International Journal of Research in Medical Sciences
Asiquzzaman M et al. Int J Res Med Sci. 2026 Dec;14(12):5612-5619
www.msjonline.org
pISSN 2320-6071 | eISSN 2320-6012
DOI: https://dx.doi.org/10.18203/2320-6012.ijrms20264459

Original Research Article

Machine learning–based screening for autism spectrum disorder in early childhood: diagnostic accuracy and feasibility

M. Asiquzzaman[1]*, Naima Sultana[2], Jahan Ahmed Porag[3], G. M. Mohaimenul Islam[4], M. Nazmul Hossain[5], Mohammad Abed Abbas[6]

[1]Rajshahi Medical College, Rajshahi, Bangladesh
[2]Ispahani Cantonment Public School and College, Cumilla, Bangladesh
[3]Jalalabad Ragib-Rabeya Medical College, Sylhet, Bangladesh
[4]Department of Pediatric Neurology, National Institute of Neurosciences and Hospital, Dhaka, Bangladesh
[5]Lalmai Upazila Health Complex, Cumilla, Bangladesh
[6]Department of Neurosurgery, Bangladesh Medical University, Dhaka, Bangladesh

Received: 14 September 2026
Revised: 16 October 2026
Accepted: 21 October 2026

***Correspondence:**
Dr. M. Asiquzzaman,
E-mail: dr.asiquzzaman@proton.me

Copyright: © the author(s), publisher and licensee Medip Academy. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT

Background: Autism spectrum disorder (ASD) can be reliably identified in the second year of life, yet diagnosis is often delayed where specialist assessment capacity is limited. We evaluated the diagnostic accuracy and feasibility of a machine learning (ML)–based screening pipeline combining a brief caregiver questionnaire with automated video-derived behavioral features.
Methods: This cross-sectional study included 156 children aged 12–36 months from four child-development sites. Each child completed a 10-item caregiver questionnaire and, where feasible, an app-guided video from which five behavioral features were extracted automatically. The reference standard was clinician diagnosis informed by ADOS-2. Logistic regression, random forest, gradient boosting and support-vector models were trained on questionnaire-only and multimodal feature sets and evaluated with 20 repetitions of stratified five-fold cross-validation. Outcomes were AUC, sensitivity, specificity, predictive values and feasibility indicators.
Results: Twenty-nine children (18.6%) had ASD. The multimodal model outperformed the questionnaire-only model (AUC 0.98 vs 0.85; ΔAUC 0.13, 95% CI 0.07–0.20; $p<0.001$), achieving a sensitivity of 0.93, specificity of 0.99 and negative predictive value of 0.99. Reduced social orienting and gesture count were the strongest predictors. The questionnaire required a median of 6.2 minutes, usable video was obtained in 90.6% of attempts, and 68.6% of caregivers rated usability ≥4/5.
Conclusions: A multimodal ML screening pipeline achieved high discrimination for early-childhood ASD and was feasible with minimal staff time, with automated video features adding substantial accuracy over caregiver report alone. Prospective external validation is required before clinical deployment.

Keywords: Autism spectrum disorder, Early screening, Machine learning, Digital phenotyping, Diagnostic accuracy, Feasibility, Child development

INTRODUCTION

Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized by persistent differences in social communication together with restricted, repetitive patterns of behavior and interests.[1] Population surveillance indicates that roughly one in 30–40 children are affected, and behavioral signs are frequently detectable before the

second birthday.[2] Early identification matters because timely, developmentally focused intervention during a period of high neuroplasticity is associated with improved communication, adaptive and social outcomes.[3]

Despite guidance recommending developmental surveillance at every well-child visit and formal autism-specific screening at 18 and 24 months, a large gap persists between the age at which reliable identification is possible and the age at which children actually receive a diagnosis.[4] The gap is widest in low- and middle-income settings, where specialist developmental assessment is concentrated in a small number of tertiary centers, waiting lists are long, and caregivers may travel considerable distances for evaluation.[5] In Bangladesh, as in many comparable settings, the number of children who could benefit from evaluation greatly exceeds the capacity of trained diagnosticians, so a substantial proportion of children are identified only after the optimal window for early intervention has narrowed.[6] Given these systemic constraints, augmenting traditional screening with automated, scalable digital tools may effectively reduce the burden on specialists by facilitating early risk stratification in frontline pediatric care.[7,8] Recent advances in computer vision and behavioral signal processing suggest that such automated approaches can yield diagnostic performance metrics comparable to established clinical instruments.[9,10] Integrating these machine learning algorithms with standard behavioral observations allows for the identification of objective biomarkers that can refine and accelerate the traditional diagnostic trajectory.

