Supplementary Materials

Supplementary Table 1. Performance for screening of ADHD

Model	AUROC, mean (95% CI)	Sensitivity, mean (95% CI)	Specificity, mean (95% CI)	PPV, mean (95% CI)	NPV, mean (95% CI)	F1-Score, mean (95% CI)	NLL, mean (95% CI)	Brier score, mean (95% CI)	Youden's Index
Random forest classifier	95.5% (94.6%–96.5%)	87.6% (85.4%–90.0%)	89.2% (87.1%–91.4%)	88.3% (86.6%–90.0%)	89.0% (86.8%–91.4%)	87.8% (85.6%– 90.2%)	0.37 (0.35–0.38)	0.11 (0.10–0.11)	0.53
XGBoost	96.9% (96.2%–97.7%)	91.6% (89.7%–93.7%)	92% (90.1%–93.9%)	91.8% (90.5%–93.2%)	91.9% (90.1%–93.9%)	91.7% (89.8%–93.6%)	0.24 (0.2–0.27)	0.07 (0.06–0.07)	0.41
Extra trees classifier	96.4% (95.6%–97.2%)	90.7% (88.8%–92.8%)	89.2% (87.2%–91.4%)	90.0% (88.5%–91.6%)	89.3% (87.3%–91.6%)	90.6% (88.6%–92.8%)	0.38 (0.36–0.39)	0.11 (0.10–0.12)	0.52
Logistic regression	96.5% (95.7% –97.4%)	91.3% (89.3%–93.4%)	90.4% (88.4%–92.7%)	90.9% (89.4%–92.4%)	90.5% (88.5%–92.7%)	91.2% (89.2%–93.3%)	0.25 (0.22–0.29)	0.07 (0.06–0.08)	0.44

AUROC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value; NLL: negative log-likelihood, XGBoost, extreme gradient boosting classifier.

Supplemental Table 2. AUROC performance results by machine learning models and cross-validation datasets.

Model	Training dataset					Validation dataset					Test Dataset				
	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4
Random Forest	1.00	1.00	1.00	1.00	1.00	0.96	0.94	0.93	0.94	0.95	0.98	0.95	0.93	0.95	0.95
XGBoost	1.00	1.00	1.00	1.00	1.00	0.98	0.95	0.96	0.95	0.96	0.99	0.97	0.95	0.95	0.96
Extra Trees Classifier	1.00	1.00	1.00	1.00	1.00	0.97	0.95	0.95	0.96	0.95	0.98	0.96	0.95	0.96	0.96
Logistic Regression	0.98	0.98	0.98	0.99	0.98	0.96	0.95	0.97	0.95	0.95	0.97	0.97	0.94	0.92	0.97

AUROC, area under the receiver operating characteristic curve;

The dataset used in our analysis was divided into five subsets using patient-level five-fold cross-validation. For each fold, an average of 193.6 participants with 331.2 images from the ADHD group and 193.4 participants with 331.6 images from the TD group were used for training. Validation involved 64.8 participants with 112 ADHD images and 65.0 participants with 111.6 TD images while testing involved 64.6 participants with 110.8 ADHD images and 64.6 participants with 110.8 TD images. The small differences in the average number of participants are caused by stratification by patient level, not image level, and the selection of single-sample patients.

Supplementary Table 3. Statistical review of feature values between ADHD and TD participants

