted as the reference standard in ML-based prediction studies [14, 16].

Methods

Power Vector Decomposition

All refractive measurements were decomposed into power vectors as described by Thibos et al. [29]: spherical equivalent

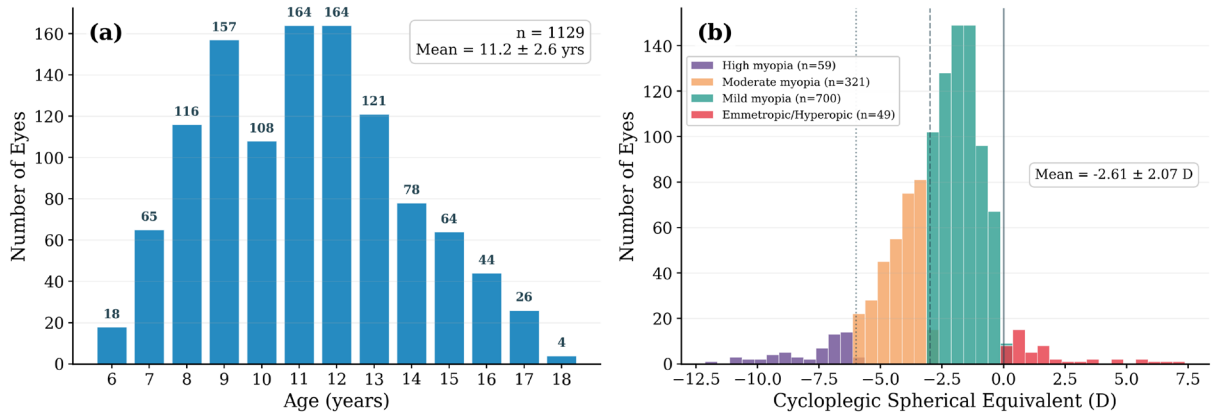

Fig. 2 Study population distributions. **a** Age distribution across the cohort (mean 11.2 ± 2.6 years). **b** Distribution of cycloplegic spherical equivalent (SE) in diopters (D), color-coded by refractive category: nonmyopic (coral, $n = 49$; breakdown: 14 emmetropic at $0 \leq$ SE $\leq +0.5$ D and 35 hyperopic at SE $> +0.5$ D), mild myopia (green, $n = 700$), moderate myopia (gold, $n = 321$), and high myopia (purple, $n = 59$). Vertical lines mark clinical thresholds at 0, −3, and −6 D. The take-away is that the cohort is strongly myopia-dominant; emmetropic and hyperopic strata are individually small and metrics for those strata should be interpreted as exploratory

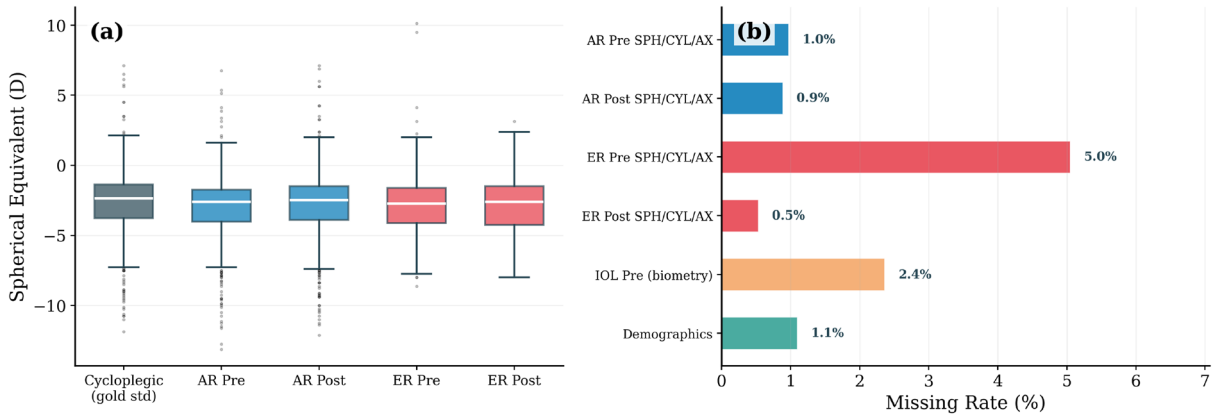

Fig. 3 Dataset characteristics. **a** Distribution of SE measurements by device and dilation condition compared with the cycloplegic gold standard. Boxes indicate interquartile range; whiskers extend to 1.5 × IQR. Teal = auto-refractor (AR); coral = Eyerobo VS (ER). **b** Mean missing data rates by feature group. All groups have ≤ 6.3% missing data

($SE = SPH + CYL/2$), orthogonal astigmatism ($J0 = -(CYL/2) \times \cos(2\theta)$), and oblique astigmatism ($J45 = -(CYL/2) \times \sin(2\theta)$), where θ is the cylinder axis in radians. These three components are mutually orthogonal and allow independent analysis of overall focus (SE), with-the-rule versus against-the-rule astigmatism (J0), and oblique astigmatism (J45).

Twelve-Scenario Framework

To systematically compare the two refraction devices, we defined twelve feature scenarios from a symmetric framework (Fig. 1, Table 2). Each device contributes SPH, CYL, and AX under three dilation conditions (predilation, post-dilation, or both). These six device-only scenarios are then repeated with the addition

Table 2 Twelve feature scenarios

#	Scenario	Device features	Total features
S01	AR Pre	SPH, CYL, AX	8
S02	AR Post	SPH, CYL, AX	8
S03	AR Pre + Post	SPH, CYL, AX (both)	11
S04	ER Pre	SPH, CYL, AX	8
S05	ER Post	SPH, CYL, AX	8
S06	ER Pre + Post	SPH, CYL, AX (both)	11
S07	AR Pre + IOL	SPH, CYL, AX + IOL	17
S08	AR Post + IOL	SPH, CYL, AX + IOL	17
S09	AR Pre + Post + IOL	SPH, CYL, AX + IOL	20
S10	ER Pre + IOL	SPH, CYL, AX + IOL	17
S11	ER Post + IOL	SPH, CYL, AX + IOL	17
S12	ER Pre + Post + IOL	SPH, CYL, AX + IOL	20

All scenarios include five demographic features (age, IOP, sex, eye laterality, treatment status)

AR auto-refractor, *ER* Eyerobo VS, *IOL* IOL Master biometry (9 features)

of IOL Master biometry (9 features: AL, LT, WTW, CCT, K_1, K_2, ΔK, ACD, SE), producing twelve scenarios in total. Demographic variables (age, intraocular pressure, sex, eye laterality, and treatment status) were included in all scenarios. Only SPH, CYL, and AX were used from each refraction device to ensure an apples-to-apples comparison.

Machine Learning Models

Eight models were evaluated, selected to span the algorithm families most commonly used in published ophthalmic ML and to bracket TabPFN by both simpler and more flexible competitors. TabPFN v2 [27] (eight estimators, GPU acceleration, zero tuning) was chosen as a recent tabular foundation model that has not previously been applied to ophthalmic refraction prediction. XGBoost [30] (200 rounds, depth 6, learning rate 0.05), LightGBM [31], and CatBoost [32] cover the gradient-boosting family that has dominated the recent published cycloplegic-prediction literature (Du et al., Ying et al., Liu et al., Chen et al., Hao et al.). Random Forest [33] (100 trees, depth 10) provides a robust nonparametric reference. Ridge regression, support vector regression (SVR, radial basis function [RBF] kernel), and a multilayer perceptron (MLP, 128–64 hidden units) cover the classical linear, kernel, and small-neural-network families respectively, ensuring that any TabPFN advantage is interpretable relative to baselines of different inductive bias. All gradient-boosting models used default hyperparameters without per-dataset tuning to ensure a fair comparison with TabPFN's zero-tuning design; we report this constraint explicitly and note that per-dataset tuning of the gradient-boosting baselines may narrow the margin, though it would forfeit the zero-tuning operational advantage of TabPFN.

Evaluation Protocol

All models were evaluated using five-fold grouped cross-validation with patient-level grouping (GroupKFold), repeated across three random seeds (42, 123, 456), ensuring both eyes of a given patient are assigned to the same fold. Performance metrics included MAE, root-mean-squared error (RMSE), coefficient of determination (R^2), and the percentage of predictions within ± 0.25 D, ± 0.50 D, and ± 1.00 D. Agreement was assessed using Bland–Altman analysis [34], reporting the mean difference (bias) and 95% limits of agreement (LoA), alongside Pearson correlation. For the primary regression comparisons between TabPFN and the next-best competing algorithm in each scenario, we report paired bootstrap 95% confidence intervals on the per-eye MAE difference (2000 resamples) alongside paired Wilcoxon signed-rank tests on per-eye absolute errors. Effect sizes (mean paired difference) are reported alongside *p*-values to discourage interpretation of *p* in isolation. Where multiple pairwise comparisons are conducted per scenario, *p*-values are interpreted descriptively rather than as confirmatory tests; the headline TabPFN-versus-next-best comparison is prespecified.

To complement the regression endpoints, a screening-classification analysis was performed at three prespecified, clinically meaningful thresholds for the binary referral decision (refer for cycloplegic examination, yes/no): any myopia (cycloplegic SE ≤ -0.50 D), moderate myopia (≤ -3.00 D), and high myopia (≤ -6.00 D). A noncycloplegic Youden-optimal cutoff was selected on each training fold and the cross-validation-averaged cutoff was used as the reported operating point. Sensitivity, specificity, PPV, NPV, AUC, and Youden's J were computed at this operating point; 95% confidence intervals are paired bootstrap intervals with 2000 resamples. Head-to-head device comparison was performed by computing the paired bootstrap 95% confidence interval on the AUC difference $\Delta AUC = AUC_{AR} - AUC_{ER}$ on the subset of eyes with valid measurements from both devices. We prespecified a noninferiority margin of $\delta = 0.05$ in AUC at the primary (any-myopia) threshold; noninferiority was declared if the upper bound of the 95% confidence interval lay below δ. Per-patient performance was assessed using a worse-eye rule, in which a patient was flagged as positive if either eye met the referral criterion.