Conventional caregiver-report screening instruments are inexpensive and easy to deploy, but their real-world performance is constrained by reliance on subjective recall, variable literacy, and limited specificity, which generates large numbers of low-yield referrals into already-saturated diagnostic services.[11] Two complementary developments offer a path forward. First, inexpensive smartphones and tablets are now widely available even in resource-limited communities, enabling app-guided collection of standardized stimuli and short video recordings at the point of care. Second, machine learning (ML) applied to automatically extracted behavioral features so-called digital phenotyping can quantify gaze, orienting, gesture and movement in a reproducible, examiner-independent way. Combining structured caregiver report with objective, automatically scored behavioral markers could improve screening accuracy while preserving the low cost and scalability that make screening viable.[12] Digital tools represent a promising mechanism to bridge the diagnostic gap, as they can automate the analysis of complex behavioral patterns without requiring extensive training in computer vision or clinical statistics.[13] By leveraging these data-driven algorithms, such mobile screening frameworks can effectively democratize access to diagnostic services, thereby transcending geographical and socioeconomic barriers that frequently impede early detection.[14] Furthermore, by integrating these digital biomarkers into the standard waiting list process, clinicians can obtain ecologically valid data prior to formal assessment, which facilitates the prioritization of children exhibiting more severe clinical presentations.[15] Nevertheless, most reported ML autism classifiers have been developed on convenience samples, evaluated without rigorous internal validation, or reported only discrimination without the calibration, predictive values and implementation data that determine whether a test is usable in practice.[16] Diagnostic-accuracy metrics also depend critically on prevalence, and enriched research samples can give misleadingly optimistic predictive values relative to community screening.[17] Moreover, shifting toward larger, more representative cohorts is essential to address the current lack of diversity in demographic data and ensure the generalizability of these diagnostic models.[18] Beyond demographic diversity, it is imperative that future research evaluates the performance of these classifiers in real-world clinical workflows to ensure that technical accuracy translates into improved patient outcomes.[19]

We therefore set out to (i) quantify the diagnostic accuracy of an ML-based screening pipeline that fuses a brief caregiver questionnaire with automated video behavioral features, benchmarked against a clinician reference standard; (ii) test whether the automated video component adds accuracy beyond caregiver report alone; and (iii) assess the practical feasibility of the pipeline in a pediatric child-development service, including administration time, video-capture success and caregiver usability.

METHODS

Study design and setting

This was an analytical, cross-sectional diagnostic-accuracy and feasibility study conducted at the Child Development Centre (Shishu Bikash Kendra) of National Shishu Hospital, Dhaka, together with three collaborating child-development units (designated sites A–D). The study was conducted from February 2025 to February 2026, and children were enrolled consecutively from developmental-assessment clinics during this period. Reporting follows the STARD 2015 guidance for diagnostic-accuracy studies and the TRIPOD statement for prediction-model development and validation.

Participants

Children aged 12–36 months who attended the participating clinics for a developmental concern or for routine developmental follow-up were eligible for inclusion, provided that a caregiver gave written informed consent.

Children were excluded from the study if they had a previously confirmed genetic syndrome or a major sensory impairment, such as severe hearing or visual impairment, that precluded standard developmental and observational assessment. Children whose caregivers declined to give

informed consent were not enrolled. Children in whom a usable video could not be obtained were not excluded; they were retained in the analysis, and their video features were treated as missing, as described below.

