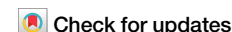




Automated AI based identification of autism spectrum disorder from home videos



Dong Yeong Kim^{1,2,24}, Ryemi Do^{3,24}, Youmin Shin^{1,2}, Hewoen Sim³, Hanna Kim³, Sungchul Cho³, Geonhee Lee^{2,4}, Seyeon Park^{2,5}, Boa Jang^{1,2}, Hyojeong Lim³, Sungji Ha⁶, Jaeun Yu⁶, Hangnyoung Choi^{6,7}, Junghan Lee^{6,7}, Min-Hyeon Park⁸, Ayeong Cho⁸, Chan-Mo Yang^{9,10}, Dongho Lee⁹, Heejeong Yoo^{11,12}, Yoojeong Lee¹¹, Guiyoung Bong¹¹, Johanna Inhyang Kim¹³, Haneul Sung¹⁴, Hyo-Won Kim¹⁵, Eunji Jung¹⁶, Seungwon Chung¹⁷, Jung-Woo Son¹⁷, Jae Hyun Yoo¹⁸, Sekye Jeon¹⁸, Jinseong Jang¹⁹, You Bin Lim²⁰, Jeeyoung Chun²⁰, Wooseok Choi²⁰, Sooyeon Lee²⁰, Sohyun Park²⁰, Jisung Ahn²⁰, Chae Rim Lee²⁰, Keun-Ah Cheon^{7,21,25} ✉, Young-Gon Kim^{2,22,23,25} ✉ & Bung-Nyun Kim^{20,25} ✉

Autism spectrum disorder (ASD) is a prevalent childhood-onset neurodevelopmental condition. Early diagnosis remains challenging by the time, cost, and expertise required for traditional assessments, creating barriers to timely identification. We developed an AI-based screening system leveraging home-recorded videos to improve early ASD detection. Three task-based video protocols under 1 min each—name-response, imitation, and ball-playing—were developed, and home videos following these protocols were collected from 510 children (253 ASD, 257 typically developing), aged 18–48 months, across 9 hospitals in South Korea. Task-specific features were extracted using deep learning models and combined with demographic data through machine learning classifiers. The ensemble model achieved an area under the receiver operating characteristic curve of 0.83 and an accuracy of 0.75. This fully automated approach, based on short home-video protocols that elicit children’s natural behaviors, complements clinical evaluation and may aid in prioritizing referrals and enabling earlier intervention in resource-limited settings.

Autism spectrum disorder (ASD) is a neurodevelopmental condition characterized by differences in social communication and interaction, as well as patterns of restricted and repetitive behaviors that emerge early in development¹. Recent meta-analyses estimate that globally approximately 0.6% of the population is on the autism spectrum². Moreover, data from the Global Burden of Disease Study suggest that ASD affected around 61.8 million individuals worldwide in 2021—one in 127 people—placing ASD among the top causes of non-fatal health burden in children and adolescents under 20³. ASD influences cognitive and socio-emotional functioning across the lifespan, and early identification is critical for enhancing adaptive functioning and social outcomes. Initiation of personalized intervention between 2.5 and 3 years of age was associated with the greatest gains in cognitive functioning after 1 year, with younger age at onset significantly predicting improved outcomes⁴. However, ASD is typically diagnosed at an average age of 3.5–4 years worldwide⁵, which is considerably later than the ideal window for early intervention, generally regarded as before age 2. Such delays are even more pronounced in low- and middle-income countries,

where the average age of diagnosis is approximately 45.5 months, with some regions in Asia and Africa reporting mean diagnostic ages exceeding 5 years due to systemic barriers and limited access to specialized services⁵. According to the Centers for Disease Control and Prevention (CDC)⁶, most children with ASD in the U.S. are not diagnosed until approximately 54 months of age, with 70% being diagnosed after 51 months. In South Korea, although developmental screenings are conducted regularly between 4 months and 5 years of age, significant delays in diagnosis and intervention persist even when early parental concerns are present^{7,8}. At tertiary hospitals, the waiting time for diagnostic evaluation can extend from 1–2 years⁹. Given the global prevalence and the widespread delays in diagnosis, there is an urgent international need for scalable, automated screening tools that can support early identification and intervention.

Conventional diagnostic tools, such as the Autism Diagnostic Observation Schedule (ADOS)¹⁰ and the Autism Diagnostic Interview-Revised (ADI-R)¹¹, are resource-intensive, reliant on trained professionals, and may introduce observer bias¹². These standardized instruments, while considered

gold standards, require in-person administration, are time-consuming, and are often limited in accessibility due to high cost and the need for specialized training. Compared to the diagnostic tool of the ADOS-2, parent-report screening tools for ASD, such as the M-CHAT and Q-CHAT, demonstrate limited accuracy, with insufficient sensitivity and positive predictive value¹³. In addition, caregiver-report instruments such as the SRS-2 and SCQ-2 have been reported to show limited specificity in distinguishing ASD from other developmental or psychiatric conditions, indicating their potential risk for over-identification when used without clinician-administered assessments¹⁴. The reduced accuracy of caregiver-report screening tools may stem from variability in caregivers' recall and subjective interpretation of behaviors, which can influence item responses and compromise diagnostic precision. Conversely, although clinician-administered tools like the ADOS and ADI-R offer higher diagnostic validity, they may not fully capture behaviors that manifest in naturalistic home or community settings, as children's behavior can differ across contexts and over time.

In contrast, home videos offer high ecological validity by capturing children's behaviors in familiar, everyday settings¹⁵. Children spend most of their time at home, where they are generally more relaxed and likely to display their typical behaviors. In contrast, clinical or laboratory settings may evoke atypical behaviors due to their unfamiliarity. For instance, toddlers on the autism spectrum have been shown to exhibit more repetitive behaviors in clinical environments compared to the home¹⁶. Observing children during spontaneous interactions in their natural environment allows for a more representative and context-sensitive assessment of their developmental functioning¹⁷, which also aligns with the neurodiversity paradigm^{18,19}. Nevertheless, the manual coding of home videos is labor-intensive and prone to inter-rater variability¹⁵. This reduces the scalability and reliability of video-based assessment.

Recent research has focused on artificial intelligence (AI) and machine learning (ML) for the automated analysis of home videos, which offer scalable and objective alternatives²⁰. While promising, most AI studies face limitations, such as small sample sizes^{20–22} or relying on integrating questionnaires with home videos²³ and manually annotating them^{24–28}, which introduces subjectivity and limits generalizability. Some studies have adopted automated feature extraction methods, but they typically focus only on specific features such as stimming behaviors²⁹ or facial analysis^{20,30} and often require unfamiliar environments, such as controlled laboratory settings^{21,28}. Many approaches depend on specific behavioral categories or constrained protocols, which may fail to capture the variability and complexity of naturalistic behaviors. These methodological constraints reduce the scalability, objectivity, and ecological validity of automated screening systems, thereby limiting their utility in real-world, early ASD screening contexts.