Category	Feature	Description	ADHD	TD	p-value
Disc-related Parameters	Disc_height	Vertical dimensions of the optic disc	93169 (10498)	101023 (16945)	<0.001
	Disc_width	Horizontal dimensions of the optic disc	84432 (10245)	89663 (13295)	<0.001
	Cup_height	Vertical dimensions of the optic cup	41777 (10366)	54834 (18131)	<0.001
	Cup_width	Horizontal dimensions of the optic cup	40661 (9963)	50571 (13986)	<0.001
	CDR_vertical	Ratio of cup-to-disc vertical dimensions	0.44 (0.08)	0.534 (0.104)	<0.001
	CDR_horizontal	Ratio of cup-to-disc horizontal dimensions	0.48 (0.08)	0.558 (0.092)	<0.001
Global Vascular Metrics	Fractal_dimension	Complexity of retinal vascular branching.	1.5 (0.03)	1.519 (0.036)	<0.001
	Vessel_density	Proportion of the area occupied by vessels.	0.08 (0.01)	0.078 (0.011)	<0.001
	Average_width	Mean vessel diameter.	3451 (251)	3333 (425)	<0.001
	Distance_tortuosity	Ratio of vessel path length to its straight-line distance.	3.2 (1.0)	2.838 (0.801)	<0.001
	Squared_curvature_tortuosity	Vessel curvature emphasizing sharp changes.	19.1 (17.1)	15.081 (12.324)	<0.001
	Tortuosity_density	Density of vessel tortuosity in a given area.	0.7 (0.03)	0.686 (0.032)	0.51
Artery-Specific Metrics	Artery_fractal_dimension	Complexity of arterial branching.	1.3 (0.04)	1.316 (0.048)	<0.001
	Artery_vessel_density	Area covered by arteries.	0.034 (0.005)	0.037 (0.007)	<0.001
	Artery_average_width	Mean arterial diameter.	3688 (309)	3706 (484)	0.48
	Artery_distance_tortuosity	Distance-based arterial tortuosity.	4.8 (2.9)	3.900 (2.077)	<0.001
	Artery_squared_curvature_tortuosity	Arterial curvature measure using squared curvature.	51 (79)	35.886 (68.134)	<0.001
	Artery_tortuosity_density	Density of arterial tortuosity.	0.7 (0.05)	0.725 (0.052)	<0.001
Vein-Specific Metrics	Vein_fractal_dimension	Complexity of venous branching.	1.33 (0.037)	1.341 (0.048)	0.004
	Vein_vessel_density	Area covered by veins.	0.043 (0.006)	0.042 (0.009)	0.004
	Vein_average_width	Mean venous diameter.	3814 (294)	3853 (486)	0.108
	Vein_distance_tortuosity	Distance-based venous tortuosity.	3.4 (1.4)	2.876 (1.150)	<0.001
	Vein_squared_curvature_tortuosity	Venous curvature measure using squared curvature.	22 (21.8)	16 (16.9)	<0.001
	Vein_tortuosity_density	Density of venous tortuosity.	0.7 (0.04)	0.74 (0.04)	0.578
B Zone-Specific Metrics	Artery_fractal_dimension_zone_B	Complexity of arterial branching in zone B.	1.01 (0.055)	0.99 (0.059)	<0.001
	Artery_vessel_density_zone_B	Proportion of the area covered by arteries in zone B.	0.005 (0.001)	0.004 (0.001)	<0.001
	Artery_average_width_zone_B	Mean arterial vessel width in zone B.	3997 (456)	4384 (754)	<0.001
	Artery_distance_tortuosity_zone_B	Distance-based tortuosity for arteries in zone B.	2.61 (2.07)	2.52 (2.42)	0.53

	Artery_squared_curvature_tortuosity_zone_B	Curvature measure of arteries using squared curvature in zone B.	9.23 (14.05)	8.81 (22.45)	0.71
	Artery_tortuosity_density_zone_B	Density of arterial vessel tortuosity in zone B.	0.69 (0.11)	0.71 (0.11)	0.009
	Vein_fractal_dimension_zone_B	Complexity of venous branching in zone B.	1.03 (0.05)	1.02 (0.05)	<0.001
	Vein_vessel_density_zone_B	Proportion of the area covered by veins in zone B.	0.006 (0.001)	0.005 (0.001)	<0.001
	Vein_average_width_zone_B	Mean venous vessel width in zone B.	4344 (464)	4591 (701)	<0.001
	Vein_distance_tortuosity_zone_B	Distance-based tortuosity for veins in zone B.	2.61 (1.94)	2.503 (2.044)	0.373
	Vein_squared_curvature_tortuosity_zone_B	Curvature measure of veins using squared curvature in zone B.	8.98 (14.2)	8.740 (17.398)	0.802
	Vein_tortuosity_density_zone_B	Density of venous vessel tortuosity in zone B.	0.66 (0.097)	0.667 (0.101)	0.186
Central Retinal Vessel Equivalents and Ratios in B zone	CRAE_Hubbard_zone_B	Central Retinal Arteriolar Equivalent calculated using the Hubbard method within Zone B.	8938 (1286)	9278 (2539)	0.005
	CRAE_Knudtson_zone_B	Central Retinal Arteriolar Equivalent calculated using the Knudtson method within Zone B.	7933 (1159)	8221 (2266)	0.008
	CRVE_Hubbard_zone_B	Central Retinal Venular Equivalent calculated using the Hubbard method within Zone B.	10551 (1067)	10971 (2217)	<0.001
	CRVE_Knudtson_zone_B	Central Retinal Venular Equivalent calculated using the Knudtson method within Zone B.	11675 (1220)	12118 (2495)	<0.001
	AVR_Hubbard_zone_B	Arteriolar-to-Venular Ratio calculated using the Hubbard method within Zone B	-13 (333)	-68 (731)	0.109
	AVR_Knudtson_zone_B	Arteriolar-to-Venular Ratio calculated using the Knudtson method within Zone B	-11 (294)	-59 (631)	0.113
C Zone-Specific Metrics	Artery_fractal_dimension_zone_C	Complexity of arterial branching in zone C.	1.14 (0.046)	1.152 (0.056)	<0.001
	Artery_vessel_density_zone_C	Proportion of the area covered by arteries in zone C.	0.014 (0.003)	0.013 (0.003)	<0.001
	Artery_average_width_zone_C	Mean arterial vessel width in zone C.	3944 (384)	4195 (665)	<0.001
	Artery_distance_tortuosity_zone_C	Distance-based tortuosity for arteries in zone C.	3.89 (2.95)	3.688 (3.084)	0.276
	Artery_squared_curvature_tortuosity_zone_C	Curvature measure of arteries using squared curvature in zone C.	29 (48)	28.765 (75.434)	0.903
	Artery_tortuosity_density_zone_C	Density of arterial vessel tortuosity in zone C.	0.73 (0.075)	0.735 (0.077)	0.625
	Vein_fractal_dimension_zone_C	Complexity of venous branching in zone C.	1.17 (0.044)	1.182 (0.054)	<0.001
	Vein_vessel_density_zone_C	Proportion of the area covered by veins in zone C.	0.016 (0.003)	0.015 (0.004)	<0.001
	Vein_average_width_zone_C	Mean venous vessel width in zone C.	4218 (368)	4330 (585)	<0.001