A total of 156 children formed the analysis cohort. Because the cohort was assembled to include an adequate number of ASD cases for model development, the observed ASD proportion is enriched relative to community prevalence; the implications of this for predictive values are addressed explicitly in the analysis and discussion.

Index test: the screening pipeline

Caregiver questionnaire

Caregivers completed a 10-item structured questionnaire administered on a tablet or smartphone, each item probing an early social-communication or behavioral domain (response to name, eye contact, pointing/showing, pretend play, social smile, repetitive behaviors, shared interest, response to others' emotion, sensory sensitivity, and imitation). Each item was scored 0 (typical) or 1 (concern), and items were summed to a total screen score (range 0–10).

Automated video behavioral features

Where the child cooperated and the environment allowed, a short standardized, app-guided video was recorded while developmentally appropriate social stimuli were presented. Five behavioral features were extracted automatically by the analysis software: social-orienting score and eye-contact proportion (attention to social versus non-social content), communicative gesture count, facial-affect variability, and a motor-stereotypy index (0–1 scales, except gesture count). Video was captured on the caregiver's personal device or a clinic tablet. When video could not be obtained or was not usable, the child was retained in the analysis, and the video features were treated as missing.

Reference standard

The reference standard was a clinical best-estimated diagnosis of ASD versus non-ASD, made by experienced developmental clinicians and informed by the Autism Diagnostic Observation Schedule, Second Edition (ADOS-2), with the calibrated severity score recorded. Reference-standard assessors were independent of the automated feature-extraction pipeline and its outputs.

Feasibility outcomes

Prespecified implementation outcomes were questionnaire-completion time (minutes), the proportion of children in whom usable video was obtained among those attempted, caregiver-rated usability (1–5 Likert scale), clinician review time (minutes), the device type used, and downstream service outcomes (referral generation and initiation of intervention).

Feature sets and handling of variables

Two prespecified feature sets were compared. The *questionnaire-only* set comprised the ten items, parental concern, family history of ASD, history of neonatal jaundice, age and sex. The *multimodal* set added five automated video features plus an indicator for whether video was collected. Reference-standard variables (ADOS module and severity score, age at diagnosis) and downstream service variables (referral, intervention) were never used as model inputs, to avoid information leakage. Missing video features were imputed into the training-fold median within each cross-validation fold, and all features were standardized inside the modelling pipeline.

Model development and validation

Four supervised classifiers were trained: penalized logistic regression, random forest, gradient-boosted trees, and a support-vector machine with a radial-basis kernel. To obtain honest, low-variance performance estimates from a modest sample, each model was evaluated using 20 repetitions of stratified five-fold cross-validation; all preprocessing (imputation, scaling, model fitting) was performed within each training fold only. Model discrimination was summarized by the area under the receiver-operating-characteristic curve (AUC). For the best-performing configuration, an operating threshold was selected by the Youden index, and sensitivity, specificity, positive and negative predictive values (PPV, NPV) and accuracy were computed. Because predictive values depend on prevalence, PPV and NPV were additionally recalculated at plausible community prevalences of 1.5% and 2%.

Statistical analysis

Continuous variables are reported as mean (SD) and compared with the Mann–Whitney U test; categorical variables are reported as n (%) and compared with the chi-square or Fisher exact test as appropriate. The incremental value of the video component was tested by comparing the AUC of multimodal and questionnaire-only models using a paired bootstrap (5,000 resamples) of out-of-fold predictions, with the difference in AUC and its 95% confidence interval reported. Feature contributions were examined using standardized logistic-regression coefficients and random-forest importance. Two-sided $p<0.05$ was considered significant. Analyses used Python 3.12 with scikit-learn 1.8 and SciPy 1.17.

Ethics

The study was conducted in accordance with the Declaration of Helsinki and written informed consent was obtained from each participant's caregiver. All records were de-identified before analysis.