To overcome these limitations, we developed short, structured home-video protocols that parents can record in familiar settings to naturally elicit each child's unique ASD-related behaviors. In contrast to methods based on manual video coding, our fully automated AI pipelines objectively extract clinically meaningful behavioral indicators from these videos. By combining parent-friendly, naturalistic behavior elicitation with objective AI-based feature extraction, our method addresses the objectivity, scalability, and ecological validity gaps in previous research, offering a practical solution for earlier and more accessible ASD screening.

Results

Dataset demographic

In Table 1, we present the total number of young children in each class along with the corresponding (training/testing) allocation for clarity and to ensure a consistent model evaluation across videos. Notably, a male predominance was observed, with male participants accounting for more than twice the number of female participants³¹. This imbalance is consistent with the higher prevalence of ASD in males, reflecting the composition of the dataset.

In addition, the number of participants varied across the videos. Ninety children were included in the test dataset. Ten children recorded two videos, two children recorded all three videos, and the remaining 78 children recorded only one video. For the final ensemble model, we integrated the

predictions for each child by averaging their model-predicted confidence scores across one or more videos to ensure the comprehensive integration of predictions across scenarios.

To evaluate whether the training and test sets were comparable in terms of demographic and behavioral variables, we conducted independent two-sample *t*-tests across each feature. This analysis was performed to confirm that any observed performance differences in model evaluation would not be attributable to confounding population disparities. As shown in Supplementary Table 1, all *p* values exceeded 0.05, indicating no statistically significant differences between the training and test sets.

Task-specific classification performance

Table 2 summarizes the classification results across name-response, imitation, and ball-playing tasks, reporting the area under the receiver operating characteristic curve (AUROC), accuracy (ACC), precision (PRE), and sensitivity (SEN) for each model. Each model was evaluated with stepwise inclusion of task-specific features, common clinical features, and demographic metadata (age and sex) to assess the incremental value of feature integration.

- For the name-response task, which targets social orienting behaviors characteristic of ASD, LightGBM³² achieved an AUROC of 0.72 without additional features. Incorporating metadata improved performance to an AUROC of 0.81, with accuracy increasing from 0.69 to 0.73.
- For the imitation task, designed to assess differences in social imitation, the logistic regression model³³ improved from an AUROC of 0.65 (baseline) to 0.75 with common features, and further to 0.78 with both common features and metadata.
- For the ball-playing task, measuring reciprocal turn-taking, LightGBM³² improved from an AUROC of 0.62 to 0.78 with common features, and ultimately to 0.81 with full feature integration.

These findings demonstrate that incorporating multi-domain social behavioral features enhances classification performance, reflecting the multi-faceted nature of ASD symptomatology.

Ensemble model performance

The ensemble model, integrating predictions across videos from multiple tasks, achieved an AUROC of 0.80 at baseline, increasing to 0.83 with metadata inclusion. This ensemble approach provided the most robust and generalizable classification performance, underscoring the benefit of aggregating diverse behavioral dimensions. External validation on noisy video samples achieved an AUROC of 0.73, supporting the feasibility of applying the model under variable home-recording conditions. Detailed performance metrics are provided in Supplementary Table 2.

Interpretation of model predictions

SHapley Additive exPlanations (SHAP)³⁴ analysis was conducted to identify key feature contributions aligned with clinically recognized ASD behaviors:

- Name-response task (Fig. 1b): longer response latency and elevated variability in parental calling attempts were strongly associated with ASD predictions, reflecting the ability to orient to social stimuli.
- Imitation task (Fig. 1d): reduced eye contact duration, diminished physical engagement, and delayed imitation responses were key drivers of ASD classification, consistent with behaviors in motor imitation and joint attention.
- Ball-playing task (Fig. 1f): prolonged turn-taking durations and reduced eye contact contributed to ASD predictions, reflecting ability to engage in reciprocal social engagement and coordination.

Behavioral signature of ASD in extracted features

Significant group-level differences were observed between ASD and TD children in both task-specific and common clinical features (Tables 3 and 4).

- For task-specific features (Table 3): children with ASD demonstrated significantly longer response latencies during name-response (5.29 ± 5.66 vs. 3.62 ± 3.25 s; $p = 0.017$). Although similar trends were

Table 1 | Demographics of the final dataset used for AI analysis

Videos	Training data/Test data number			Age (month)	Sex (M/F)
	TD	ASD	Total		
Name-response	89 (71/18)	88 (70/18)	177 (141/36)	37.3 ± 8.0	124/53
Imitation	84 (67/17)	82 (65/17)	166 (132/34)	39.4 ± 6.9	117/49
Ball-playing	84 (67/17)	83 (66/17)	167 (133/34)	39.2 ± 6.9	116/51
Total	257 (205/52)	253 (201/52)	510 (406/104)	38.6 ± 7.3	357/153

Age was recorded as the mean and standard deviation in months.
 TD typically developing, ASD autism spectrum disorder, AI artificial intelligence.

Table 2 | Machine-learning model results

	Meta info (sex, age)	Common clinical features	Unique clinical features	AUROC	ACC	PRE	SEN
Name-response			✓	0.72 ± 0.04	0.69 ± 0.04	0.67 ± 0.05	0.77 ± 0.08
Name-response	✓		✓	0.81 ± 0.02	0.73 ± 0.03	0.72 ± 0.03	0.77 ± 0.05
Imitation			✓	0.65 ± 0.00	0.65 ± 0.02	0.64 ± 0.02	0.69 ± 0.02
Imitation		✓	✓	0.75 ± 0.00	0.69 ± 0.06	0.71 ± 0.08	0.69 ± 0.02
Imitation	✓	✓	✓	0.78 ± 0.01	0.74 ± 0.02	0.73 ± 0.02	0.78 ± 0.04
Ball-playing			✓	0.62 ± 0.02	0.55 ± 0.02	0.54 ± 0.02	0.76 ± 0.02
Ball-playing		✓	✓	0.78 ± 0.03	0.69 ± 0.05	0.69 ± 0.09	0.72 ± 0.06
Ball-playing	✓	✓	✓	0.81 ± 0.03	0.75 ± 0.03	0.72 ± 0.03	0.84 ± 0.04
Ensemble		✓	✓	0.80 ± 0.02	0.74 ± 0.02	0.76 ± 0.04	0.77 ± 0.05
Ensemble	✓	✓	✓	0.83 ± 0.01	0.75 ± 0.02	0.76 ± 0.01	0.80 ± 0.02

The mean and standard deviation of each value were calculated.
 ACC accuracy, AUROC area under the receiver operating characteristic curve, PRE precision, SEN sensitivity.

observed in imitation and ball-playing tasks, these differences did not reach statistical significance ($p = 0.064$ and $p = 0.116$, respectively).

- For common clinical features (Table 4): children with ASD exhibited greater lack of eye contact (4.25 ± 6.47 vs. 1.30 ± 3.78 s; $p < 0.001$), increased non-engaged movements (2.59 ± 9.23 vs. 0.69 ± 5.94 s; $p < 0.001$), and prolonged physical contact duration (3.78 ± 7.69 vs. 0.69 ± 2.52 s; $p = 0.026$).

These results statistically support the discriminative utility of both task-specific and common behavioral features, reflecting a coherent profile of delayed and disrupted social engagement in ASD.