	Vein_distance_tortuosity_zone_C	Distance-based tortuosity for veins in zone C.	3.3 (2.02)	2.8 (1.71)	<0.001
	Vein_squared_curvature_tortuosity_zone_C	Curvature measure of veins using squared curvature in zone C.	20 (32)	14 (23)	0.002
	Vein_tortuosity_density_zone_C	Density of venous vessel tortuosity in zone C.	0.714 (0.063)	0.711 (0.062)	0.462
Central Retinal Vessel Equivalents and Ratios in C zone	CRAE_Hubbard_zone_C	Central Retinal Arteriolar Equivalent calculated using the Hubbard method within Zone C.	10017 (1047)	10744 (1824)	<0.001
	CRAE_Knudtson_zone_C	Central Retinal Arteriolar Equivalent calculated using the Knudtson method within Zone C.	8951 (949)	9603 (1649)	<0.001
	CRVE_Hubbard_zone_c	Central Retinal Venular Equivalent calculated using the Hubbard method within Zone C.	12000 (964)	12730 (1933)	<0.001
	CRVE_Knudtson_zone_c	Central Retinal Venular Equivalent calculated using the Knudtson method within Zone C.	13401 (1083)	14221 (2181)	<0.001
	AVR_Hubbard_zone_c	Arteriolar-to-Venular Ratio calculated using the Hubbard method within Zone C.	0.837 (0.087)	0.846 (0.091)	0.09
	AVR_Knudtson_zone_c	Arteriolar-to-Venular Ratio calculated using the Knudtson method within Zone C.	0.67 (0.07)	0.68 (0.08)	0.082

ADHD, attention-deficit/hyperactivity disorder; TD, typical development; CRAE, central retinal arteriolar equivalent; CRVE, central retinal venular equivalent; AVR, arteriolar-venular ratio; zone B, the annulus 0.5-1 optic disc diameter from the disc margin; zone C, the annulus 0.5-2 optic disc diameter from the disc margin.