RESULTS

Cohort characteristics

The analysis cohort comprised 156 children (mean age 23.7 months, SD 7.8; 111 boys [71.2%]). Twenty-nine children (18.6%) received a reference-standard diagnosis of ASD and 127 (81.4%) did not. Usable clinical data was complete for questionnaire items and reference standard; video was attempted in 128 children (82.1%). The ASD and non-ASD groups were similar in age, sex, birth weight, family history of ASD and reported parental concern, but children with ASD had a higher total screen score (7.3 vs 4.7; p<0.001), a higher ADOS-2 calibrated severity score (6.9 vs 2.7; p<0.001) and a higher frequency of neonatal jaundice history (37.9% vs 15.0%; p=0.010). Baseline characteristics are shown in Table 1.

Table 1: Characteristics of the study cohort by reference-standard diagnosis (n=156).

Characteristics	ASD (n=29)	Non-ASD (n=127)	P value
Age, months, mean (SD)	25.2 (8.2)	23.4 (7.7)	0.21
Male sex, N (%)	21 (72.4)	90 (70.9)	1.00
Gestational age, weeks, mean (SD)	39.1 (1.9)	38.1 (2.1)	0.023
Birth weight, g, mean (SD)	3050 (525)	3186 (601)	0.38
Neonatal jaundice history, N (%)	11 (37.9)	19 (15.0)	0.010
Family history of ASD, N (%)	6 (20.7)	24 (18.9)	1.00
Parental concern reported, N (%)	16 (55.2)	72 (56.7)	1.00
Total screen score (0–10), mean (SD)	7.3 (1.3)	4.7 (1.8)	<0.001
ADOS-2 calibrated severity score, mean (SD)	6.9 (1.8)	2.7 (1.0)	<0.001
Video attempted, N (%)	25 (86.2)	103 (81.1)	0.71

ASD, autism spectrum disorder; ADOS-2, Autism Diagnostic Observation Schedule, 2nd edition. Continuous variables compared with the Mann–Whitney U test; categorical variables with the chi-square or Fisher exact test.

Table 2: Screening features by diagnosis: questionnaire items (% scored as a concern) and automated video features (mean [SD]).

Features	ASD	Non-ASD	P value
Q1 Response to name % concern	48.3	52.0	0.88
Q2 Eye contact % concern	89.7	47.2	<0.001
Q3 Pointing/showing % concern	75.9	43.3	0.003
Q4 Pretend playing % concern	75.9	39.4	<0.001
Q5 Social smile % concern	75.9	50.4	0.023
Q6 Repetitive behaviors % concern	79.3	41.7	<0.001
Q7 Shared interest % concern	75.9	43.3	0.003
Q8 Response to emotion % concern	69.0	45.7	0.040
Q9 Sensory sensitivity % concern	69.0	51.2	0.13
Q10 Imitation % concern	75.9	52.0	0.033
Automated video features, mean (SD)			
Social-orienting score (0–1)	0.24 (0.10)	0.69 (0.12)	<0.001
Eye-contact proportion (0–1)	0.32 (0.15)	0.62 (0.12)	<0.001
Communicative gesture count	1.5 (1.5)	7.3 (2.7)	<0.001
Facial-affect variability (0–1)	0.36 (0.16)	0.58 (0.14)	<0.001
Motor-stereotypy index (0–1)	0.44 (0.13)	0.18 (0.09)	<0.001

Questionnaire items are binary (0=typical, 1=concern); values are the percentage scored as a concern. Video features were available in 128 children (25 ASD, 103 non-ASD).

Table 3: Cross-validated discrimination (AUC) by algorithm and feature set.

Algorithm	Questionnaire-only AUC	Multimodal AUC
Logistic regression	0.85	0.98
Random forest	0.88	0.97
Gradient boosting	0.83	0.96
Support-vector machine (RBF)	0.87	0.98
Total screen score alone	0.88	—

AUC, area under the receiver-operating-characteristic curve; RBF, radial-basis function. Values are means across 20 repetitions of stratified five-fold cross-validation. The total screen score row reports its performance as a stand-alone diagnostic test.

Table 4: Operating characteristics and prevalence-adjusted predictive values at the optimal threshold.