Clinical evaluation of misclassified cases

To further examine model behavior, clinical experts reviewed misclassified test videos (104 participants: 52 ASD, 52 TD), categorized into true positive (TP), true negative (TN), false positive (FP), and false negative (FN) groups. TD children (TN and FP) consistently outperformed ASD children (TP and FN) across psychological assessments. Notably, within the ASD group, FN cases exhibited milder symptom profiles compared to TP, particularly on CBCL³⁵ domains assessing withdrawal and internalizing problems. These observations suggest that the model may exhibit sensitivity to subthreshold behavioral phenotypes and may capture a broader risk spectrum, supporting its potential utility for early identification of at-risk children even in borderline cases (Supplementary Note 1 and Supplementary Tables 3–5).

Discussion

This study presents a fully automated AI model for the early identification of ASD using short home videos based on a large cohort of young children, without relying on manual coding or parent-report measures. We developed structured videos protocols, each under 1 min, recorded by parents in familiar settings to elicit core social behaviors relevant for ASD screening. The research team predefined key features—such as response latency, parental attempts, sequential turn-taking, and gaze—which were extracted from three videos tasks (name-response, imitation, and ball-playing) using

deep learning, and then used in machine learning classifiers to build an AI model for ASD classification. Feature integration across these tasks resulted in robust diagnostic performance (AUROC = 0.83 for the ensemble model). Critically, the extracted features were not only discriminative but aligned with core clinical constructs of ASD, including reduced social orienting, diminished eye contact, and delayed imitation. SHAP-based feature attribution analyses confirmed that response latency and differences in gaze behavior consistently emerged as key discriminators across tasks, reinforcing their clinical relevance. Beyond discriminative performance, our approach demonstrated strong practical feasibility for real-world deployment, with an average inference time of approximately 14.2 s per video on standard GPU-equipped systems (RTX 3090 Ti, 24 GB VRAM). The pipeline relies entirely on open-source models—including COCO-based pose estimation, YOLOv8 for object detection, and Whisper for speech-to-text—enabling rapid, cost-free, and license-independent ASD risk estimation. In contrast to traditional diagnostic pathways such as ADOS or ADI-R, which require hours of expert-administered testing in clinical settings, our model offers fully automated ASD risk estimation in approximately 14 s per video, substantially improving accessibility and scalability.

Compared to prior research, our study offers several distinct methodological improvements that can provide an objective, low-cost, and ecologically valid approach for early ASD risk detection. Unlike earlier studies, which frequently depended on subjective assessments such as parent-reported questionnaires^{23,26,27} or manual annotation of videos^{24,25,27,28}, our approach implements a fully automated pipeline for feature extraction, significantly reducing human biases and inter-rater variability while remaining interpretable and clinically grounded. Furthermore, whereas recent automated methods predominantly target specific body features or rely on controlled laboratory settings^{20,21,28–30}, our method utilizes deep learning to comprehensively analyze rich, full-body behavioral indicators captured in naturalistic home settings. By integrating multiple tasks, our model successfully captures ASD-related behaviors, thereby substantially enhancing ecological validity and accessibility. These innovations enable

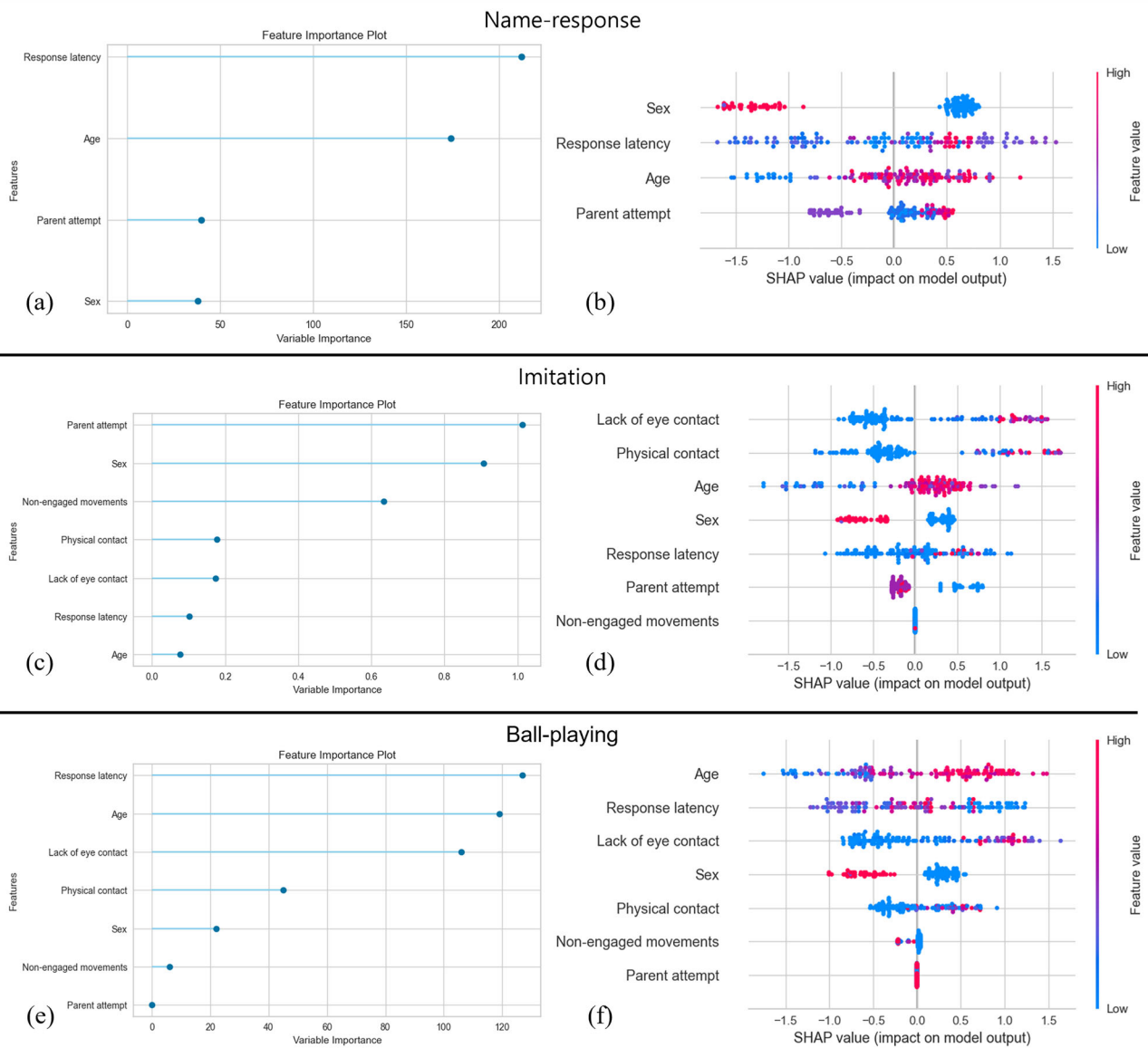


Fig. 1 | Feature importance and SHAP explainability for each model. a, c, e illustrate the relative feature importance, representing each feature’s contribution to the model’s overall predictions. b, d, f display SHAP values, providing detailed insights into how individual features affect specific predictions. SHAP SHapley Additive exPlanations.