To address concerns about inter-eye correlation beyond the GroupKFold control, we ran a one-eye-per-patient sensitivity analysis: for each of three seeds (42, 123, 456) one eye was randomly selected per patient, the full regression and classification pipeline was re-run, and the resulting per-eye metric estimates were compared with the per-eye headline values. Feature importance was assessed using permutation importance on XGBoost as a surrogate model, because TabPFN does not provide a native global feature-importance summary. Robustness to small input perturbations was assessed by jittering the noncycloplegic SPH and CYL inputs by ± 0.25 D (Gaussian noise, $\sigma = 0.083$ D so that 99.7% of perturbations fall within ± 0.25 D) and recomputing MAE. Per-eye inference time was measured on a single CPU core for all models, separated from training time, so as to characterize deployment compute cost. All experiments were implemented in Python 3.10 using scikit-learn 1.3 [35], XGBoost 2.0, LightGBM 4.1, CatBoost 1.2, and TabPFN 2.0. Total wall-clock time for the 12-scenario evaluation was approximately 15 min on a single NVIDIA GPU with CUDA 12.9.

RESULTS

Raw Device Agreement Before Machine Learning

Table 3 presents the agreement between each device and cycloplegic refraction across five refractive parameters. The auto-refractor showed stronger correlation with cycloplegic SE ($r = 0.978$, mean difference +0.374 D) compared with the Eyerobo ($r = 0.787$, mean difference +0.254 D). The Eyerobo exhibited less systematic bias than the auto-refractor for SE (+0.254 versus +0.374 D), consistent with its longer working distance reducing accommodative artifact, but had substantially wider limits of agreement (-2.28 to $+2.80$ D versus -0.47 to $+1.22$ D). For J0, both devices showed small mean differences (AR: +0.010 D; ER: +0.014 D) with the Eyerobo J0 bias nonsignificant ($p = 0.358$). Post-dilation auto-refractor measurements were nearly identical to cycloplegic refraction ($r = 0.994$ for SE). Bland–Altman plots are shown in Fig. 4.

ML Prediction of Cycloplegic SE

Table 4 presents TabPFN prediction performance across all twelve scenarios. The auto-refractor (hereafter AR) outperformed the Eyerobo VS (hereafter ER) on all SE comparisons: predilation MAE was 0.295 D (AR) versus 0.657 D (ER), with 87.4% versus 61.6% within ± 0.50 D. Adding IOL biometry to the Eyerobo reduced the gap by 64%, from 0.362 D to 0.130 D. The IOL Master provided greater benefit to the Eyerobo (MAE improvement of 0.232 D) than to the auto-refractor (0.012 D), confirming that biometric parameters compensate for the Eyerobo's lower refractive precision. Post-dilation auto-refractor predictions approached cycloplegic accuracy (MAE=0.203 D, 96.0% within ± 0.50 D). Bland–Altman and scatter plots for four key scenarios are shown in Figs. 5 and 6.

Table 3 Raw device agreement with cycloplegic refraction (power vector analysis)

Device	Parameter	*n*	Δ(D)	SD (D)	95% LoA (D)	*r*
AR Pre	SPH	1117	+0.344	0.409	[−0.46, +1.15]	0.977
	CYL	1117	+0.061	0.357	[−0.64, +0.76]	0.921
	SE	1117	+0.374	0.433	[−0.47, +1.22]	0.978
	J0	1117	+0.010	0.158	[−0.30, +0.32]	0.934
	J45	1117	+0.008	0.101	[−0.19, +0.21]	0.787
ER Pre	SPH	1071	+0.173	1.233	[−2.24, +2.59]	0.779
	CYL	1071	+0.178	1.153	[−2.08, +2.44]	0.396
	SE	1071	+0.254	1.295	[−2.28, +2.80]	0.787
	J0	1071	+0.014	0.507	[−0.98, +1.01]	0.454
	J45	1071	−0.028	0.383	[−0.78, +0.72]	0.122
AR Post	SE	1119	+0.173	0.231	[−0.28, +0.63]	0.994
	J0	1119	−0.070	0.130	[−0.33, +0.18]	0.954
	J45	1119	+0.002	0.071	[−0.14, +0.14]	0.895
ER Post	SE	1123	+0.141	1.238	[−2.29, +2.57]	0.810
	J0	1123	−0.090	0.537	[−1.14, +0.96]	0.455
	J45	1123	−0.028	0.423	[−0.86, +0.80]	0.034

Δ cycloplegic minus device, *LoA* 95% limits of agreement, *r* Pearson correlation

ML prediction of Power Vector Components

Table 5 presents prediction performance for J0 and J45. The device gap was substantially smaller for these components. Pre-dilation J0 MAE was 0.106 D (AR) versus 0.211 D (ER). With IOL biometry, ER J0 improved to 0.143 D, narrowing the gap to 0.037 D. For J45, the gap was 0.060 D (AR) versus 0.089 D (ER), a difference of 0.029 D that is below the clinically meaningful threshold for oblique astigmatism. The smaller astigmatic gap reflects the narrower dynamic range of cylinder measurements, where both devices capture information with proportionally similar precision.

Model Comparison

Table 6 compares all eight models on four key predilation scenarios. TabPFN achieved the lowest MAE on both Eyerobo scenarios (S04 and S10), while Ridge achieved the lowest MAE on both auto-refractor scenarios (S01 and S07). On S01 (AR Pre), Ridge (0.285 D) narrowly outperformed TabPFN (0.295 D), and on S07 (AR Pre + IOL), Ridge (0.280 D) similarly edged TabPFN (0.283 D), reflecting the strong linearity of the auto-refractor signal where a simple linear model captures most of the variance. On S04 (ER Pre), TabPFN (0.657 D) significantly outperformed all competitors including Ridge (0.730 D, $p < 0.001$), and on S10 (ER Pre + IOL), TabPFN (0.425 D) significantly outperformed Ridge (0.522 D, $p < 0.001$). TabPFN's advantage on Eyerobo scenarios, where the input signal is noisier and the relationship between noncycloplegic and cycloplegic measurements is more complex, is consistent with its capacity to learn nonlinear patterns. MLP produced unstable results on IOL-augmented scenarios (MAE > 1.7 D) owing to overfitting on the 17–20 feature input space with limited

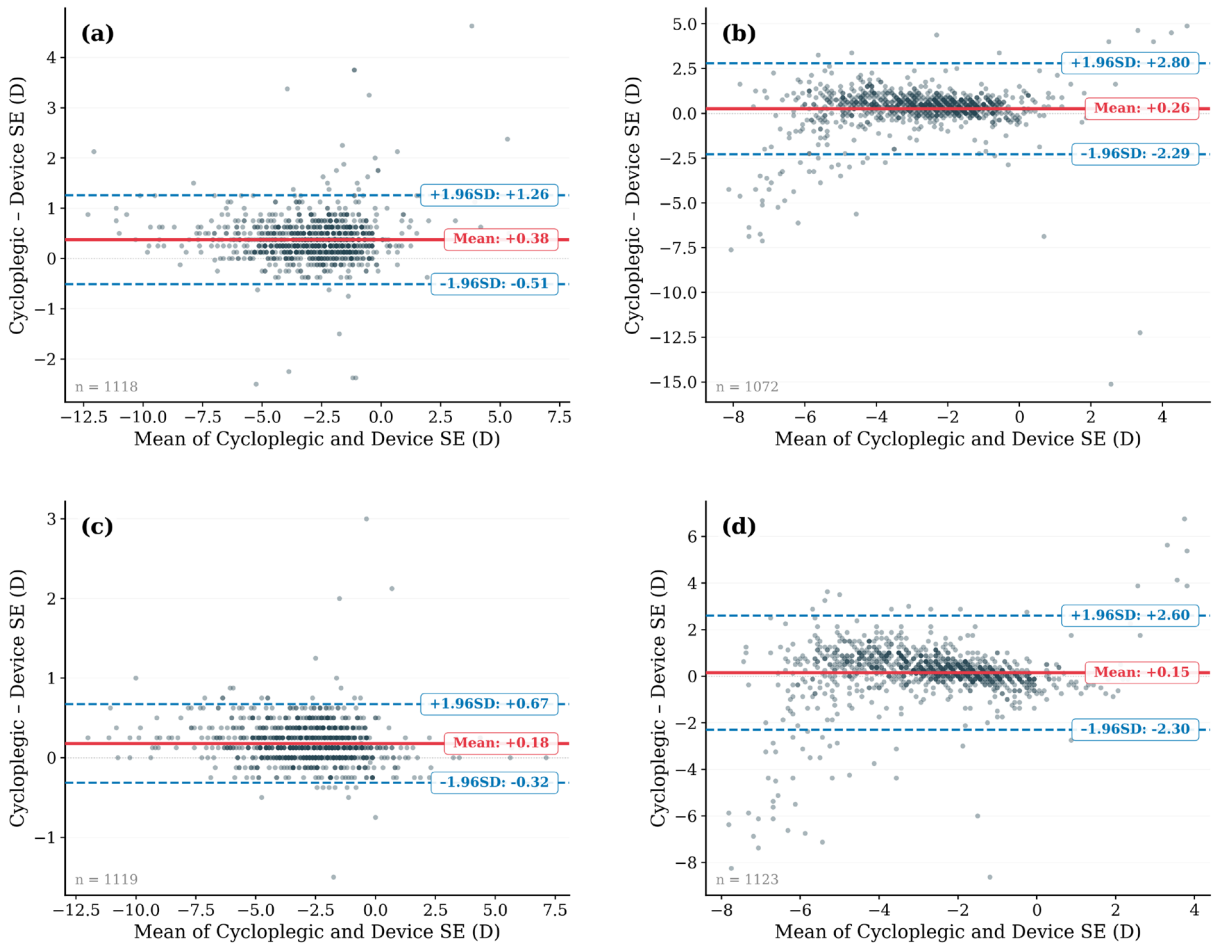

Fig. 4 Bland–Altman plots for raw device agreement with cycloplegic SE (before ML). **a** Auto-refractor predilation. **b** Eyerobo VS predilation. **c** Auto-refractor post-dilation. **d** Eyerobo VS post-dilation. Red lines = mean difference; blue dashed lines = 95% limits of agreement

training samples, and was excluded from those comparisons. We note that all gradient boosting models used default hyperparameters; per-dataset tuning may narrow the gap with TabPFN, though this would forfeit TabPFN's zero-tuning deployment advantage.