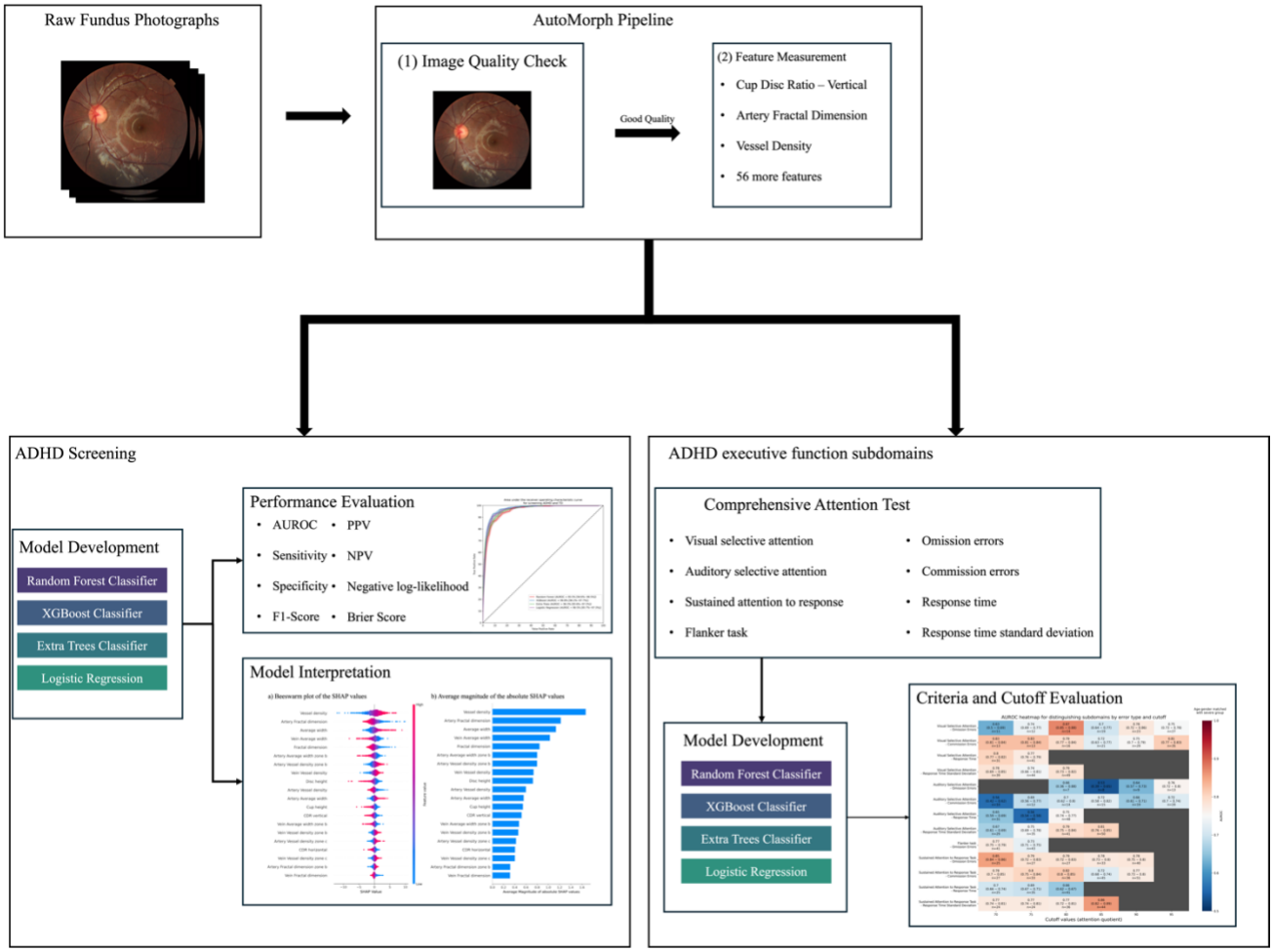
Supplementary Table 4. Hyperparameter setting for machine learning models

ML algorithms	Hyperparameter
Random Forest	RandomForestClassifier(n_estimators = 1000, random_state=2024, n_jobs=8, verbose=-1)
XGBoost	XGBClassifier(n_estimators = 1000, random_state=2024, n_jobs=8, verbose=-1)
Extra Trees Classifier	ExtraTreesClassifier(n_estimators = 1000, random_state=2024, n_jobs=8, verbose=0)
Logistic Regression	LogisticRegression(random_state=2024, n_jobs=8, verbose=0)

Tree-based models, such as Random Forest (RF), XGBoost(XGB), and Extra Trees (EXT). /The number of estimators in the ensemble has a significant impact on these models. We raised each model's default number of trees (n_estimators) from 100 to 1,000 in order to improve generalization performance.