Table 3 | Analysis of extracted unique features

		Response latency (second)	p value	Parent attempts (number)	p value
Name-response	TD	3.62 ± 3.25	0.017	1.67 ± 1.36	0.101
	ASD	5.29 ± 5.66		2.08 ± 1.88	
Imitation	TD	2.90 ± 5.82	0.064	1.10 ± 1.43	0.216
	ASD	5.48 ± 11.09		0.87 ± 0.89	
Ball-playing	TD	9.79 ± 7.01	0.116	–	–
	ASD	13.42 ± 19.74			

The mean and standard deviation of each value were calculated. An independent t-test was conducted to compare the TD and ASD groups. TD typically developing, ASD autism spectrum disorder.

cost-effective, scalable deployment in diverse, real-world environments and advance the democratization of neurodevelopmental screening by extending interpretable, clinically grounded AI tools to settings with limited specialty resources.

Another key strength of this study was the use of a standardized video recording protocol and large-scale data collection. Our sample comprised 510 children aged 18–48 months (253 with ASD and 257 typically developing), systematically recruited from 9 hospitals and community sites across South Korea, providing a relatively diverse cohort that enhances the generalizability of our findings. Unlike studies using preexisting datasets with small sample sizes^{22,29}, broad age ranges, or age imbalances between groups^{24,36}, our protocol ensured consistency in data quality and demographics. Detailed video-recording instructions delivered via a mobile app further improved data uniformity. While some studies provide only general guidelines (e.g., keeping the child’s face visible, using toys, and including social interactions)^{24,25,36}, our study emphasizes the importance of structured, standardized instruction to reduce variability in home video environments¹⁵.

This study has several limitations that should be addressed in future research. First, the sample included only children with ASD and TD, with limited clinical diversity and demographic representation (e.g., predominantly male and under age four). This may restrict the generalizability of findings, as early-diagnosed children often show more pronounced

Table 4 | Analysis of extracted common features

	Lack of eye contact	<i>p</i> value	Non-engaged movements	<i>p</i> value	Physical contact	<i>p</i> value
TD	1.30 ± 3.78	<0.001	0.69 ± 5.94	<0.001	0.69 ± 2.52	0.026
ASD	4.25 ± 6.47		2.59 ± 9.23		3.78 ± 7.69	

The mean and standard deviation of each value were calculated. An independent *t*-test was conducted to compare the TD and ASD groups. TD typically developing, ASD autism spectrum disorder.

symptoms³⁷, and females with ASD—who may present differently—were underrepresented³¹. Future studies should aim to recruit more heterogeneous samples, including children with language delays, attention difficulties, or early anxiety symptoms, and ensure a better balance across gender and age groups. Second, while ASD diagnoses were based on standardized assessments such as the ADOS^{38,39}, the absence of clinician consensus may reduce diagnostic certainty. In addition, the TD group was not followed longitudinally, raising the possibility that some participants may later receive ASD diagnoses. Incorporating long-term follow-up for all groups would improve the reliability and clinical applicability of future models. Third, in terms of data collection, although standardized instructions were given for video collection, uncontrolled variables in home environments may have introduced variability. More standardized or semi-structured recording environments should be considered in future studies to reduce noise and improve reliability. Moreover, not all children had all three video tasks available. As a result, the ensemble model utilized between one and three videos per child, potentially affecting consistency. A more uniform data collection protocol ensuring complete multimodal input per subject would strengthen comparative analyses. Fourth, regarding AI analysis, several technical and performance-related limitations were noted. Comparing psychological assessment outcomes with AI model predictions revealed that children correctly classified as having ASD by the model (true positives) exhibited more severe symptoms than those misclassified as TD (false negatives). This suggests the model may currently be optimized for identifying high-risk cases but is less sensitive to subtler or borderline ASD presentations. Incorporating data across the full spectrum of ASD features may improve the accuracy and generalizability of future models. In addition, several task-specific limitations were observed. In name-response videos, the STT model showed imprecise response timing due to variation in caregiver speech. In imitation tasks, keypoint-detection errors reduced the reliability in detecting gestures. In ball-playing tasks, object detection was inconsistent due to ball variability. Manual review also revealed systematic overestimation of task duration and occasional misclassifications. Future studies should consider training domain-adapted STT models using caregiver-child interaction data. Enhancing pose estimation with child-specific gesture datasets, standardizing task materials, and implementing automated quality control for object recognition. Addressing these limitations in future research will be critical for advancing clinically applicable AI-based diagnostic tools for ASD.

In summary, this study demonstrates the feasibility of an automated, video-based AI model for early ASD screening using short home videos. By leveraging deep learning to extract clinically meaningful behaviors from three types of task videos, our machine learning models provide a scalable and accessible alternative to traditional assessments. Enhancing diagnostic validity and sample representativeness in future studies could increase the practical applicability of AI-driven video analysis as a promising tool to assist early identification of ASD in real-world settings, particularly where clinical resources are limited.

Methods

Ethics approval

The research protocol was approved by the Institutional Review Boards (IRB) of all participating hospitals, including the Seoul National University College of Medicine/Seoul National University Hospital (IRB No. 2209-096-1360), Severance Hospital, Yonsei University Health System (IRB No. 4-2022-1468), Bundang Seoul National University Hospital (IRB No. 2305-

829-401), Hanyang University Hospital (IRB No. 2022-12-007-001), Eunpyeong St. Mary's Hospital (IRB No. 2022-3419-0002), Asan Medical Center (IRB No. 2023-0114), Chungbuk National University Hospital (IRB No. 2023-04-034), Wonkwang University Hospital (IRB No. 2022-12-023-001), and Seoul St. Mary's Hospital (IRB No. KC24ENDI0198). Written informed consent was obtained from all parents and/or legal guardians of participating children.

Study design overview

The study followed a stepwise design described in Fig. 2: home videos were first screened through a selection process to ensure protocol compliance and quality. Screened videos were then processed through deep learning-based modules, depending on the task: STT (speech-to-text)⁴⁰ was applied to capture verbal responses in name-response videos, Key-point Detector (pose estimation)⁴¹ was used to track 17 body keypoints in all three videos, and Ball Detector (object detection)⁴² was employed to detect ball position in ball-playing videos. These sub-features were subsequently transformed into clinically meaningful behavioral metrics, developed collaboratively by AI and clinical experts and informed by prior ASD studies^{38,43}. Finally, the extracted behavioral features served as inputs to machine learning classifiers trained for ASD screening, and predictions were integrated through an ensemble method based on confidence scores.

The following sections describe the participant recruitment process, the video collection and selection protocol, clinical feature extraction, and machine learning classification in detail.

Participants and recruitment

We recruited children aged 18–48 months who visited the pediatric or psychiatric departments at nine tertiary care hospitals in South Korea between October 2022 and May 2024. The participating institutions included Seoul National University Hospital, Severance Hospital, Eunpyeong St. Mary's Hospital, Wonkwang University Hospital, Bundang Seoul National University Hospital, Hanyang University Hospital, Asan Medical Center, Chungbuk National University Hospital, and Seoul St. Mary's Hospital. Recruitment was conducted through outpatient clinics, community outreach, and online promotions.