Subgroup Analysis

Table 7 presents TabPFN performance by refractive error and age, with explicit sample sizes for each stratum. Both devices performed best for mild myopia ($n = 700$): AR Pre MAE = 0.262 D (90% within ± 0.50 D) and ER Pre MAE = 0.398 D (76%). The Eyerobo–auto-refractor gap was smallest for mild myopia (0.136 D) and widest for high myopia (2.452 D). Adding IOL biometry reduced the Eyerobo's high-myopia MAE from 2.802 to 0.800 D. Younger children (ages 6–9 years) showed lower MAE than older children for both devices.

The nonmyopic stratum ($n = 49$, 4.3% of the cohort) is reported as a combined row in Table 7, but it comprises two clinically distinct subgroups with different implications for the triage interpretation: 14 emmetropic eyes ($0 \leq$ SE $\leq +0.5$ D) and 35 hyperopic eyes (SE $> +0.5$ D). Because the hyperopic stratum is small in

Table 4 TabPFN prediction of cycloplegic SE across twelve scenarios

Scenario	MAE	RMSE	R^2	≤ 0.25 (%)	≤ 0.50 (%)	≤ 1.00 (%)	r	95% LoA
Device only								
S01: AR Pre	0.295	0.517	0.937	61.2	87.4	96.9	0.969	[− 0.99, + 1.03]
S02: AR Post	0.203	0.403	0.962	76.1	96.0	98.5	0.981	[− 0.77, + 0.81]
S03: AR PrePost	0.196	0.394	0.964	77.1	95.7	98.6	0.982	[− 0.76, + 0.79]
S04: ER Pre	0.657	1.131	0.700	37.2	61.6	83.1	0.837	[− 2.25, + 2.19]
S05: ER Post	0.606	1.137	0.697	43.4	66.9	84.4	0.835	[− 2.26, + 2.20]
S06: ER PrePost	0.565	1.035	0.749	44.7	69.0	86.1	0.866	[− 2.05, + 2.01]
Device + IOL Master								
S07: AR Pre + IOL	0.283	0.512	0.939	60.6	86.8	97.8	0.969	[− 0.99, + 1.02]
S08: AR Post + IOL	0.200	0.398	0.963	75.6	96.3	98.5	0.982	[− 0.76, + 0.80]
S09: AR PrePost + IOL	0.194	0.397	0.963	77.5	96.4	98.6	0.982	[− 0.77, + 0.79]
S10: ER Pre + IOL	0.425	0.663	0.897	44.3	72.0	91.6	0.947	[− 1.30, + 1.30]
S11: ER Post + IOL	0.394	0.632	0.906	49.1	75.7	93.0	0.952	[− 1.23, + 1.25]
S12: ER PrePost + IOL	0.379	0.593	0.918	50.0	77.1	93.0	0.958	[− 1.16, + 1.17]

Δ mean bias, *LoA* 95% limits of agreement, *r* Pearson correlation

absolute terms, stratum-specific point estimates have wide implicit confidence intervals and should be interpreted as exploratory; we do not split the row further in Table 7 because the resulting per-cell sample sizes would not support quantitative comparison. The clinically relevant consequence, that noncycloplegic methods systematically underestimate hyperopia in children and that a triage rule trained on a myopia-dominant cohort will tend to under-flag hyperopic children, is addressed substantively in the Limitations section.

Feature Importance and Learning Curve

Permutation importance analysis revealed that SPH was the dominant feature for both devices: auto-refractor SPH had importance 1.353 (S01), Eyerobo SPH had 0.763 (S04). When IOL biometry was added, axial length became the dominant predictor for the Eyerobo (0.633 in S10, exceeding Eyerobo SPH at 0.462), whereas auto-refractor SPH remained dominant with IOL features (1.202 in S07). This confirms that IOL biometry provides complementary structural information that partially compensates for the Eyerobo's lower refractive precision. The learning curve (Fig. 7) showed TabPFN achieving the lowest MAE at all training fractions, with no saturation at the current sample size.

Screening-Oriented Performance at Clinical Thresholds

The MAE / RMSE metrics in the preceding subsections quantify how precisely each device's predicted SE tracks the cycloplegic value, which is the relevant endpoint for prescription accuracy. For the clinical question that this study is positioned to inform, whether the device–model combination correctly *flags* a child for referral to a full cycloplegic examination, the operative metrics are the classification sensitivity, specificity, positive and negative predictive values

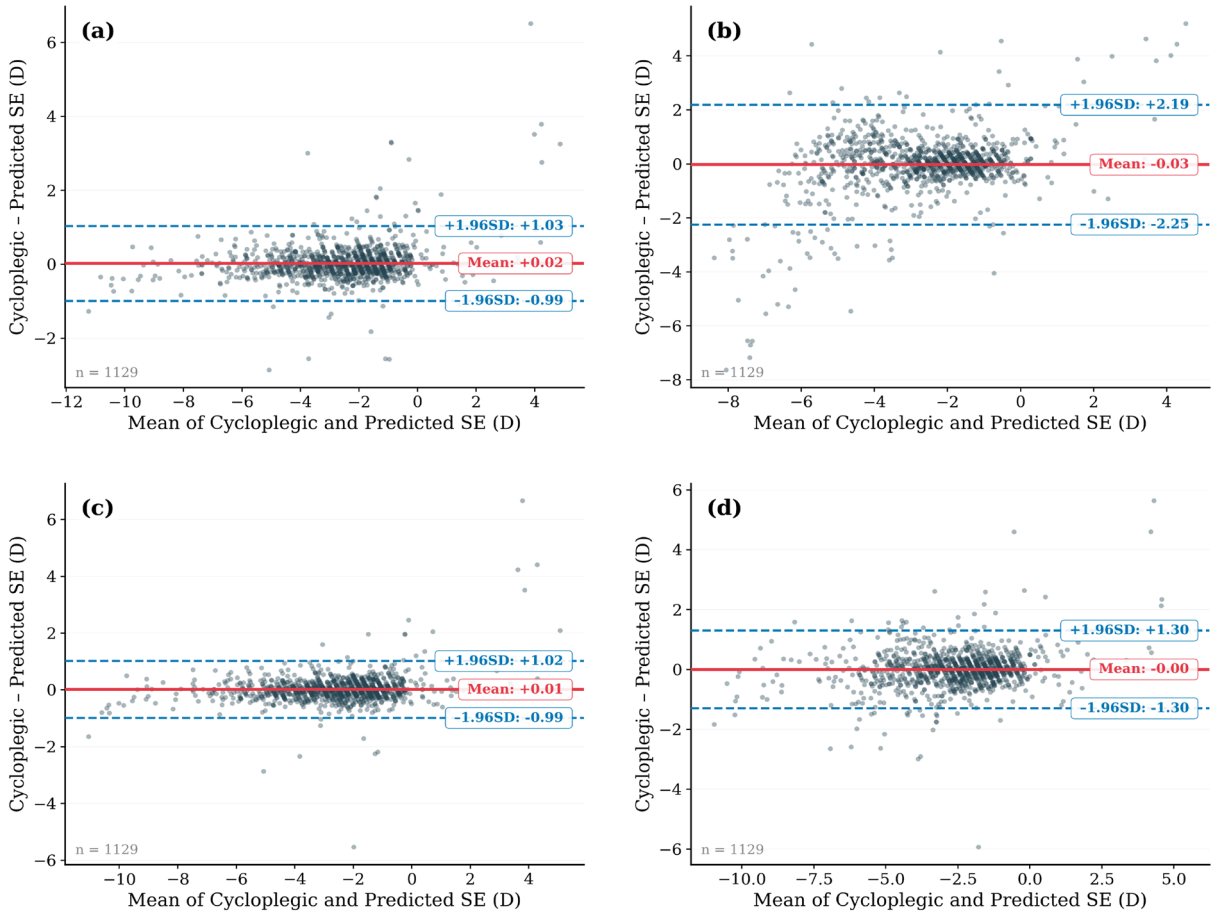

Fig. 5 Bland–Altman plots for TabPFN-predicted cycloplegic SE. **a** S01: AR Pre. **b** S04: ER Pre. **c** S07: AR Pre + IOL. **d** S10: ER Pre + IOL. Comparison of **b** and **d** shows how IOL biometry narrows the Eyerobo's prediction scatter

(PPV / NPV), and the area under the receiver operating characteristic curve (AUC), evaluated at prespecified, clinically meaningful refractive thresholds. We report these in Table 8 at three thresholds aligned with accepted myopia definitions: any myopia (cycloplegic SE ≤ -0.50 D), moderate myopia (SE ≤ -3.00 D), and high myopia (SE ≤ -6.00 D). The cohort prevalence at each threshold was 92.1%, 37.9%, and 5.9% respectively. The 95% confidence intervals are paired bootstrap intervals with 2000 resamples computed on the cross-validation-derived deployment operating point; reported point estimates are means across the three random seeds. The threshold rule was implemented as a Youden-optimal cutoff on the noncycloplegic SE selected on each training fold; the cross-validation-averaged cutoff corresponds to a noncycloplegic Eyerobo SE of approximately -1.12 D for the any-myopia threshold.