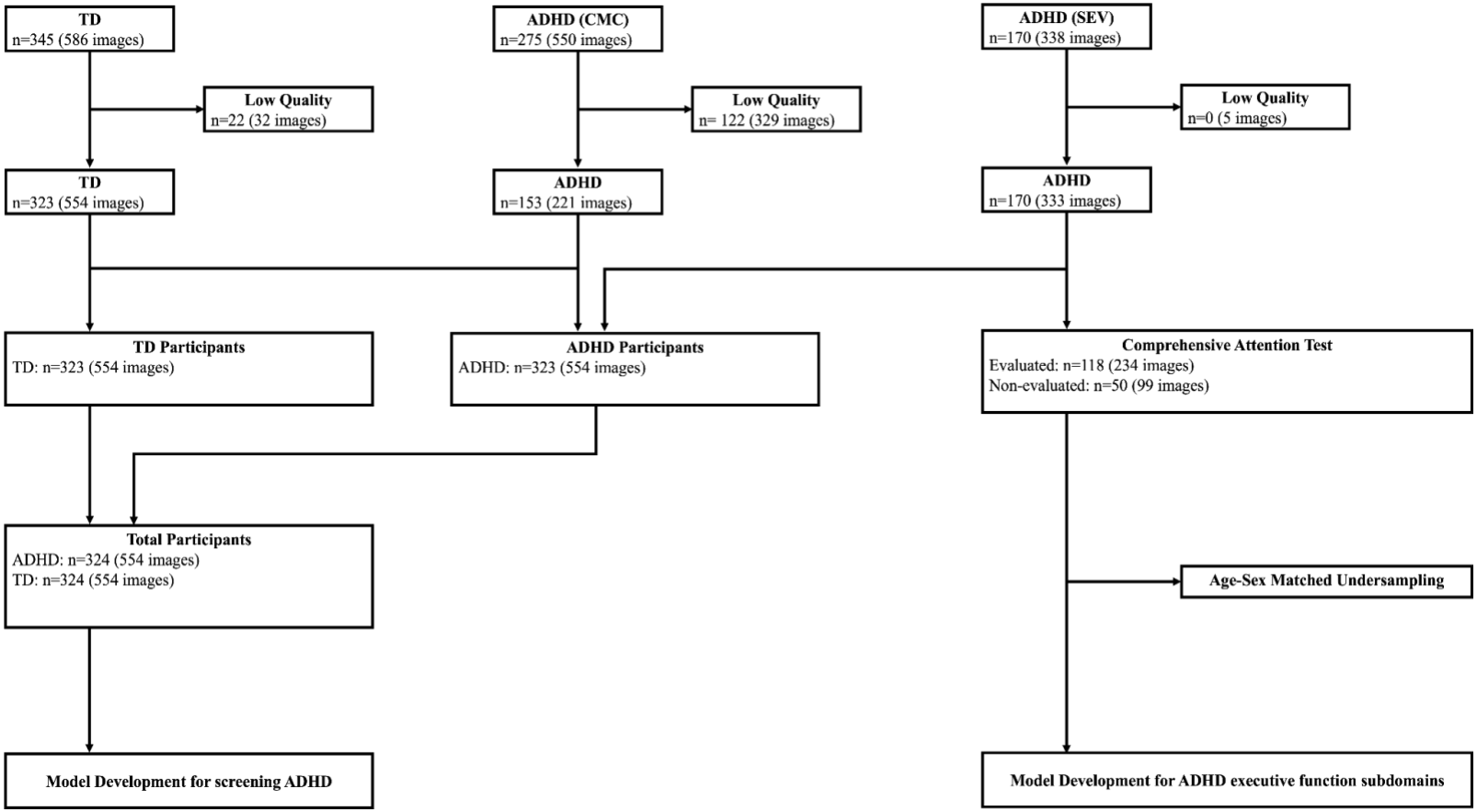
We evaluated 4 different machine learning algorithms-random forest classifier (RF), extreme gradient boosting classifier (XGBoost), extra trees classifier (EXT), and logistic regression (LR) for the development of the screening of ADHD from TD and prediction of EF subdomains in ADHD using CAT scores, AQ scores, and K-ARS scores. Scikit-Learn (version 1.4.1), XGBoost (version 2.0.3) were used to construct these models in Python (version 3.9.15).