Children were excluded if they met any of the following exclusion criteria: (1) <18 months or >49 months of age; (2) congenital genetic disorders; (3) history of acquired brain injury (e.g., cerebral palsy); or (4) seizure disorders or other neurological conditions. After applying these eligibility criteria, 315 children diagnosed with ASD and 127 children classified as typically developing (TD) were included.

All participants underwent psychological assessments, including developmental screenings and ASD-specific evaluations, tailored by age (Table 5). The screening tools included the Korean Developmental Screening Test for Infants and Children (K-DST)⁴⁴, Behavior Development Screening for Toddlers-Interview/play (BeDevel-I/P)^{7,45}, Modified Checklist for Autism in Toddlers (M-CHAT)⁴⁶, Quantitative Checklist for Autism in Toddlers (Q-CHAT)⁴⁷, Sequenced Language Scale for Infants (SELSI)⁴⁸, Child Behavior Checklist (CBCL)³⁵, Korean Vineland Adaptive Behavior Scales (K-VABS)⁴⁹, Social Communication Questionnaire Lifetime Version (SCQ-L)⁵⁰, Social Responsiveness Scale (SRS-2)⁵¹, and Preschool Receptive-Expressive Language Scale (PRES)⁵². If any screening result exceeded clinical thresholds, diagnostic evaluations were conducted using the Autism Diagnostic Observation Schedule, Second Edition (ADOS-2)³⁸ and the Korean Childhood Autism Rating Scale, Second Edition (K-CARS-2)⁵³. All ADOS-2

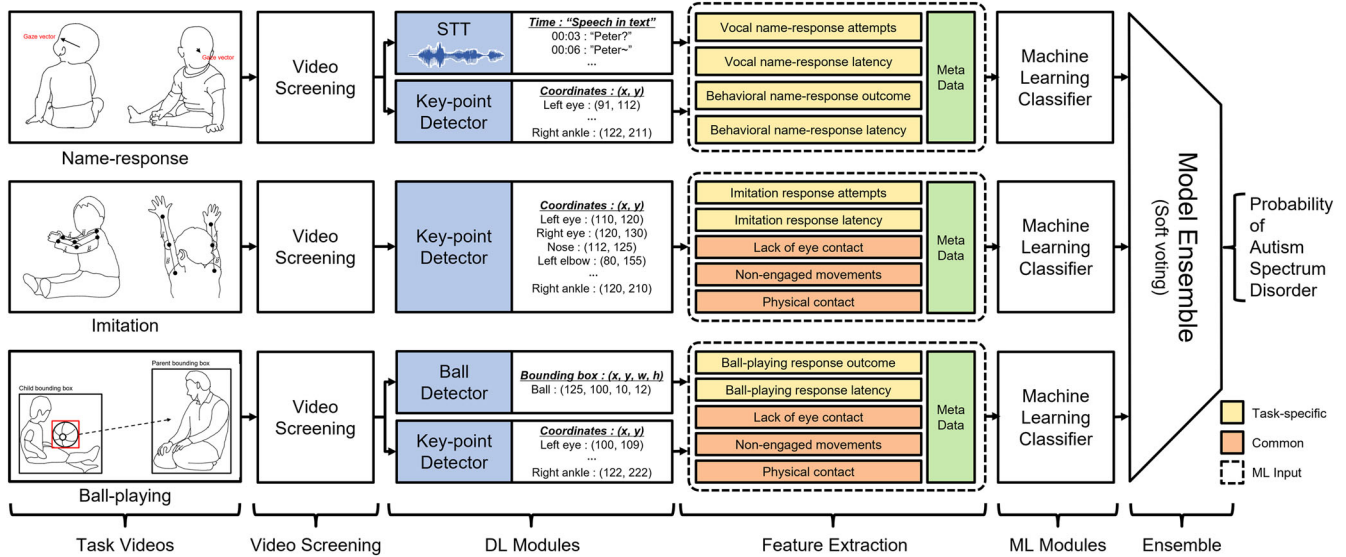


Fig. 2 | Overview of our AI approach. Each home video first undergoes a selection process to ensure protocol compliance and quality. Selected videos are then processed through deep learning (DL)-based modules such as STT (speech-to-text), Key-point Detector (pose estimation), and Ball Detector (object detection) to extract sub-features. These sub-features are then transformed into clinically interpretable

behavioral features. The extracted features are used to train machine learning (ML) classifiers for each of the three structured video tasks (name-response, imitation, and ball-playing). Finally, predictions from the task-specific models are integrated through an ensemble method based on confidence scores to yield the probability of ASD.

assessments were administered by examiners who achieved research reliability under certified supervision.

Children were classified as ASD if they met either of the following: (1) ADOS-2 score equal to or above the autism spectrum cutoff; or (2) K-CARS-2 score ≥ 30 ⁵³. TD children met all of the following: (1) normal range across all screening tools; (2) no evidence of language delay; (3) no medical, surgical, or neurological conditions; (4) no first-degree relatives diagnosed with ASD; and (5) no history of prematurity (gestational age < 36 weeks).

Video recording protocol

Following enrollment and diagnostic classification, participants completed structured home video recordings using a standardized mobile application. This application was designed to capture core social interaction behaviors relevant for early ASD screening, informed by validated clinical frameworks such as the Early Social Communication Scale (ESCS)⁵⁴ and the ADOS-2³⁸.

The mobile application provided parents with comprehensive recording instructions, embedded instructional videos, automated framing guides, and upload functions to ensure standardization across home environments. Detailed technical protocols for device setup, environmental controls, and task execution are described in Supplementary Note 2.

Parents recorded three structured interaction tasks at home:

- Name-response task: parents called the child’s name from outside the child’s visual field to assess social orienting. Repetitions or familiar sounds were used if no response was observed within 5 s.
- Imitation task: parents demonstrated simple motor actions (hand-raising and clapping) to assess imitation skills, with variations depending on the child’s age. Multiple prompts were allowed when necessary.
- Ball-playing task: parents engaged in reciprocal turn-taking by rolling a ball to the child, initially using non-verbal gestures, followed by verbal encouragement if needed.

Each video was approximately 1 min in duration, and recordings were restricted from being paused during the first 5 s to capture spontaneous responses.

All videos were reviewed for protocol compliance and quality control. Videos with critical protocol deviations or technical issues were excluded. Re-recordings were permitted when protocol violations were identified.

Only one video per task per child was retained. Vertically recorded videos were excluded. Following quality control and group balancing through random sampling, the final dataset consisted of 253 ASD and 257 TD videos (Fig. 3). An independent validation set containing 158 additional videos (90 ASD and 68 TD) was also prepared for external model evaluation.

Clinical feature extraction overview

We implemented a structured feature extraction pipeline developed collaboratively by AI and clinical experts. This process comprised (1) task-specific feature extraction and (2) common clinical features extraction in detail.

Task-specific feature extraction

We extracted features from each structured video task, namely name-response, imitation, and ball-playing, using a combination of gaze-based, audio-based and motion-based cues, as described below.