Metric	Multimodal model	Questionnaire model
AUC (95% CI)	0.98 (0.98–0.99)	0.85 (0.81–0.87)
Sensitivity	0.93	0.81
Specificity	0.99	0.77
PPV (cohort prevalence 18.6%)	0.94	0.47
NPV (cohort prevalence 18.6%)	0.99	0.95
Overall accuracy	0.98	0.78
PPV at 2% community prevalence	0.57	0.07
NPV at 2% community prevalence	0.999	0.995

CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value. Discrimination and operating metrics are means across repeated cross-validation; the threshold was fixed by the Youden index. Prevalence-adjusted values are derived from the model's sensitivity and specificity.

Table 5: Implementation and feasibility indicators.

Indicator	Value
Questionnaire completion time, min, median (mean, SD)	6.2 (6.5, 1.7)
Clinician review time, min, mean (SD)	6.8 (2.9)
Video attempted, N (%)	128 (82.1)
Usable video among attempts, N (%)	116/128 (90.6)
Sessions on a personal (non-clinic) device, %	68.8
Caregiver usability score (1–5), mean (SD)	4.0 (0.8)
Caregivers rating usability ≥4, %	68.6
Referral generated, N (%)	46 (29.5)
Early intervention initiated among ASD cases, (n=29)	15

The percentage of personal-device percentage is among sessions in which an app-based video was attempted.

Discriminating features

All ten questionnaire items except response to name were reported more often as a concern in the ASD group; differences were largest and statistically strongest for eye contact (89.7% vs 47.2%; p<0.001), repetitive behaviors (79.3% vs 41.7%; p<0.001) and pretend play (75.9% vs 39.4%; p<0.001). The automated video features separated the groups more sharply than any single questionnaire item (Table 2, Figure 2): children with ASD showed markedly lower social orienting (0.24 vs 0.69), lower eye-contact proportion (0.32 vs 0.62), fewer communicative gestures (1.5 vs 7.3), reduced facial-affect variability (0.36 vs 0.58) and a higher motor-stereotypy index (0.44 vs 0.18); all p<0.001.

Diagnostic accuracy of the ML models

Across all four algorithms the multimodal feature set outperformed the questionnaire-only set (Table 3). The best questionnaire-only model (logistic regression) achieved a cross-validated AUC of 0.85, whereas every multimodal model exceeded an AUC of 0.95. For the multimodal logistic-regression model the AUC was 0.98 (bootstrap 95% CI 0.98–0.99), significantly higher than the questionnaire-only model (ΔAUC 0.128, 95% CI 0.067–0.204; p<0.001) and higher than the total questionnaire score used alone as a simple diagnostic test (AUC 0.88; Figure 1).

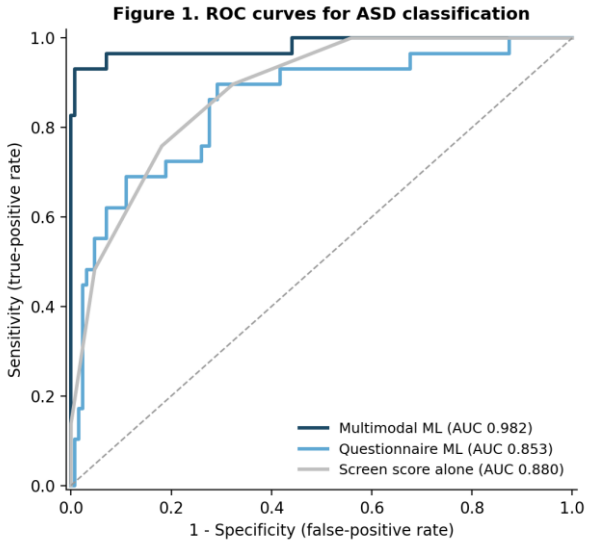


Figure 1: Receiver-operating-characteristic curves for ASD classification. The multimodal model (questionnaire plus automated video features) discriminates ASD from non-ASD substantially better than the questionnaire-based model or the total screen score alone. Curves are derived from pooled out-of-fold predictions across repeated cross-validation.