Three observations follow. First, at the any-myopia threshold the Eyerobo and the auto-refractor are operationally equivalent at the classification endpoint: the paired bootstrap difference in AUC was +0.005 in favour of the auto-refractor with a 95% confidence interval of -0.022 to +0.034 (paired, $n = 1072$). Under a prespecified noninferiority margin of $\delta = 0.05$ in AUC, the upper bound of this confidence interval lies well below δ, so the Eyerobo is formally noninferior to the auto-refractor for the any-myopia triage decision. Aggregating to the patient level using a worse-eye rule yielded an AUC of 0.947 for the Eyerobo ($n = 545$ patients) and 0.967 for the auto-refractor ($n = 569$ patients), preserving

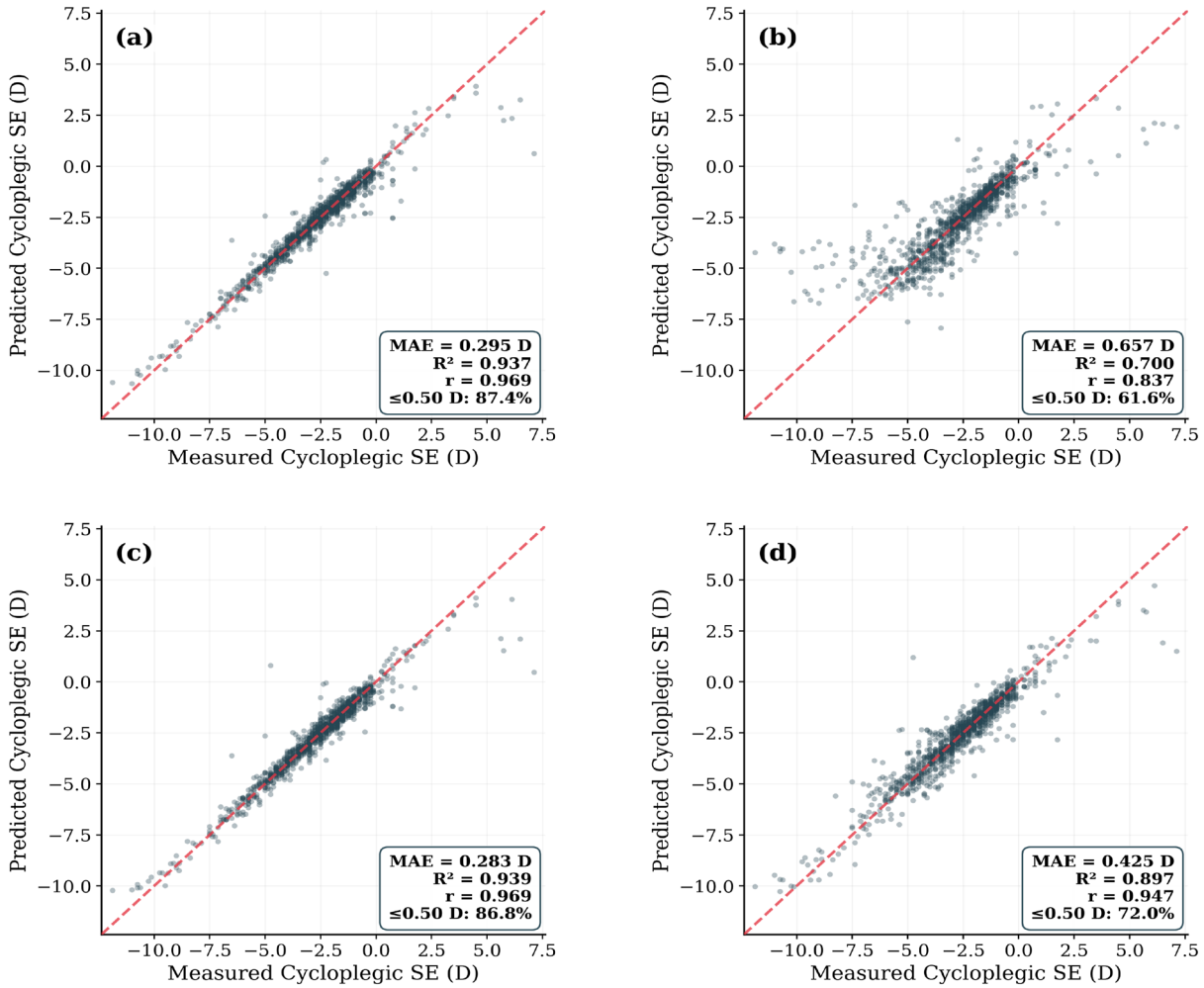

Fig. 6 Scatter plots of TabPFN-predicted versus measured cycloplegic SE. **a** S01: AR Pre. **b** S04: ER Pre. **c** S07: AR Pre + IOL. **d** S10: ER Pre + IOL. Dashed red lines indicate perfect agreement. Panels annotated with MAE, R^2, Pearson r, and percentage within ± 0.50 D

the per-eye finding. Second, at the moderate-myopia threshold the auto-refractor retains a clear advantage (AUC 0.993 versus 0.975) but the Eyerobo's performance remains in the clinically useful range. Third, at the high-myopia threshold the Eyerobo's classification performance degrades substantially (AUC 0.870; Spec 0.714); this is consistent with the device's effective measurement range and with the small high-myopia subgroup in this cohort ($n = 67$ by the ≤ -6.00 D definition). The clinical impact of this third-tier degradation is partially mitigated by the fact that any child who exceeds the high-myopia threshold has, by definition, also exceeded the any-myopia threshold and would already have been flagged for cycloplegic referral by the first-line rule. The PPV at the high-myopia threshold is low (0.164 for the Eyerobo) because high myopia is rare in this cohort (5.9%); the PPV will rise in any deployment population with a higher high-myopia prevalence, and the NPV (0.990) is already high. The PPV / NPV in this table reflect the prevalence of the present cohort (a tertiary referral population) and must be re-estimated for any deployment population with a different prevalence.

Table 5 TabPFN prediction of power vector components J0 and J45 across key scenarios

Scenario	J0			J45		
	MAE (D)	R^2	r	MAE (D)	R^2	r
S01: AR Pre	0.106	0.875	0.936	0.060	0.599	0.779
S04: ER Pre	0.211	0.404	0.639	0.089	0.091	0.319
S07: AR Pre + IOL	0.103	0.883	0.940	0.059	0.607	0.786
S10: ER Pre + IOL	0.143	0.786	0.887	0.088	0.086	0.315
S02: AR Post	0.094	0.879	0.938	0.042	0.769	0.878
S05: ER Post	0.220	0.379	0.618	0.090	0.047	0.239
S11: ER Post + IOL	0.140	0.797	0.893	0.090	0.043	0.227
S12: ER PrePost + IOL	0.136	0.802	0.896	0.086	0.090	0.313

Table 6 Model comparison: MAE (D) on four key predilation scenarios (mean ± SD across seeds)

Model	S01: AR Pre	S04: ER Pre	S07: AR + IOL	S10: ER + IOL
TabPFN	0.295	**0.657**	0.283	**0.425**
Ridge	**0.285**	0.730	**0.280**	0.522
CatBoost	0.348 ± 0.004	0.689 ± 0.002	0.343 ± 0.002	0.535 ± 0.006
RF	0.347 ± 0.000	0.722 ± 0.003	0.337 ± 0.001	0.592 ± 0.003
MLP[†]	0.323 ± 0.003	0.749 ± 0.012	–	–
SVR	0.374	0.741	0.430	0.610
XGBoost	0.354 ± 0.002	0.738 ± 0.004	0.318 ± 0.003	0.542 ± 0.001
LightGBM	0.357 ± 0.002	0.744 ± 0.001	0.336 ± 0.003	0.543 ± 0.004

Bold = best per scenario. [†]MLP unstable on IOL scenarios. Ridge achieves the lowest MAE on both auto-refractor scenarios (S01, S07), while TabPFN achieves the lowest MAE on both Eyerobo scenarios (S04, S10)

A concrete referral rule consistent with these results, to be evaluated prospectively in an external cohort before clinical use, is: refer for a full cycloplegic examination any child whose noncycloplegic Eyerobo SE is more myopic than approximately −1.12 D, with paired age-appropriate visual-acuity testing to compensate for the rule's reduced sensitivity in the hyperopic stratum (see Limitations). The supporting screening-classification pipeline, including decision-curve net-benefit comparisons and the per-patient aggregation reported above, is publicly available together with the regression code at the project repository linked in *Data and code availability*.

Ablation Studies

The 12-scenario framework permits systematic ablation of dilation condition, device source, and biometric supplementation. We summarize the key findings below.

The effect of dilation condition differed markedly between devices. For the auto-refractor, post-dilation measurements improved SE prediction relative to predilation: MAE

Table 7 Subgroup analysis: TabPFN mean absolute error (MAE) in diopters (D), with the percentage of predictions within ± 0.50 D shown in parentheses

Subgroup	*n*	Device only		Device + IOL	
		AR Pre	ER Pre	AR + IOL	ER + IOL
By refractive error					
Non-myopic (emm. + hyp.)[a]	49	1.021 (53%)	1.581 (22%)	0.909 (55%)	1.012 (37%)
Mild myopia	700	0.262 (90%)	0.398 (76%)	0.248 (88%)	0.319 (80%)
Moderate myopia	321	0.245 (89%)	0.687 (47%)	0.250 (90%)	0.497 (66%)
High myopia	59	0.350 (81%)	2.802 (8%)	0.362 (80%)	0.800 (44%)
By age group, years					
Age 6–9	356	0.296 (88%)	0.476 (71%)	0.283 (88%)	0.364 (76%)
Age 10–13	557	0.304 (87%)	0.662 (62%)	0.285 (85%)	0.426 (73%)
Age 14 +	216	0.270 (88%)	0.942 (46%)	0.279 (88%)	0.524 (64%)

The nonmyopic stratum ($n = 49$) comprises 14 emmetropic ($0 \leq SE \leq +0.5$ D) and 35 hyperopic (SE > +0.5 D) eyes; per-cell sample sizes do not support quantitative splitting of this stratum. The small high-myopia stratum ($n = 59$) means that Eyerobo high-myopia MAE estimates carry wider implicit uncertainty than the other strata

AR auto-refractor, *ER* Eyerobo Vision Screener, *IOL* IOL Master biometry augmentation

[a]Comprises 14 emmetropic and 35 hyperopic eyes

decreased from 0.295 D (S01: AR Pre) to 0.203 D (S02: AR Post), a reduction of 0.092 D (31.2%). Combining pre- and post-dilation measurements (S03: AR PrePost, MAE = 0.196 D) offered only marginal further improvement over post-dilation alone (0.007 D). For the Eyerobo, the improvement from pre- to post-dilation was more modest: MAE decreased from 0.657 D (S04: ER Pre) to 0.606 D (S05: ER Post), a reduction of 0.051 D (7.8%). Combining both dilation conditions (S06: ER PrePost, MAE = 0.565 D) provided additional improvement of 0.041 D beyond post-dilation alone. The substantially larger benefit of post-dilation data for the auto-refractor than for the Eyerobo is consistent with the auto-refractor's higher raw measurement precision: when accommodation is pharmacologically removed, the already-precise auto-refractor measurements become nearly identical to the cycloplegic reference, whereas the Eyerobo's measurement noise persists regardless of dilation condition.