- Name-response task: in a name-response video, the child’s response to the parent’s call can be vocal or behavioral. The STT⁴⁰ model was used to convert the video audio into text, and specific keywords in the STT output were used to identify children’s vocal responses. To evaluate behavioral responses, we first established a *gaze estimation framework*: a “gaze vector” was defined as a line orthogonal to the ear-to-ear axis and passing through the nose (Fig. 4a), derived from 17 keypoints. Shifts in the vector’s length and orientation were interpreted as changes in gaze direction, serving as a non-intrusive proxy for joint attention and social engagement. Based on this framework, changes in gaze direction, indicated by a shortened gaze vector and altered eye positions, suggested that the child had turned toward the parent. Response latency was defined as the time from parental prompt to the child’s initial vocal or behavioral response to initiate a vocal or behavioral response. Parental attempts were captured by analyzing the frequency of parents’ calls in the STT output.
- Imitation task: in the imitation video, the parent performed gestures, such as clapping and arm-raising, which the child was encouraged to imitate. Movements were recognized based on predefined rules applied to key point configurations, particularly arm-raising actions. The alignment of the key points for the wrist, elbow, and shoulder along the Y-axis, the angle between the elbow and shoulder, and the elbow-wrist

Table 5 | Age-specific psychological assessments for participant selection

Age	Screening test		Additional assessment
	ASD screening tests	Language and functional level assessment	
18–23 months	K-DST, BeDevel, M-CHAT, Q-CHAT	SELSI, CBCL 1.5-5, K-VABS	ADOS-2
24–29 months	K-DST, BeDevel, M-CHAT, Q-CHAT, SCQ	SELSI or PRES, CBCL 1.5-5, K-VABS	ADOS-2, K-CARS-2
30 months	K-DST, BeDevel, M-CHAT, Q-CHAT, SCQ, SRS	SELSI or PRES, CBCL 1.5-5, VABS	ADOS-2, K-CARS-2
31–42 months	K-DST, BeDevel, SCQ, SRS	SELSI or PRES, CBCL 1.5-5, VABS	ADOS-2, K-CARS-2
43–48 months	K-DST, SCQ, SRS-2	SELSI or PRES, CBCL 1.5-5, VABS	ADOS-2, K-CARS-2

ADOS-2 Autism Diagnostic Observation Schedule, K-CARS Korean version of the Childhood Autism Rating Scale, Second Edition, K-DST Korean Developmental Screening Test for Infants and Children, BeDevel-//P Behavior; Development Screening for Toddlers-Interview/play, M-CHAT Modified Checklist for Autism in Toddlers, Q-CHAT Quantitative Checklist for Autism in Toddlers, SELSI Sequenced Language Scale for Infants, CBCL Korean version of the Child Behavior Checklist 1.5–5, K-VABS Korean Vineland Adaptive Behavior Scales, Second Edition, SCQ-L Social Communication Questionnaire Lifetime Version, SRS-2 Social Responsiveness Scale, Second Edition, Preschool version, PRES Preschool Receptive-Expressive Language Scale.

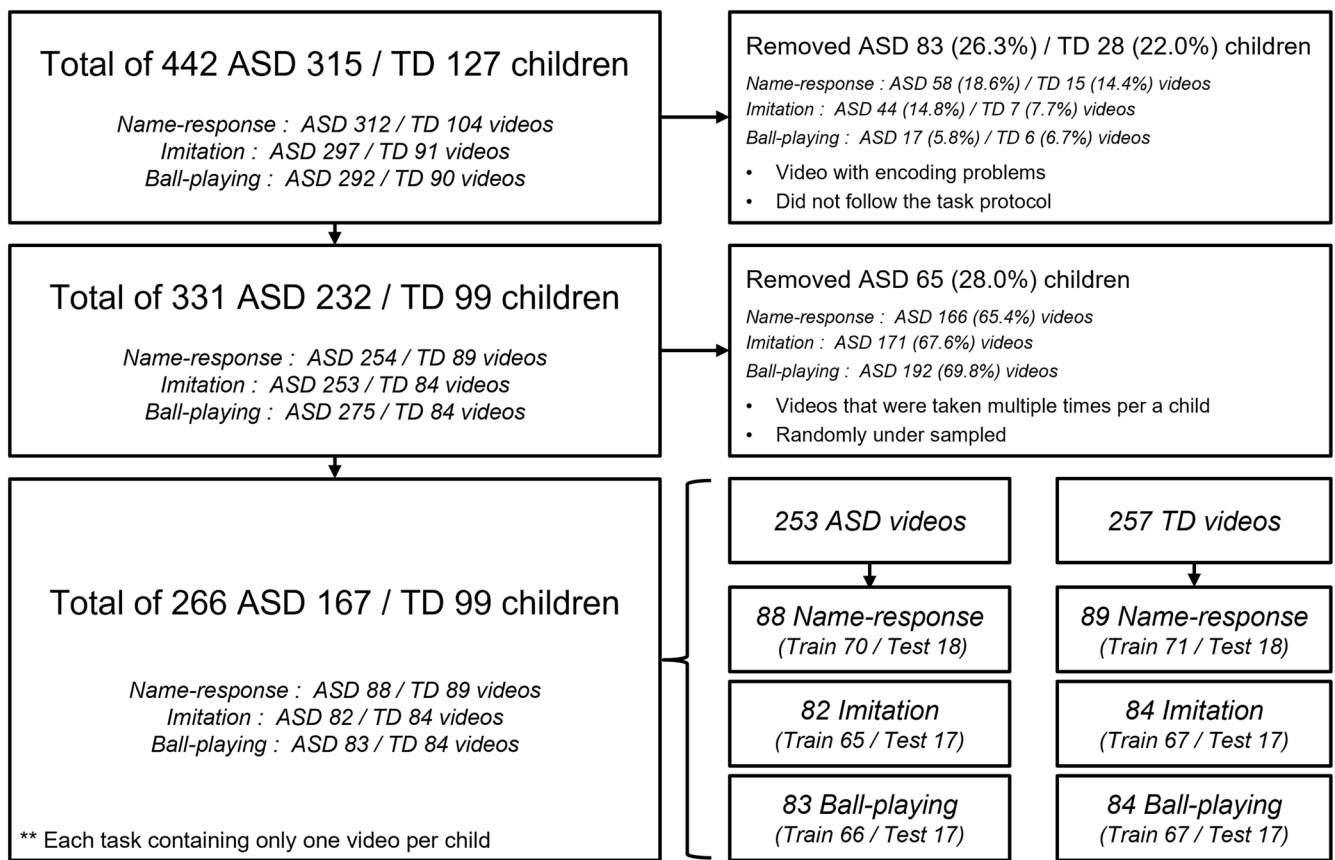


Fig. 3 | Dataset workflow. ASD autism spectrum disorder, TD typically developing.