Supplementary Figure 1. Overview of the study



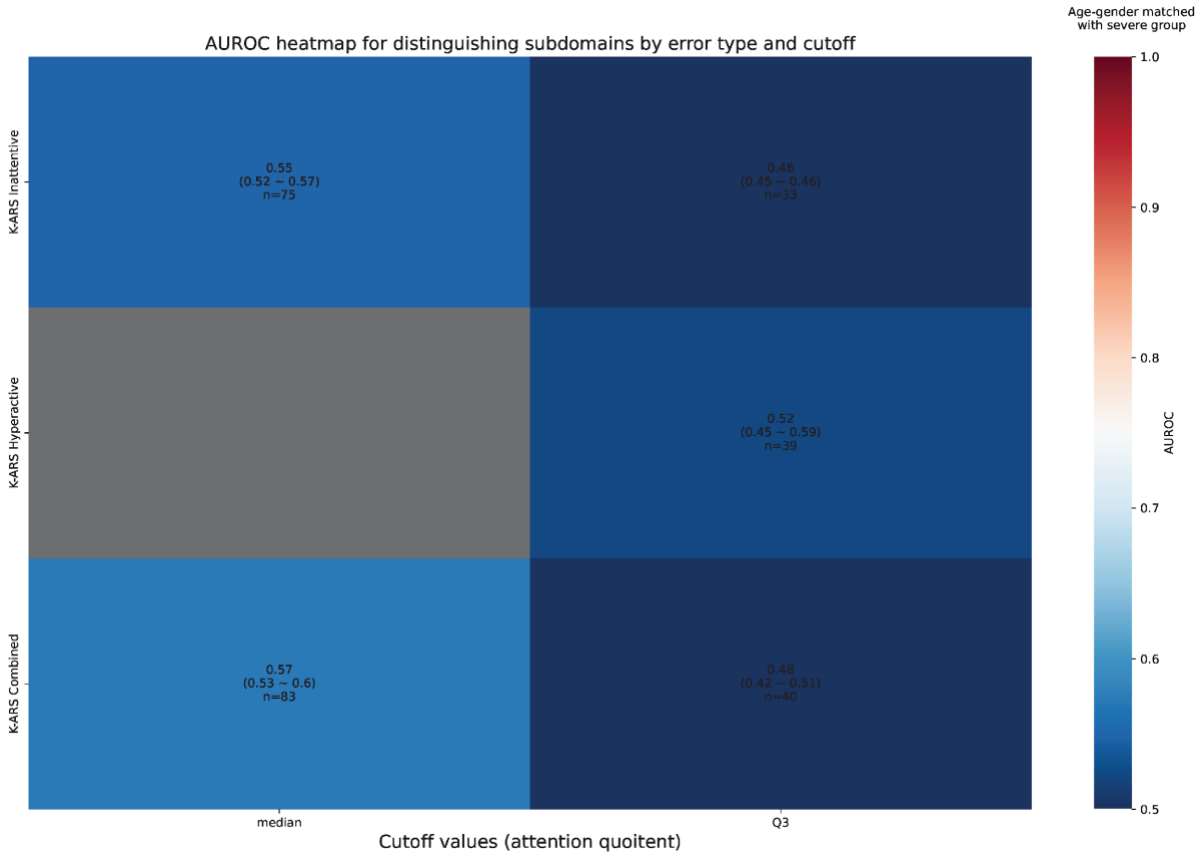
This figure provides a comprehensive overview of the study workflow, focusing on the methodology used for attention-deficit/hyperactivity disorder(ADHD) screening and the subdomains of executive function. We used AutoMorph pipeline to remove bad-quality images and extract feature values. We developed four types of machine learning models, random forest classifier, XGboost classifier, extra trees classifier, and logistic regression, using the feature values for two objectives.

Supplementary Figure 2. Study dataset flow diagram



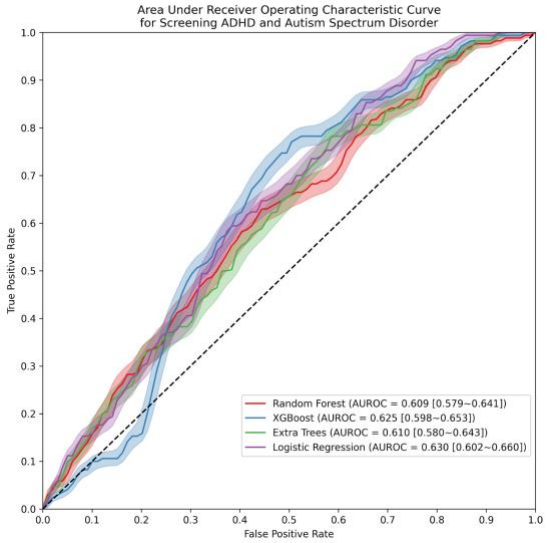
The study flow diagram shows the number of participants from the initial to the final analysis. Two different groups of typical development (TD) and attention-deficit/hyperactivity disorder (ADHD) individuals were collected at Severance Hospital (SEV) and Catholic Medical Center (CMC). They all took retinal photography, and the images with low quality were dropped out with AutoMorph image quality assessment. The remaining 323 ADHD individuals and age- and sex-matched TD individuals were used for screening ADHD, and 118 individuals were used for subdomains of executive function deficits.