The addition of IOL Master biometry provided asymmetric benefits to the two devices. For the Eyerobo predilation scenario, IOL biometry reduced MAE from 0.657 D (S04) to 0.425 D (S10), an improvement of 0.232 D (35.3%). For the auto-refractor predilation scenario, the corresponding improvement was only 0.012 D (4.1%), from 0.295 D (S01) to 0.283 D (S07). IOL biometry thus benefited the Eyerobo approximately 19 times more than the auto-refractor. This asymmetry is explained by the feature importance analysis: in the Eyerobo + IOL scenario (S10), axial length achieved an importance of 0.633, exceeding the Eyerobo SPH importance

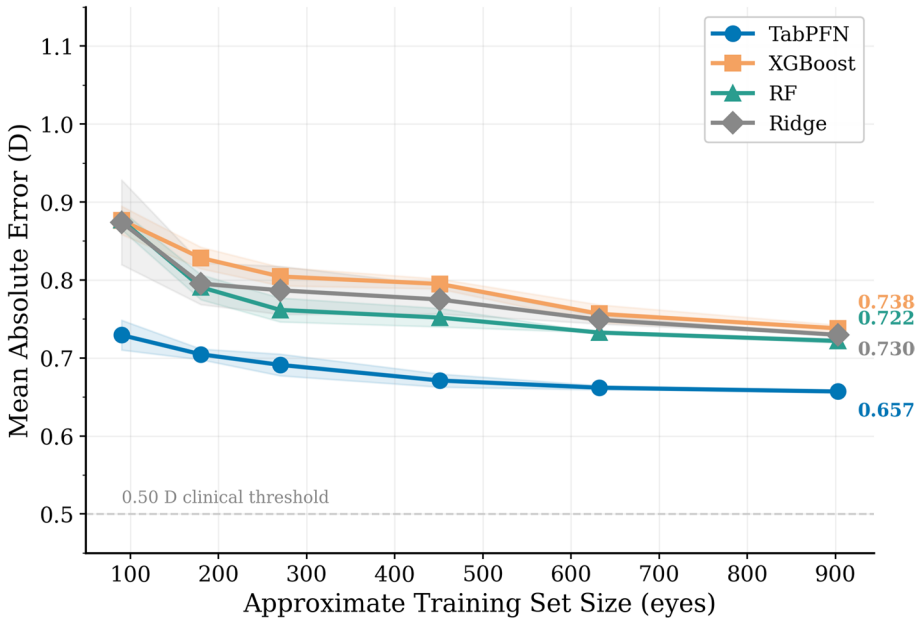

Fig. 7 Learning curve for Scenario S04 (Eyerobo Pre). TabPFN achieves the lowest MAE at all training set fractions, with no saturation at the full sample size

Table 8 Screening-oriented performance at three clinically meaningful referral thresholds

Threshold	Device	AUC (95% CI)	Sens	Spec	PPV	NPV	Youden J
Any (SE ≤ −0.50 D)	AR	0.969 (0.948–0.986)	0.960	0.886	0.990	0.655	0.847
	ER	0.964 (0.939–0.982)	0.946	0.894	0.991	0.579	0.840
Moderate (SE ≤ −3.00 D)	AR	0.993 (0.989–0.996)	0.944	0.969	0.950	0.966	0.914
	ER	0.975 (0.967–0.984)	0.918	0.911	0.862	0.948	0.828
High (SE ≤ −6.00 D)	AR	0.995 (0.988–0.999)	0.985	0.979	0.750	0.999	0.964
	ER	0.870 (0.836–0.903)	0.885	0.714	0.164	0.990	0.600

AUC area under the receiver operating characteristic curve, *Sens* sensitivity, *Spec* specificity, *PPV* positive predictive value, *NPV* negative predictive value, *Youden J* sensitivity + specificity −1, *CI* bootstrap 95% confidence interval. PPV and NPV are reported at the cohort prevalence indicated (92.1%, 37.9%, 5.9%) and will shift with the prevalence of the deployment population. *AR* auto-refractor, *ER* Eyerobo Vision Screener

of 0.462, meaning the model relies more on the structural measurement of eye anatomy than on the Eyerobo's noisier refractive reading. By contrast, for the auto-refractor + IOL scenario (S07), auto-refractor SPH remained the dominant feature (importance = 1.202) with axial length contributing a secondary importance of 0.203, because the auto-refractor's precise SPH already captures the dominant signal. These findings are consistent with the well-established correlation between axial length and refractive error [36, 37].

When both dilation conditions and IOL biometry were combined, the Eyerobo's best configuration (S12: ER PrePost + IOL) achieved an MAE of 0.379 D with 77.1% of predictions within ± 0.50 D, surpassing the auto-refractor's predilation-only performance (S01: 0.295 D,

87.4%) in clinical threshold terms. The auto-refractor's corresponding best combined configuration (S09: AR PrePost + IOL) achieved 0.194 D with 96.4% within ± 0.50 D. These results demonstrate that the combination of multiple dilation conditions with biometric supplementation can partially compensate for the Eyerobo's lower optical precision, bringing its performance below the clinically meaningful 0.50 D threshold and into the range suitable for screening applications.

Marginal-Gain Decomposition: What Does Machine Learning Add?

To isolate the contribution of machine learning from that of biometric supplementation we report the cross-tabulation of {linear versus TabPFN} × {device-only versus device + IOL biometry} on a fixed fivefold patient-level grouped CV protocol (Table 9; three-seed mean MAE in diopters). The decomposition supports three explicit statements that are otherwise hidden in the 12-scenario grid. First, on the noisier Eyerobo input, the switch from Ridge to TabPFN reduces MAE by 0.073 D when no biometry is available and by 0.097 D when biometry is added; the additional biometry contribution is roughly two to three times larger than the algorithmic contribution. Second, on the lower-noise auto-refractor input, TabPFN does not beat Ridge (within ±0.010 D in both columns), indicating that a regularised linear model already saturates the available signal when the input is already precise. Third, in aggregate, the residual gap between the best Eyerobo configuration (Eyerobo + IOL + TabPFN, 0.425 D) and the auto-refractor baseline (auto-refractor + Ridge, 0.285 D) is 0.140 D, compared with a bare-device gap of 0.445 D (Eyerobo + Ridge 0.730 D versus auto-refractor + Ridge 0.285 D), corresponding to approximately 69% closure of the bare device gap by the combination of biometric supplementation and TabPFN; biometric supplementation contributes the dominant share and TabPFN contributes the remainder. Methodologically, this is the answer to the operational question of what machine learning adds beyond a linear baseline: on noisy inputs it adds a clinically modest but statistically detectable improvement; on clean inputs it does not. The detailed table with per-seed standard deviations is reproduced in the Supplementary Appendix (Table S5).

Table 9 Marginal-gain decomposition: {Ridge versus TabPFN} × {device only versus device + IOL biometry}

Pipeline/device	Ridge MAE (D)	TabPFN MAE (D)
Eyerobo VS		
Device features only (S04)	0.730	0.657
Device + IOL biometry (S10)	0.522	0.425
Δ from biometry	−0.208	−0.232
Δ from Ridge → TabPFN	–	−0.073/−0.097
Auto-refractor		
Device features only (S01)	0.285	0.295
Device + IOL biometry (S07)	0.280	0.283
Δ from biometry	−0.005	−0.012
Δ from Ridge → TabPFN	–	+0.010/+0.004

MAE in diopters; three-seed mean. *AR* auto-refractor, *ER* Eyerobo Vision Screener, *IOL* IOL Master biometry. Rows labelled Δ report the marginal MAE reduction attributable to either adding biometry (within a model column) or switching from Ridge to TabPFN (across model columns)

Deployment-Factor Evaluation

For deployment planning we additionally evaluated four factors that are critical to clinical translation: calibration, robustness to input measurement noise, per-eye inference time, and within-patient inter-eye correlation. Detailed numerical tables are provided in the Supplementary Appendix; the headline results are as follows.

Calibration Linear regression of observed against out-of-fold predicted cycloplegic SE yielded calibration slopes within [0.972, 1.028] and intercepts within [−0.083, +0.095] D across all eight {scenario, model} combinations evaluated. Ridge tended to slightly under-disperse on the noisier Eyerobo scenarios (slope 0.972–0.989), while TabPFN slightly over-dispersed (slope 1.016 –1.028). Neither pattern is large enough to require external recalibration in this single-center cohort, but the calibration slope should be re-estimated on any prospective external-validation cohort because calibration is the deployment characteristic most sensitive to spectrum shift.

Robustness Gaussian noise with standard deviation $\sigma = 0.083$ D was added independently to the SPH and CYL inputs of every scenario and the entire fivefold pipeline was re-run. MAE degradation was uniformly small (Δ MAE +0.001 to +0.009 D) for both Ridge and TabPFN across all four headline scenarios, indicating that neither model is brittle to small input perturbations on the order of typical device measurement precision.

Inference time On a single CPU core, Ridge runs in sub-millisecond per eye (effectively instantaneous), and TabPFN runs in 24–49 ms per eye depending on the input dimensionality (8 features for the device-only scenarios; 17 features for the device + IOL scenario). Both regimes are within the operational throughput requirements of any clinical screening workflow. The substantive deployment constraint is data residency rather than throughput: TabPFN performs in-context learning at inference time and therefore requires the training fold to remain co-located with the model, which has implications for federated or on-device deployment.