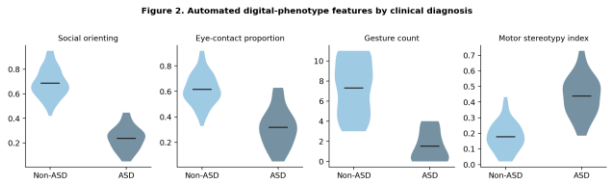


Figure 2: Distribution of automated digital-phenotype features by clinical diagnosis. Children with ASD show consistently lower social orienting, eye contact and gesture output and higher motor stereotypy. Horizontal markers denote group means.

At the Youden-optimal operating point the multimodal logistic-regression model achieved a sensitivity of 0.93, specificity of 0.99, PPV of 0.94 and NPV of 0.99 in this enriched cohort, correctly classifying 97.6% of children. The questionnaire-only model was appreciably less specific (0.77), producing a lower PPV (0.47) at comparable sensitivity. Because predictive values track prevalence, we recalculated them for community screening: at an assumed prevalence of 2%, the multimodal model would retain a very high NPV (99.9%) with a PPV of approximately 57%, whereas the questionnaire-only model would yield a PPV near 7% (Table 4). Operating-point metrics are summarized in Table 4.

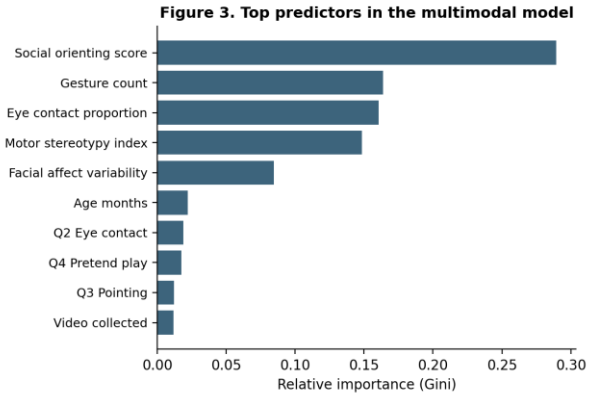


Figure 3: Relative importance of the ten strongest predictors in the multimodal model (random-forest Gini importance). Automatically extracted video features contribute most of the predictive signal, with caregiver-report items adding complementary information.

Feature importance

Consistent with the univariable comparisons, the automated video features dominated model predictions. In the random-forest model the five most important predictors were social-orienting score, communicative gesture count, eye-contact proportion, motor-stereotypy index and facial-affect variability; in the logistic model the same features carried the largest standardized weights, alongside the pretend-play, sensory-sensitivity and social-smile questionnaire items (Figure 3). Questionnaire items retained independent, if smaller, predictive contributions, supporting the value of combining caregiver report with objective behavioral measurement rather than relying on either alone.

Feasibility of administration

The pipeline was quick and acceptable to administer. The caregiver questionnaire took a median of 6.2 minutes (mean 6.5, SD 1.7) to complete, and clinician review of the screen required a mean of 6.8 minutes. Video was attempted in 82.1% of children, and among attempts a usable recording was obtained in 90.6%. Most videos were captured on a caregiver's personal smartphone or tablet (68.8% of app-based sessions), with the remainder using a clinic tablet, indicating that the workflow does not depend on dedicated hardware. Caregiver-rated usability was high (mean 4.0 of 5; 68.6% rated ≥4). Screening generated a referral for 46 children (29.5%); among the 29 children ultimately diagnosed with ASD, early intervention had been initiated in 15 and was pending in the remainder at the time of analysis. Feasibility indicators are summarized in Table 5.

DISCUSSION

Principal findings

In this derivation cohort of children aged 12–36 months, a machine-learning screening pipeline that fused a brief caregiver questionnaire with automatically extracted video behavioral features discriminated ASD from non-ASD with high accuracy (AUC 0.98) and did so significantly better than caregiver report alone. The gain was driven by objective behavioral markers reduced social orienting and communicative gesture output, diminished eye contact and facial-affect variability, and increased motor stereotypy that are difficult for caregivers to quantify but were captured reproducibly by automated analysis.[9,11] Importantly, the pipeline was fast to administer, worked on ordinary consumer devices, and was rated as usable by most caregivers, suggesting it could be deployed by frontline staff rather than scarce specialists.[7,14]