Supplemental Figure 3. Performance for predicting severity using the Korean ADHD Rating Scale based on retinal features



The heatmap shows AUROC values for the Korean-ADHD Rating Scale(K-ARS) across median and Q3 cutoff values, with age-gender matching to the severe group. The models predicting severity based on the K-ARS scores were less successful, with AUROC values of < 60%. K-ARS Combined showed moderated performance at the median cutoff with an average AUROC of 0.57.

Supplementary Note 1. Performances for discriminating between ASD and ADHD



Supplementary Figure 4. Performance for discriminating autism spectrum disorder and attention-deficit/hyperactivity disorder. This figure shows AUROC values for screening attention-deficit/hyperactivity disorder (ADHD) and autism spectrum disorder (ASD). The models screening the disorders were less successful with AUROC values of < 65%. The logistic regression model showed a moderated performance with an average AUROC value of 63%.

	AUROC, mean (95% CI)	Sensitivity, mean (95% CI)	Specificity, mean (95% CI)	F1-Score, mean (95% CI)	PPV, mean (95% CI)	NPV, mean (95% CI)	NLL, mean (95% CI)	Brirer, mean (95% CI)
Random Forest	60.9% (58.0%-64.1%)	14.7% (11.2%-18.1%)	90.4% (88.9%-92.0%)	19.6% (15.2%-23.5%)	29.4% (22.8%-35.7%)	79.6% (77.7%-81.4%)	0.56 (0.54~0.57)	0.19 (0.18~0.19)
XGBoost	62.5% (59.8%-65.3%)	24.7% (20.5%-29.0%)	77.5% (75.5%-79.7%)	23.8% (19.9%-27.6%)	23.0% (19.0%-26.8%)	79.1% (77.2%-81.4%)	0.77 (0.72~0.82)	0.20 (0.22-0.25)
Extra Trees	61.0% (58.0%-64.3%)	13.5% (10.0%-16.8%)	91.4% (90.0%-92.8%)	18.6% (14.0%-22.6%)	29.9% (22.6%-36.4%)	79.6% (77.6%-81.4%)	0.56 (0.54~0.57)	0.19 (0.18~0.19)
Logistic Regression	63.0% (60.2%-66.0%)	18.8% (14.7%-22.4%)	87.2% (85.5%-89.0%)	22.7% (17.9%-26.5%)	28.6% (22.5%-33.8%)	79.8% (77.9%-81.9%)	0.56 (0.54~0.58)	0.19 (0.18~0.20)

Supplementary Table 5. Overall performance for classification of autism spectrum disorder and attention-deficit/hyperactivity disorder through retinal features

All ML models for distinguishing children with ADHD and autism spectrum disorder (ASD) showed relatively low performances. RF, XGBoost, EXT, and LR models achieved AUROC values of 60.9% (95% CI 58.0% - 64.1%), 62.5% (95% CI 59.8% - 65.3%), 61.0% (95% CI 58.0% - 64.3%), 63.0% (95% CI 60.2% - 66.0%), respectively (Supplementary Figure 5). The best-performing model, Logistic

Regression, showed a sensitivity of 18.8% (95% CI 14.7% - 22.4%), a specificity of 87.2% (95% CI 85.5%–89.0%), an F1-score of 22.7% (95% CI 17.9% - 26.5%), an PPV of 28.6% (95% CI 22.5% - 33.8%), and an NPV of 79.8% (95% CI 77.9% - 81.9%). The model showed calibration with 0.56 (95% CI 0.54 – 0.58), and the Brier score of 0.18 (95% CI 0.18 – 0.20).

Supplementary Note 2. Performance evaluation metrics

The probability of the instances in our dataset needs to be changed into either positive or negative, using a threshold. So that, each predicted binary label has therefore four possible designations: a true positive (TP) is a correctly predicted positive outcome, a true negative (TN) is a correctly predicted negative outcome, a false positive (FP) is a negative instance predicted to be positive, and a false negative (FN) is a positive instance predicted to be negative. In this study, we used Youden-index to set the optimal threshold Youden-index is determined by the number that achieved the highest Youden's J statistic. We tested the threshold from 0.01 to 0.99 to figure out the optimal Youden-index using the validation dataset and applied the value to the test dataset to appropriately evaluate the performances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = Recall = \frac{TP}{TP + FN} \in [0,1]$$