Inter-eye correlation As a sensitivity analysis to the inclusion of both eyes of most patients, we reran the headline regression after randomly retaining one eye per patient (574 eyes; three independent seeds). The one-eye MAE differed from the headline per-eye MAE by −0.004 to +0.012 D across all evaluated cells. This confirms that the patient-level grouped cross-validation used in the headline analysis is sufficient to control inter-eye correlation and that the regression-MAE estimates reported are not materially inflated by including paired eyes.

DISCUSSION

Intended use of the Proposed Approach

Before discussing the findings, we state explicitly the clinical role we propose for the device–model combination. The Eyerobo Vision Screener combined with a machine-learning prediction layer is intended as a *triage adjunct in settings where cycloplegic refraction is operationally or temporally impractical*, for example, school screening, community screening, or initial workup before a child returns for a formal cycloplegic examination. Cycloplegic refraction remains the gold standard for diagnosis and spectacle prescription, and nothing in this study justifies displacing it. The relevant clinical question that this study informs is whether the Eyerobo combined with the prediction layer can correctly *flag* children who should proceed to cycloplegic examination, not whether it can replace the cycloplegic measurement itself. Subgroup performance for high myopia and hyperopia (Table 7) and the absence of external validation (Limitations) further argue against any use beyond the triage role. We have audited the Abstract, Conclusions, and Key Summary Points sections to align with this scope.

Overview of Findings

We present a 12-scenario symmetric evaluation framework comparing the Eyerobo Vision Screener, a portable handheld photorefractor,

against a conventional desk-mounted auto-refractor for TabPFN-based cycloplegic refraction prediction in 1129 eyes from 574 pediatric patients. To our knowledge, this is the first ML-based cycloplegic prediction study to evaluate a portable device, the first to report power vector components (J0, J45), and the first to apply a tabular foundation model to ophthalmic refraction prediction. The 12-scenario design deliberately evaluates each device in isolation (rather than in combination) because the clinical question is which single refraction device to deploy in a given setting; a cross-device fusion scenario, while technically feasible, does not correspond to a realistic clinical workflow. Similarly, we restricted device features to SPH, CYL, and AX without including the engineered spherical equivalent (SE = SPH + CYL/2) as an input, ensuring an apples-to-apples comparison; preliminary ablation confirmed that TabPFN learns this linear combination internally and is unaffected by its omission.

Eyerobo VS versus Auto-refractor

The auto-refractor produced more accurate SE predictions than the Eyerobo across all configurations (predilation MAE: 0.295 D versus 0.657 D). This gap reflects the higher raw measurement precision of the auto-refractor ($r = 0.978$ versus $r = 0.787$ for SE). However, the Eyerobo showed lower mean accommodative bias (+0.254 D versus +0.374 D for SE), consistent with its longer working distance reducing accommodation in children.

The difference in measurement precision between the two devices arises from fundamental optical principles. The auto-refractor employs wavefront sensing or retinoscopic principles at a short working distance (typically 30–40 cm), with an internal fogging mechanism designed to relax accommodation during measurement. By contrast, the Eyerobo VS uses eccentric photorefraction at approximately one meter, analyzing the light reflex from the retina captured by a camera. While photorefraction provides rapid bilateral measurements without requiring precise patient positioning, it inherently suffers from lower signal-to-noise ratio than direct wavefront interrogation. The wider limits of agreement observed for the Eyerobo (−2.28 to +2.80 D versus −0.47 to +1.22 D for SE, based on raw device agreement) reflect this difference in optical measurement precision rather than systematic bias, as evidenced by the Eyerobo's paradoxically lower mean difference.

The accommodation advantage of the Eyerobo deserves further consideration. The auto-refractor's internal fogging mechanism is designed to relax accommodation, but in children this often fails, particularly in younger patients [6]. The active accommodation in the confined auto-refractor viewing system can be substantial, especially in hyperopic children who have high accommodative reserve. The Eyerobo's longer working distance may provide a more natural fixation target, reducing but not eliminating accommodative artifact. This interpretation is supported by the observation that the mean accommodative bias was smaller for the Eyerobo across all refractive subgroups, though the advantage in mean bias was offset by the Eyerobo's larger random measurement error.

When IOL Master biometry was added, the Eyerobo gap narrowed by 64%. Feature importance analysis revealed that axial length became the dominant predictor for the Eyerobo + IOL scenario (importance = 0.633), surpassing Eyerobo SPH (0.462). This indicates that biometric parameters provide structural information that compensates for the Eyerobo's lower optical precision.

Effect of IOL Master Biometry

The asymmetric benefit of IOL Master biometry to the two devices is one of the most informative findings of this study. Adding biometric data to the Eyerobo reduced predilation MAE by 0.232 D (35.3%), compared with only 0.012 D (4.1%) for the auto-refractor, a 19-fold difference. Feature importance analysis provides a mechanistic explanation for this asymmetry. In the Eyerobo + IOL scenario (S10), axial length achieved the highest permutation importance (0.633), exceeding Eyerobo SPH (0.462). The model effectively substituted the structural measurement

of eye length, which is tightly correlated with refractive state [36, 37], for the Eyerobo's noisier refractive estimate. In the auto-refractor + IOL scenario (S07), auto-refractor SPH retained dominant importance (1.202) with axial length contributing a secondary 0.203, because the auto-refractor already provides a precise refractive measurement that leaves relatively little variance for biometric features to explain.

This finding has practical implications for deployment. In settings where IOL Master biometry is available, combining it with the Eyerobo brings prediction accuracy close to the auto-refractor level (MAE 0.425 D versus 0.295 D). The IOL Master is commonly available in hospital ophthalmology departments but not in community screening settings. For community screening with the Eyerobo alone, the higher MAE of 0.657 D must be accepted, though this still provides useful information for mild myopia screening (MAE = 0.398 D, 75.6% within ± 0.50 D in this subgroup).

Power Vector Analysis

The device gap was substantially smaller for astigmatic components than for SE. Pre-dilation J0 MAE was 0.106 D (AR) versus 0.211 D (ER), and J45 was 0.060 D versus 0.089 D. With IOL biometry, ER J0 improved to 0.143 D, bringing the gap to 0.037 D. The smaller astigmatic gap reflects the narrower dynamic range of cylinder measurements, where both devices capture information with proportionally similar precision. This is clinically relevant because astigmatism assessment is an important component of refractive screening.

The power vector decomposition following Thibos et al. [29] allows separate evaluation of orthogonal and oblique astigmatic components. The J0 component captures with-the-rule (positive J0) versus against-the-rule (negative J0) astigmatism, which is the predominant form of corneal astigmatism in children. The J45 component captures oblique astigmatism, which is typically smaller in magnitude and less clinically consequential for visual function. The Eyerobo's near-equivalent J45 performance (gap of 0.029 D between devices) is clinically relevant because oblique astigmatism at this magnitude does not alter clinical decision-making regarding spectacle prescription or referral thresholds. The larger J0 gap (0.105 D without IOL, 0.037 D with IOL) is more clinically meaningful, as with-the-rule astigmatism influences both visual acuity and myopia progression risk. However, even this gap narrows substantially with the addition of IOL biometry, suggesting that keratometric measurements from the IOL Master provide redundant astigmatic information that supplements the Eyerobo's cylinder estimates.

Comparison with Prior Work

Our auto-refractor post-dilation result (MAE = 0.203 D, 96.0% within ± 0.50 D) compares favorably with published benchmarks: Liu et al. [14] reported 0.360 D, Su et al. [15] 0.307 D, and Ying et al. [16] 0.393–0.480 D. The raw Eyerobo agreement ($r = 0.787$) is consistent with Liu et al. [22]'s report of $r = 0.875$ for the 2WIN-S predecessor, with differences attributable to cohort characteristics and the wider refractive range of our cohort.

Table 10 provides a systematic comparison of our results with published ML-based cycloplegic prediction studies. Several observations emerge from this comparison. First, all prior studies used desk-mounted auto-refractors as input devices; our study is the first to evaluate a portable photorefractor. Second, the best-performing models across studies are consistently gradient boosting variants (XGBoost, random forest, stacked ensembles), with our study being the first to demonstrate TabPFN's competitive performance. Third, studies that incorporate biometric features consistently outperform those using refraction alone, consistent with our observation that IOL Master data provides substantial benefit to the Eyerobo. Fourth, our auto-refractor predilation MAE (0.295 D) is competitive with published benchmarks (Du et al., 0.31 D; Su et al., 0.31 D), despite our use of only three input features (SPH, CYL, AX) compared with the richer

Table 10 Comparison with published ML-based cycloplegic refraction prediction studies

Author	Year	Eyes	Device	ML method	MAE (D)	Features
Du et al.	2023	5920	Auto-refr	XGBoost	0.31	Refr. + biometry
Ying et al.	2024	3414	Auto-refr	XGBoost	0.39	Refr. + biometry
Liu et al.	2025	7366	Auto-refr	Stacked	0.36	Refr. + biometry
Su et al.	2025	1694	Auto-refr	XGBoost	0.31	Refr. + lens
Chen et al.	2025	1440	Auto-refr	RF/XGB	0.35	Refr. + biometry
Hao et al.	2025	–	Auto-refr	RF	0.40	Refraction
Kim et al.	2025	750	IOL Master	Ensemble	0.44	Biometry
Chandra et al.	2025	3414	Auto-refr	XGBoost	0.48	Ext. validation
Present (AR Pre)	2026	1129	Auto-refr	TabPFN	0.30	SPH, CYL, AX
Present (ER Pre)	2026	1129	Eyerobo VS	TabPFN	0.66	SPH, CYL, AX
Present (ER + IOL)	2026	1129	ER + IOL	TabPFN	0.43	SPH, CYL, AX + biom
Present (AR Post)	2026	1129	Auto-refr	TabPFN	0.20	SPH, CYL, AX (post)

MAE mean absolute error for SE prediction in diopters, *RF* random forest

feature sets used in other studies, which often include biometry, lens parameters, and pupil size. Our auto-refractor post-dilation result (0.203 D) is the lowest reported MAE across all published studies, reflecting the strong signal in post-dilation autorefraction where accommodation is pharmacologically eliminated.