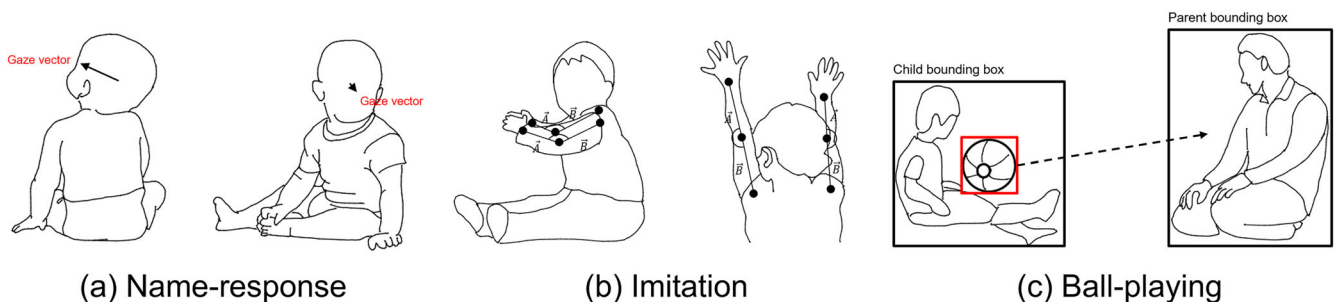


Fig. 4 | Examples of scenarios from the three video types. Our study utilizes three types of scenario videos: **a** an example of a “name-response” video with a gaze vector overlaid, **b** an example of a “imitation” video showing key points for the wrist, elbow, and shoulder, and **c** an example of a “ball-playing” video with bounding boxes drawn around each participant and the ball.

vectors (Fig. 4b) were used to detect arm-raising gestures. The extracted features included the number of parental attempts before the child successfully imitated the action and the response latency from the start of the parent’s action to the child’s imitation.

- Ball-playing task: in the ball-playing video, the parents placed a ball between themselves and the child, observing whether the child would pass the ball back through hand extension without verbal prompts (Fig. 4c). A pre-trained object detector⁴² based on the COCO⁵⁵ dataset was used to detect the position of the ball, and ownership was determined by calculating the intersection over union (IoU) values between the bounding boxes of the child, parent, and ball. A valid interaction was defined as a sequential transfer of the ball from the child to the parent. The time taken for the ball to be transferred from the child to the parent was recorded, along with whether the action was performed.

Common clinical features extraction

In addition to task-specific features, several common behavioral markers were extracted across imitation and ball-playing tasks, capturing broader social engagement indicators such as lack of eye contact, non-engaged movements, and physical contact.

- Lack of eye contact: reduced eye contact is a well-established early indicator of ASD, reflecting difficulties in social attention and joint engagement⁵⁶. Using the previously defined gaze vector in the name-response, we estimated the child’s visual attention toward the parent. When the child focused on the parent, the dot product of the gaze vector increased. The cumulative duration for which the child did not focus on the parent was recorded as the clinical feature. This unobtrusive approach allowed for consistent estimation of gaze direction across videos without requiring specialized eye-tracking hardware.
- Non-engaged movements: motor restlessness and lack of sustained engagement are often observed in children with ASD, especially during social tasks^{57,58}. These behaviors may reflect underlying challenges with self-regulation or attention. We analyzed the centroid, height-to-width ratio, and detection status of the child’s bounding box to quantify the time the child spent moving during the video. This provided a measure of the child’s activity level and inability to remain still. Dynamics of the bounding box were utilized as a non-intrusive proxy for overall movement and restlessness in naturalistic conditions.
- Physical contact: during the video tasks, we frequently observed instances where the child did not follow parental instructions or disengaged from the task (e.g., walking away or becoming unresponsive).

Given that children with ASD often show reduced responsiveness to verbal instructions and a tendency to disengage from structured tasks, and that previous studies have found that parents of children with ASD use more gestures to facilitate engagement⁵⁹, we frequently observed parents touching the child’s body, such as gently placing a hand on the shoulder or arm, while speaking to them or attempting to draw their attention back to the task. Such instances of physical contact likely reflect naturalistic parental strategies to manage noncompliance or disengagement, particularly when verbal or gestural prompts alone are insufficient. To detect instances of physical contact with the parent during the interaction, we identified when the parent’s wrist key points entered the child’s bounding box. The cumulative duration of these instances was recorded as an indicator of the degree of parental involvement. This measure serves as an indirect proxy for child compliance and regulation difficulties, as well as the caregiver’s effort to maintain engagement through physical prompts. The use of 2D wrist keypoints enables unobtrusive detection of such events in naturalistic home settings without requiring manual annotation or wearable sensors.

ML classification models

As the final step in the analysis pipeline, the extracted task-specific and common clinical features were aggregated into a tabular dataset for training ML models, including linear models (logistic regression)⁶⁰, tree-based methods (LightGBM, XGBoost, CatBoost, random forest, gradient boosting classifier, AdaBoost)^{32,61-65}, support vector machines⁶⁶, k-nearest neighbors⁶⁷, and multi-layer perceptron models⁶⁸.

To determine the optimal model for each task, we performed a stratified 10-fold cross-validation on 80% of the data, reserving 20% as an independent hold-out test set. The model with the highest mean validation AUROC was selected. Separate models were trained for each video type, and soft ensemble techniques were applied to the children who appeared in two or more videos in the test set. A detailed explanation of the model selection process is illustrated in Fig. 5.

Statistical analysis

To assess whether the extracted behavioral metrics differed significantly between ASD and TD groups, we performed group-level statistical comparisons separately for task-specific features (Table 3) and common clinical features (Table 4). As the two groups were independent, we used independent two-sample t-tests for all comparisons. This test was chosen to evaluate whether the mean value of each feature metric differed significantly

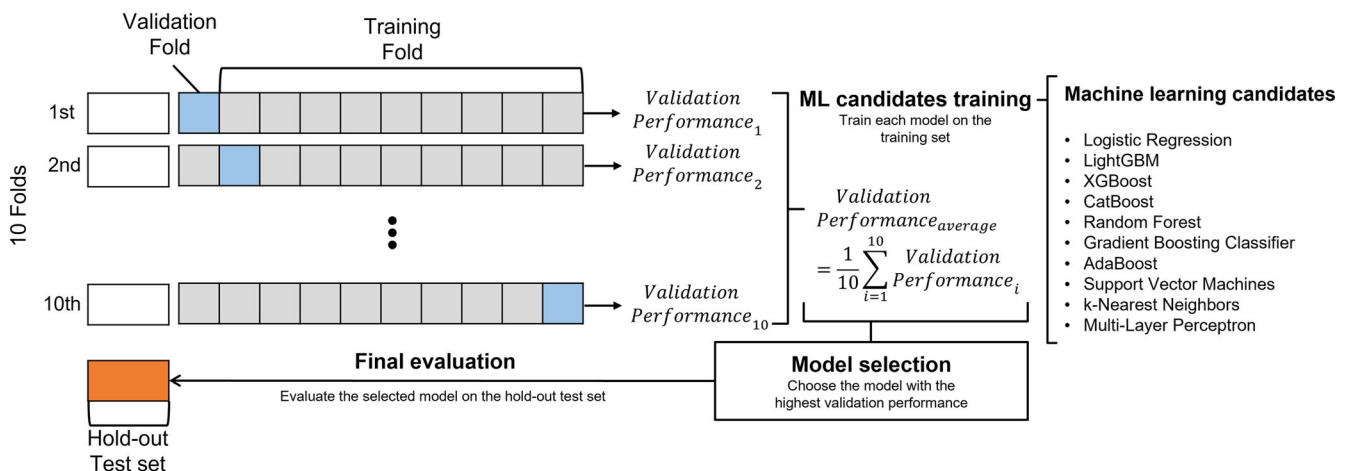


Fig. 5 | Overview of the model selection process. The dataset was split into a training-validation set (80%) and an independent hold-out test set (20%). Candidate models were trained separately for each video task (Name-response, imitation, and ball-playing) using stratified ten-fold cross-validation, and the model with the

highest mean validation area under the receiver operating characteristic curve (AUROC) was selected. The selected models were then evaluated on the hold-out test set. ML machine learning.

between groups, under the assumption of independent observations. All analyses were conducted using Python (v3.8.19) with the scikit-learn (v1.2.2) and SciPy (v.1.10.1) packages. A significance threshold of $p < 0.05$ (two-tailed) was applied for all tests.