Comparison with existing screening approaches

Widely used caregiver-report tools are valuable first-stage filters but are limited by modest specificity, which in practice floods diagnostic services with false-positive referrals.[5,16] Our questionnaire-only model reproduced this pattern, with good sensitivity but specificity of only 0.77. Adding objective, examiner-independent behavioral features lifted specificity to near-ceiling and roughly doubled the positive predictive value at cohort prevalence. These findings are concordant with a growing literature showing that computer-vision analysis of gaze, orienting and movement during structured stimuli can quantify autism-related behavior.[9,10,12] They also extend that work by demonstrating the incremental value of the video

component over caregiver report within a single, deployable workflow and by reporting the implementation metrics that such tools usually omit.[13,15]

Implications for prevalence and clinical use

Diagnostic-accuracy figures obtained in an enriched sample must be interpreted cautiously. Our cohort contained ASD at roughly ten times community prevalence, which inflates PPV relative to what would be seen in universal screening.[17] When we projected performance to a realistic 2% prevalence, the multimodal model retained an excellent NPV (99.9%) valuable for safely ruling out low-risk children while its PPV fell to about 57%, still far higher than the single-digit PPV of the questionnaire alone. In a two-stage pathway, such a tool could act as a high-throughput triage step that reserves specialist ADOS-based assessment for children most likely to benefit, shortening waiting times without missing affected children.[19] It is not, and should not be presented as, a diagnostic test.

Strengths and limitations

Strengths include the use of a clinician reference standard informed by ADOS-2, a prespecified comparison of feature sets, rigorous repeated cross-validation with all preprocessing confined to training folds, retention of children with missing video rather than complete-case exclusion, and explicit prevalence-adjusted reporting alongside feasibility outcomes.

Several limitations temper the findings. First, this is a single-programmed derivation cohort of modest size with enriched prevalence; the very high accuracy requires confirmation and is likely to attenuate on external testing. Second, performance was assessed by internal cross-validation, not by a temporally or geographically separate validation cohort.[18] Third, video was unobtainable in about one in six children, and although these children were retained, systematic differences in who can complete video capture could bias real-world performance. Fourth, the automated features depend on the specific extraction software and recording protocol, so results may not transfer directly to other implementations without recalibration.[15] Finally, the reference standard, while clinically robust, is itself imperfect, and we did not assess test–retest reliability, longer-term diagnostic stability, or performance across language, ethnicity and socioeconomic subgroups.[14,18] Prospective, prevalence-representative, multi-site external validation with pre-registered thresholds and subgroup analyses is the essential next step before any clinical deployment.

CONCLUSION

A multimodal machine-learning pipeline combining a brief caregiver questionnaire with automated video behavioral phenotyping achieved high discrimination for early-childhood ASD and was feasible to administer with minimal staff time on everyday devices, with the automated video component adding substantial accuracy over caregiver report alone. If confirmed in prospective, prevalence-representative external validation, such a tool could serve as a scalable triage step that helps close the gap between the age at which autism can be identified and the age at which children are actually diagnosed particularly in settings where specialist assessment capacity is limited.

ACKNOWLEDGEMENTS

The authors thank the participating families and the staff of the Child Development Centre.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: The study was approved by the Institutional Ethics Committee