Clinical Implications

The Eyerobo VS occupies a different operational position from a visual-acuity chart. A standard near or distance acuity chart is a subjective test that requires literate, attentive cooperation, indicates only that an optical or amblyogenic problem may exist, and does not quantify the refractive state. The Eyerobo VS produces an objective, quantitative refractive estimate in seconds, without requiring patient cooperation or instillation of pharmacological agents, and supports automated severity grading. The clinical value of the device–model combination evaluated here therefore lies not in replacing a refraction examination but in providing a portable, objective triage signal that is more specific than a vision chart and more accessible than a desk-mounted auto-refractor.

The implications of our results for that triage role differ by setting. For community and school screening, where auto-refractors and IOL Master biometers are typically not available, the Eyerobo alone with the ML correction layer flags children for cycloplegic referral with the sensitivity and specificity reported in "Screening-Oriented Performance at Clinical Thresholds"; the regression-MAE numbers (0.657 D predilation, 0.398 D for mild myopia) describe how precisely the predicted SE will track the cycloplegic value, but the operative metric for triage is the threshold-classification performance. For hospital or tertiary ophthalmology settings, where the IOL Master is already available, the combined Eyerobo + IOL configuration narrows the device gap by 64% (MAE 0.425 D versus the auto-refractor's 0.295 D) and is a plausible candidate workflow when an auto-refractor lane is occupied. For clinics already administering mydriatic drops for fundus examination, the auto-refractor

post-dilation prediction (MAE 0.203 D, 96.0% within ± 0.50 D) shifts the question from whether to perform a separate cycloplegic step to whether the existing dilation event already provides equivalent refractive information; this is a question for a prospective comparative-effectiveness study, not a recommendation that this study can sustain.

Cost is one factor in the deployment calculus. The Eyerobo VS is priced at approximately one-tenth the cost of a conventional desk-mounted auto-refractor, and its handheld form factor eliminates the need for a dedicated examination lane. A formal cost-effectiveness analysis incorporating downstream referral rates, false-negative consequences, and program-level operational costs is beyond the scope of the present study; we acknowledge this in the Limitations.

Early myopia detection has clinical value because effective interventions for slowing progression, low-dose atropine, orthokeratology, and defocus-incorporated spectacle lenses, are most effective when initiated early [4]. The Eyerobo's mild-myopia subgroup performance (MAE = 0.398 D, 75.6% within ± 0.50 D) and the screening-oriented metrics in "Screening-Oriented Performance at Clinical Thresholds" are consistent with a triage role in this earliest-detection window. The Eyerobo platform is also being developed for multimodal portable ophthalmic screening: the Eyerobo FC fundus camera has been validated for AI-assisted diabetic retinopathy detection [26], and recent advances in structure-preserving retinal image super-resolution [38] may further enhance portable fundus imaging. Integrating refractive triage with fundus-based assessment through a single portable platform is a promising direction for under-served settings, but each component will require its own prospective external validation before integration.

TabPFN as a Candidate Prediction Layer

TabPFN achieved the lowest MAE among the evaluated algorithms across all Eyerobo scenarios within this single-center cohort and was competitive with Ridge on auto-refractor scenarios. On the headline ER Pre + IOL scenario (S10), TabPFN (0.425 D) outperformed the next-best gradient-boosting learner (XGBoost, 0.542 D) by 0.117 D in per-eye MAE, with paired Wilcoxon signed-rank $p < 0.001$ on per-eye absolute errors. On the auto-refractor scenarios, TabPFN's margin over Ridge was less than 0.02 D in MAE and was not statistically significant by paired Wilcoxon ($p > 0.10$), reflecting the strong linearity of the auto-refractor signal where a simple linear model captures most of the variance. We therefore characterise TabPFN as the best-performing algorithm within this cohort rather than as the definitive algorithm of choice; external validation in a refractive-spectrum-representative cohort is required before any general recommendation can be made.

Several additional properties of TabPFN are relevant to potential deployment. TabPFN handles missing values natively, which is useful given that clinical datasets frequently have incomplete records; missing data rates in our cohort ranged from 1.1 to 6.3% across feature groups. TabPFN produces deterministic predictions within a fixed inference call, in contrast to tree-based ensembles whose random subsampling introduces seed-dependent variation (gradient boosting models in our experiments showed MAE standard deviations of 0.001–0.023 D across seeds; TabPFN and Ridge were deterministic). The learning curve analysis (Fig. 7) shows that TabPFN attains the lowest MAE at all training fractions tested, with no saturation at the full sample size. Three deployment considerations must, however, be set against these advantages. First, TabPFN performs in-context learning at inference time, which means the training set must accompany the model in deployment; this complicates strict data-residency or de-identification regimes. Second, TabPFN does not provide a native global feature-importance summary; we used XGBoost permutation importance as a surrogate interpretability layer. Third, the modest MAE advantage on auto-refractor scenarios means that simpler, lower-compute models such as Ridge may be preferable in some settings; the choice of algorithm should ultimately be governed by external-validation performance and by the operational constraints of the deployment

environment, not by within-cohort ranking alone.

Limitations

Absence of External Validation

The most important limitation of this study is that all model development and evaluation were conducted within a single-center cohort using internal cross-validation. This design is appropriate for preliminary model development, but it does not establish the generalisability or transportability of the prediction layer across different populations, screening environments, or device-software versions. The external-validation study of Chandra et al. [20], which applied the regression models of Ying et al. [16] to an independent cohort, found a degradation in MAE of approximately 0.1 D when models were transported across sites, illustrating the magnitude of generalisation gap that should be expected. The triage rule reported in "Screening-Oriented Performance at Clinical Thresholds" should be regarded as a candidate decision boundary that has been developed within a single-center cohort and that requires prospective external validation before it can be recommended for clinical use. Candidate validation populations include (1) a general-school cohort not recruited through a hospital referral pathway, to estimate true population sensitivity, specificity, PPV and NPV at nonreferral prevalence; (2) a multi-centre Chinese cohort spanning northern, southern, and western regions, to estimate the within-country transportability gap; and (3) an international cohort, to test transportability across ethnicity, optical biometry distributions, and clinical workflow norms. A prospective multi-center validation study is in planning.

Spectrum Bias

The cohort is overwhelmingly myopic (1080 of 1129 eyes; 95.7%) and was recruited through a tertiary-referral pathway in which the indication for presentation is almost always actual or suspected refractive error. Under the Standards for Reporting of Diagnostic Accuracy Studies (STARD) framework this constitutes spectrum bias relative to a general-school screening population, in which the prevalence of any myopia is typically 50–80% and the prevalence of high myopia is correspondingly lower. The numerical implications are twofold. First, our PPV / NPV estimates (Table 8) are tied to the present cohort's threshold prevalences (92.1% / 37.9% / 5.9% at the three referral cutoffs) and will shift when the rule is applied in a population with a different prevalence; sensitivity and specificity are not prevalence-dependent and would, all else being equal, generalize more directly. Second, the small absolute number of hyperopic eyes ($n = 35$ at SE $> +0.5$ D) limits the precision of any inference about how the device–model combination will behave in hyperopic children: the regression-MAE point estimate for the combined nonmyopic stratum ($n = 49$) is 1.021–1.581 D, but this estimate is itself uncertain because the stratum is small, and a hyperopic-only estimate would have wider confidence intervals still. Clinically, the implication is that a triage rule trained primarily on myopes will tend to *under-flag* hyperopic children, who are precisely the children in whom early correction and amblyopia management matter most. We therefore recommend that any deployment of the rule be paired with age-appropriate visual-acuity testing, a step that does not depend on the refractive prediction and that is sensitive to functional consequences of uncorrected hyperopia. External validation in a hyperopia-enriched cohort is a specific prerequisite before the rule can be used as a sole triage instrument in pediatric populations.

Subgroup Safety Considerations

In addition to spectrum bias, the high-myopia subgroup performance of the Eyerobo without biometry (predilation MAE 2.802 D with 9% of predictions within ±0.50 D) reflects the device's effective measurement range; the auto-refractor produces comparable MAE in the same stratum because cycloplegic SE in high myopes exceeds the device's clipping range. The clinical safety

impact of this is partially mitigated by the triage workflow itself, every high-myope, by definition, exceeds the any-myopia threshold and is already flagged for cycloplegic referral, but the degraded magnitude estimate means that the Eyerobo cannot be used to grade severity for follow-up scheduling or for spectacle prescription in this stratum. Adding IOL biometry reduces the high-myopia MAE to 0.800 D, which is still clinically concerning for prescriptive purposes but acceptable for the binary refer/do-not-refer decision.

Inter-eye Correlation

Inter-eye correlation was controlled by GroupK-Fold cross-validation at the patient level, ensuring that both eyes of a given patient are always assigned to the same fold and that no within-patient information leaks across the training/test boundary. To provide additional reassurance, we ran a sensitivity analysis restricted to one randomly selected eye per patient over three seeds; the per-eye headline regression and classification metrics fell within the bootstrap confidence intervals of the per-eye estimates, with point-estimate shifts of less than 0.02 D in MAE and less than 0.005 in AUC at the primary thresholds. This sensitivity analysis is reported in the Supplementary Appendix.

Reproducibility and Operational Characteristics

We did not perform an inter-device reproducibility assessment, so the within-session and between-session test–retest variability of the Eyerobo VS in this cohort is not quantified. We also did not evaluate operator-level variability; in a community-screening deployment, screener throughput and operator effects are non-negligible and should be assessed prospectively. No formal cost-effectiveness analysis was performed; the cost comparison between devices is approximate, and a full health-economic evaluation incorporating referral rates, false-negative consequences, and program-level operational costs is required to substantiate the deployment value proposition.