$$Specificity = \frac{TN}{TN + FP} \in [0,1]$$

$$PPV = \frac{TP}{TP + FP} \in [0,1]$$

$$NPV = \frac{TN}{TN + FN} \in [0,1]$$

$$Precision = \frac{TP}{TP + FP} \in [0,1]$$

$$F - Score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \in [0,1]$$

$$J = sensitivity + specificity - 1$$

$$J_{max} = max_t\{sensitivity(t) + specificity(t) - 1\}$$

On the other hand, performance metrics, including AUROC, the NLL, and the Brier score, are not evaluated with hard prediction labels but with soft probabilities. NLL is the measure of how well a probabilistic model predicts the true outcomes. The lower NLL value indicates that the model's predicted probabilities closely align with the actual values, and a higher NLL value indicates the opposite, leading to worse predictions. The Brier score is a metric used to evaluate the accuracy of probabilistic predictions. This metric measures the mean squared error (MSE) between the predicted probabilities and the actual outcomes and evaluates both the calibration and discrimination.

$$NLL = -\sum_{i=1}^{N} log\, P(y_i | x_i, \theta)$$

$$Brier\ Score = \frac{1}{N}\sum_{i=1}^{N} \sum_{c=1}^{C} \left(p_{i,c} - y_{i,c}\right)^2$$

Supplementary Note 3 STARD 2015 Checklist

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	1
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	2
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	3-5
	4	Study objectives and hypotheses	5
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	13
Participants	**6**	Eligibility criteria	13
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, and inclusion in the registry)	13
	8	Where and when potentially eligible participants were identified (setting, location, and dates)	13
	9	Whether participants formed a consecutive, random, or convenience series	13
Test methods	**10a**	Index test, in sufficient detail to allow replication	13-15
	10b	Reference standard, in sufficient detail to allow replication	13-15
	11	Rationale for choosing the reference standard (if alternatives exist)	14-15
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	-
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	13
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	-
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	14
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	14
	15	How indeterminate index test or reference standard results were handled	14-15
	16	How missing data on the index test and reference standard were handled	14-15
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	14-15
	18	Intended sample size and how it was determined	14
RESULTS			
Participants	**19**	Flow of participants, using a diagram	Supplemental Figure 4
	20	Baseline demographic and clinical characteristics of participants	Table 1
	21a	Distribution of severity of disease in those with the target condition	5-6
	21b	Distribution of alternative diagnoses in those without the target condition	-

	22	Time interval and any clinical interventions between index test and reference standard	-
Test results	**23**	Cross-tabulation of the index test results (or their distribution) by the results of the reference standard	-
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	6-7
	25	Any adverse events from performing the index test or the reference standard	6-7
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	11-12
	27	Implications for practice, including the intended use and clinical role of the index test	12-13
OTHER INFORMATION			
	28	Registration number and name of registry	1
	29	Where the full study protocol can be accessed	15
	30	Sources of funding and other support; role of funders	16

Supplementary Note 4. Detailed information on the computerized comprehensive attention test

The computerized comprehensive attention test (CAT) is a computer-based continuous performance test (CPT) for individuals aged 4–49 years old. The CAT has four subsets: visual selective attention (VSA), auditory selective attention (ASA), sustained attention to response task (SART), and Flanker task (FT). Each visual or auditory stimulus was presented on a black screen at 2-s intervals. In the VSA test, participants were instructed to press a button as quickly as possible whenever a circle appeared while ignoring other shapes. In total, 150 stimuli were presented, of which 75 were circles. In the ASA test, participants were instructed to press a button as quickly as possible upon hearing a specific bell sound while ignoring other auditory stimuli. In total, 150 stimuli were presented, of which 75 were the target sounds. In the SART, participants were instructed to ignore only the X shapes and press the button as quickly as possible for all other shapes. In total, 300 stimuli were presented, of which 75 were X shapes. In the FT, participants were instructed to press the arrow key (left or right) corresponding to the direction of the open side of the middlebox as quickly as possible when five boxes with one open side appeared on the screen. In total, 150 stimuli were presented. Each subset exhibits four indicators: omission errors (OE) for the number of missed responses, commission errors (CE) for the number of wrong responses, mean reaction time (RT) for the average response time to the stimuli, and standard deviation of reaction time (RTSD) for response time variability. These measurements, 16 in total, were normalized to obtain an attention quotient (AQ), adjusting for sex and age, with an average value of 100 and a standard deviation of 15.