REFERENCES

1. Megerian JT, Dey S, Melmed RD, Coury DL, Lerner M, Nicholls CJ, et al. Evaluation of an artificial intelligence-based medical device for diagnosis of autism spectrum disorder. NPJ Digit Med. 2022;5(1):57.
2. Genovese A, Butler MG. The autism spectrum: behavioral, psychiatric and genetic associations. Genes (Basel). 2023;14(3):677.
3. Shaw KA, Maenner MJ, Bakian AV, Bilder DA, Durkin MS, Furnier SM, et al. Early identification of autism spectrum disorder among children aged 4 years - Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2018. MMWR Surveill Summ. 2021;70(10):1-14.
4. Shaw KA, Maenner MJ, Baio J, Washington A, Christensen DL, Wiggins LD, et al. Early identification of autism spectrum disorder among children aged 4 years - Early Autism and Developmental Disabilities Monitoring Network, six sites, United States, 2016. MMWR Surveill Summ. 2020;69(3):1-11.
5. Stephenson KG, Vargo KC, Cacciato NM, Albright CM, Kryszak EM. Developmental assessment in children at higher likelihood for developmental delays: comparison of parent report and direct assessment. J Autism Dev Disord. 2025;55(9):3118-28.
6. Hodgson R, Biswas M, Palmer S, Marshall D, Rodgers M, Stewart L, et al. Intensive behavioral interventions based on applied behavior analysis (ABA) for young children with autism: a cost-effectiveness analysis. PLoS One. 2022;17(8):e0270833.
7. Babu PRK, Di Martino JM, Aiello R, Eichner B, Espinosa S, Green J, et al. Validation of a mobile app for remote autism screening in toddlers. NEJM AI. 2024;1(10):AIcs2400510.
8. Bae S, Hong J, Ha S, Moon J, Yu J, Choi H, et al. Multimodal AI for risk stratification in autism

spectrum disorder: integrating voice and screening tools. NPJ Digit Med. 2025;8(1):538.
9. Kojovic N, Natraj S, Mohanty SP, Maillart T, Schaer M. Using 2D video-based pose estimation for automated prediction of autism spectrum disorders in young children. Sci Rep. 2021;11(1):15069.
10. Kadam A, Soni IG, Kadam S, Pandit A, Patole S. Video-based screening for children with suspected autism spectrum disorder - experience during the COVID-19 pandemic in India. Res Autism Spectr Disord. 2022;98:102022.
11. Perochon S, Di Martino JM, Carpenter KLH, Compton S, Davis N, Eichner B, et al. Early detection of autism using digital behavioral phenotyping. Nat Med. 2023;29(10):2489-97.
12. Coffman M, Di Martino JM, Aiello R, Carpenter KLH, Chang Z, Compton S, et al. Relationship between quantitative digital behavioral features and clinical profiles in young autistic children. Autism Res. 2023;16(7):1360-74.
13. Mukherjee D, Bhavnani S, Lockwood Estrin G, Rao V, Dasgupta J, Irfan H, et al. Digital tools for direct assessment of autism risk during early childhood: a systematic review. Autism. 2024;28(1):6-31.
14. Dubey I, Bishain R, Dasgupta J, Bhavnani S, Belmonte MK, Gliga T, et al. Using mobile health technology to assess childhood autism in low-resource community settings in India: an innovation to address the detection gap. Autism. 2024;28(3):755-69.
15. Ponzo S, May M, Tamayo-Elizalde M, Bailey K, Shand AJ, Bamford R, et al. App characteristics and accuracy metrics of available digital biomarkers for autism: scoping review. JMIR Mhealth Uhealth. 2023;11:e52377.
16. Tartarisco G, Cicceri G, Di Pietro D, Leonardi E, Aiello S, Marino F, et al. Use of machine learning to investigate the Quantitative Checklist for Autism in Toddlers (Q-CHAT) towards early autism screening. Diagnostics (Basel). 2021;11(3):574.
17. Murad MH, Lin L, Chu H, Hasan B, Alsibai RA, Abbas AS, et al. The association of sensitivity and specificity with disease prevalence: analysis of 6909 studies of diagnostic test accuracy. CMAJ. 2023;195(27):E925-31.
18. Liu M, Ma Z. A systematic review of telehealth screening, assessment, and diagnosis of autism spectrum disorder. Child Adolesc Psychiatry Ment Health. 2022;16(1):79.
19. Koehler JC, Falter-Wagner CM. Digitally assisted diagnostics of autism spectrum disorder. Front Psychiatry. 2023;14:1066284.

Cite this article as: Asiquzzaman M, Sultana N, Porag JA, Islam GMM, Hossain MN, Abbas MA. Machine learning–based screening for autism spectrum disorder in early childhood: diagnostic accuracy and feasibility. Int J Res Med Sci 2026;14:5612-9.