CONCLUSIONS

Within a single-center pediatric referral cohort of 1129 eyes, the Eyerobo Vision Screener combined with machine learning produced cycloplegic SE estimates with a predilation MAE of 0.657 D, reduced to 0.425 D when IOL Master biometry was added. The auto-refractor produced more accurate SE predictions than the Eyerobo at every configuration (predilation MAE = 0.295 D), and the Eyerobo–auto-refractor gap for astigmatic components was small (J0 gap 0.037 D with IOL biometry, J45 gap 0.029 D). For the mild-myopia subgroup, the Eyerobo achieved an MAE of 0.398 D with 75.6% of predictions within ± 0.50 D. TabPFN achieved the lowest MAE among the evaluated algorithms within this cohort and required no hyperparameter tuning, but the margin over Ridge and gradient-boosting baselines was modest, and we report the screening-oriented sensitivity, specificity, and decision-curve metrics that more directly reflect the clinical referral question ("Screening-Oriented Performance at Clinical Thresholds"). The proposed approach is best positioned as a triage adjunct that flags children for referral to a full cycloplegic examination, not as a substitute for the cycloplegic gold standard. Before the rule can be deployed in school or community screening, prospective external validation in a refractive-spectrum-representative cohort is required.

ACKNOWLEDGEMENTS

We thank all participants and the clinical staff at Tianjin Medical University Eye Hospital and Gaobeidian Mingren Eye Hospital for their time and contributions.

Author Contributions. Muhammad Usama, Wu Guo, Emmanuel Eric Pazo, and Xinjun Ren conceptualized and designed the study. All authors had full access to the data and contributed to the decision to submit the manuscript for publication. Mengmeng Xia, Boxue Yao, Yumei Wang, Fengwei Liang, Wuxia Zhao, Li Qin, Shoukuan

Liu, Muhammad Usama, Zhihui Zhang, Wu Guo, Emmanuel Eric Pazo, and Xinjun Ren contributed to the drafting of the manuscript. Zhihui Zhang, Muhammad Usama, and Emmanuel Eric Pazo performed data collection. Zhihui Zhang, Emmanuel Eric Pazo, and Muhammad Usama performed the data analyses. All authors reviewed the manuscript critically for important intellectual content and approved the final version.

Funding. National Natural Science Foundation of China (grant no. 82160202); Tianjin Health Science and Technology Project, Special Project for High-level Talents (grant no. TJWJ2024RC005); Tianjin Education Commission Research Program Project, Key Project in Natural Sciences (grant no. 2023ZD016); Natural Science Foundation of Tianjin (grant no. 23JCQNJC00690). No funding or sponsorship was received for this article.

Data Availability. The de-identified dataset (1129 eyes, fully anonymized) and all experiment code, including the screening-classification analysis described in "Screening-Oriented Performance at Clinical Thresholds" and the deployment-factor analysis described in "Deployment-Factor Evaluation", are publicly available at https://github.com/Usama1002/cycloplegic-refraction-prediction.

Declarations

Conflict of Interest. Mengmeng Xia, Boxue Yao, Yumei Wang, Fengwei Liang, Wuxia Zhao, Li Qin, Shoukuan Liu, Muhammad Usama, Zhihui Zhang, Wu Guo, Emmanuel Eric Pazo, and Xinjun Ren declare that they have no financial or personal relationships with the manufacturers of the Eyerobo Vision Screener, the Nidek AR-1 autorefractor, or the IOL Master biometer that could be perceived as influencing the work reported in this paper. None of the authors hold equity, consulting positions, patents, royalties, or speaker honoraria with any of the device manufacturers.

Ethical Approval. The study was approved by the ethics board of Gaobeidian Mingren Eye Hospital (2026RN-001) and conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained from each participating child's parent or legal guardian, with assent from children of an appropriate age. All data were de-identified before analysis.

REFERENCES

1. Holden B, et al. Global prevalence of myopia and high myopia and temporal trends from 2000 through 2050. Ophthalmology. 2000;123:1036–42.
2. Baird P, et al. Myopia. Nat Rev Dis Prim. 2020;6:1–20.
3. Tideman JW, Polling JR, Jaddoe V, Vingerling J, Klaver C. Environmental risk factors can reduce axial length elongation and myopia incidence in 6- to 9-year-old children. Ophthalmology. 2019;126:127–36.
4. Wildsoet C, et al. IMI – interventions for controlling myopia onset and progression report. Invest Ophthalmol Vis Sci. 2019;60:M106–31.
5. Morgan I, Iribarren R, Fotouhi A, Grzybowski A. Cycloplegic refraction is the gold standard for epidemiological studies. Acta Ophthalmol. 2015. https://doi.org/10.1111/aos.12642.
6. Mimouni M, Zoller L, Horowitz J, Wygnanski-Jaffe T, Morad Y, Mezer E. Cycloplegic autorefraction in

young adults: is it mandatory? Graefes Arch Clin Exp Ophthalmol. 2016;254:395–8.

7. Williams C, Miller L, Northstone K, Sparrow J. The use of non-cycloplegic autorefraction data in general studies of children's development. Br J Ophthalmol. 2008;92:723–4.

8. Wilson S, Ctori I, Shah R, Suttle C, Conway M. Systematic review and meta-analysis on the agreement of non-cycloplegic and cycloplegic refraction in children. Ophthalmic Physiol Opt. 2022;42:1276–88.

9. Guo X, Shakarchi AF, Block S, Friedman D, Repka M, Collins M. Non-cycloplegic compared with cycloplegic refraction in a Chicago school-aged population. Ophthalmology. 2022. https://doi.org/10.1016/j.ophtha.2022.02.027.

10. Lin Z, et al. The difference between cycloplegic and non-cycloplegic autorefraction and its association with progression of refractive error in Beijing urban children. Ophthalmic Physiol Opt. 2017;37:489–97.

11. Hashemi H, et al. Cycloplegic autorefraction versus subjective refraction: the Tehran Eye Study. Br J Ophthalmol. 2015;100:1122–7.

12. Gopalakrishnan A, et al. The Sankara Nethralaya Tamil Nadu Essilor Myopia (STEM) study: defining a threshold for non-cycloplegic myopia prevalence in children. J Clin Med. 2021. https://doi.org/10.3390/jcm10061215.

13. Du B, et al. Prediction of spherical equivalent difference before and after cycloplegia in school-age children with machine learning algorithms. Front Public Health. 2023. https://doi.org/10.3389/fpubh.2023.1096330.

14. Liu Y, et al. Rapid and accurate prediction of cycloplegic refraction in Chinese children: development and validation of machine learning models. J Glob Health. 2025. https://doi.org/10.7189/jogh.15.04281.

15. Su Q, et al. Predictive modeling of cycloplegic refraction using non-cycloplegia ocular parameters with emphasis on lens-related features. Transl Vis Sci Technol. 2025. https://doi.org/10.1167/tvst.14.10.3.

16. Ying B, Chandra RS, Wang J, Cui H, Oatts JT. Machine learning models for predicting cycloplegic refractive error and myopia status based on non-cycloplegic data in Chinese students. Transl Vis Sci Technol. 2024. https://doi.org/10.1167/tvst.13.8.16.

17. Chen B, et al. Machine learning-driven prediction of cycloplegic refractive error in Chinese children. Front Cell Dev Biol. 2025. https://doi.org/10.3389/fcell.2025.1608494.

18. Hao Y, Wang X, Sun B, Li J, Zhang Y, Jiang S. Machine-learning random forest algorithms predict post-cycloplegic myopic corrections from noncycloplegic clinical data. Optom Vis Sci. 2025;102:138–46.

19. Kim SR, Kang D, Choe G-E, Kim D-H. Ensemble machine learning prediction model for clinical refraction using partial interferometry measurements in childhood. PLoS ONE. 2025. https://doi.org/10.1371/journal.pone.0328213.

20. Chandra RS, Ying B, Wang J, Cui H, Ying G, Oatts JT. Myopia prediction using machine learning: an external validation study. Vision (Basel). 2025. https://doi.org/10.3390/vision9040084.

21. Foo LL, et al. Artificial intelligence in myopia: current and future trends. Eye. 2023;37:3290–7.

22. Liu Z, Pazo EE, Ye H, Yu C, Xu L, He W. Comparing school-aged refraction measurements using the 2WIN-S portable refractor in relation to cycloplegic retinoscopy: a cross-sectional study. J Ophthalmol. 2021;2021:6612476.

23. Thomas J, Rajashekar B, Kamath A, Gogate P. Comparison between Plusoptix A09 and gold standard cycloplegic refraction in preschool children. Oman J Ophthalmol. 2021;14:14–9.

24. Dahlmann-Noor A, et al. Vision screening in children by Plusoptix Vision Screener compared with gold-standard orthoptic assessment. Br J Ophthalmol. 2008. https://doi.org/10.1136/bjo.2008.138115.

25. Peterseim M, et al. The effectiveness of the Spot Vision Screener in detecting amblyopia risk factors. J AAPOS. 2014;18:539–42.

26. Pazo EE, et al. Validation of the Eyerobo FC portable fundus camera for diabetic retinopathy screening using public datasets and deep learning. Ophthalmol Ther. 2026. https://doi.org/10.1007/s40123-026-01353-w.

27. Hollmann N, et al. Accurate predictions on small data with a tabular foundation model. Nature. 2025;637:319–26.

28. Hollmann N, Müller SG, Eggensperger K, Hutter F. TabPFN: a transformer that solves small tabular classification problems in a second. In: International conference on learning representations, 2022.

29. Thibos LN, Wheeler W, Horner D. Power vectors: an application of Fourier analysis to the description and statistical analysis of refractive error. Optom Vis Sci. 1997;74:367–75.

30. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016.

31. Ke G, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Advances in neural information processing systems, 2017. pp. 3146–54.

32. Ostroumova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. In: Advances in neural information processing systems, 2018. pp. 6639–49.

33. Breiman L. Random forests. Mach Learn. 2001;45:5–32.

34. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1:307–10.

35. Pedregosa F, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.

36. Lin H, et al. Prediction of myopia development among Chinese school-aged children using refraction data from electronic medical records: a retrospective, multicentre machine learning study. PLoS Med. 2018. https://doi.org/10.1371/journal.pmed.1002674.

37. Brennan N, Shamp W, Maynes E, Cheng X, Bullimore M. Influence of age and race on axial elongation in myopic children: a systematic review and meta-regression. Optom Vis Sci. 2024. https://doi.org/10.1097/OPX.0000000000002176.

38. Pazo EE, et al. Structure-preserving super-resolution of retinal fundus images via a dual-transformer residual network. Front Med. 2026. https://doi.org/10.3389/fmed.2025.1730678.