Clinical review of AI model on test videos

Although quantitative metrics such as AUROC provide objective measures of model performance, they cannot fully capture the nuanced behavioral interpretations required in real-world pediatric ASD screening. To complement these numerical evaluations and assess the model's alignment with expert clinical judgment, we conducted an independent clinical review of the AI model's predictions on the test videos. All test videos were independently reviewed by a clinical psychologist (doctoral level) and a pediatric psychiatry resident (master's level), classified by the AI model. We examined three aspects: whether the child in the video was ASD or TD, how the child responded to parental instructions, and how these clinical observations compared with the AI model's classification outcomes and evaluation metrics. The test videos were categorized into four groups based on the clinical assessment and ML classification: true positive, false negative, false positive, and true negative. Differences in psychological test results among the four groups were analyzed, with detailed results provided in Supplementary Note 1 and Supplementary Tables 3–5.

Data availability

The datasets generated and analyzed during the current study are not publicly available because of privacy and confidentiality concerns, but are available from the corresponding author upon reasonable request.

Code availability

The underlying code for this study is not publicly available, but may be made available to qualified researchers upon reasonable request from the corresponding author. All analyses were conducted using Python 3.8.19 with PyTorch 2.4.0. Key libraries included Ultralytics 8.2.76, PyCaret 3.2.0, OpenAI Whisper (20231117), OpenCV-Python 4.10.0.84, and scikit-learn 1.2.2.

Received: 18 February 2025; Accepted: 4 September 2025;
Published online: 10 October 2025

References

- Association, A. P. *Diagnostic and Statistical Manual of Mental Disorders* 5th edn, text revision (DSM-5-TR) edn (American Psychiatric Association, 2022).
- Salari, N. et al. The global prevalence of autism spectrum disorder: a comprehensive systematic review and meta-analysis. *Ital. J. Pediatr.* **48**, 112 (2022).
- Global Burden of Disease Study 2021 Autism Spectrum Collaborators. The global epidemiology and health burden of the autism spectrum: findings from the Global Burden of Disease Study 2021. *Lancet Psychiatry* **12**, 111–121 (2025).
- Robain, F., Franchini, M., Kojovic, N., Wood de Wilde, H. & Schaer, M. Predictors of treatment outcome in preschoolers with autism spectrum disorder: an observational study in the Greater Geneva Area, Switzerland. *J. Autism Dev. Disord.* **50**, 3815–3830 (2020).
- van 't Hof, M. et al. Age at autism spectrum disorder diagnosis: a systematic review and meta-analysis from 2012 to 2019. *Autism* **25**, 862–873 (2021).
- Maenner, M. J. et al. *Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years—Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2016*. 12 (CDC, MMWR Surveillance Summaries, 2020).
- Bong, G. et al. Short caregiver interview and play observation for early screening of autism spectrum disorder: behavior development screening for toddlers (BeDevel). *Autism Res.* **14**, 1472–1483 (2021).
- Bahn, G. H. The role of pediatric psychiatrists in the National Health Screening Program for infants and children in Korea. *J. Korean Neuropsychiatr. Assoc.* **59**, 176–184 (2020).
- Bahn, G. H. & Lee, K. S. *Development of the Comprehensive Assessment Inventory for Differential Diagnosis and the Evaluation of Comorbidity of Developmental Delay Kids Under 7 Years Old* (Ministry of Health and Welfare, Korea, 2017).
- Lord, C. et al. *ADOS-2: Autism Diagnostic Observation Schedule* (Hogrefe, 2016).
- Rutter, M., Le Couteur, A. & Lord, C. *Autism Diagnostic Interview—Revised (ADI-R)* (Western Psychological Services, 2003).
- Galliver, M., Gowling, E., Farr, W., Gain, A. & Male, I. Cost of assessing a child for possible autism spectrum disorder? An observational study of current practice in child development centres in the UK. *BMJ Paediatr. Open* **1**, e000052 (2017).
- Sturner, R. et al. Autism screening at 18 months of age: a comparison of the Q-CHAT-10 and M-CHAT screeners. *Mol. Autism* **13**, 2 (2022).
- Duvall, S. W. et al. Factors associated with confirmed and unconfirmed autism spectrum disorder diagnosis in children volunteering for research. *J. Autism Dev. Disord.* **55**, 1660–1672 (2025).
- Steinhart, S., Gilboa, Y., Sinvani, R. T. & Gefen, N. Home videos for remote assessment in children with disabilities: a scoping review. *Telemed. e-Health* **30**, e1629–e1648 (2024).
- Stronach, S. & Wetherby, A. M. Examining restricted and repetitive behaviors in young children with autism spectrum disorder during two observational contexts. *Autism* **18**, 127–136 (2014).
- Lindenschot, M. et al. Perceive, recall, plan and perform (PRPP)-assessment based on parent-provided videos of children with mitochondrial disorder: action design research on implementation challenges. *Phys. Occup. Ther. Pediatr.* **43**, 74–92 (2023).
- Pellicano, E. & den Houting, J. Annual research review: shifting from 'normal science' to neurodiversity in autism science. *J. Child Psychol. Psychiatry* **63**, 381–396 (2022).
- Schuck, R. K. et al. Neurodiversity and autism intervention: reconciling perspectives through a naturalistic developmental behavioral intervention framework. *J. Autism Dev. Disord.* **52**, 4625–4645 (2022).
- Cai, M. et al. An advanced deep learning framework for video-based diagnosis of ASD. In *Medical Imaging Computing and Computer Assisted Intervention – MICCAI 2022*. 434–444 (2022).
- Koehler, J. C. et al. Machine learning classification of autism spectrum disorder based on reciprocity in naturalistic social interactions. *Transl. Psychiatry* **14**, 76 (2024).
- Zunino, A. et al. Video gesture analysis for autism spectrum disorder detection. In *24th International Conference on Pattern Recognition (ICPR)* 3421–3426 (2018).
- Megerian, J. T. et al. Evaluation of an artificial intelligence-based medical device for diagnosis of autism spectrum disorder. *NPJ Digital Med.* **5**, 57 (2022).
- Tariq, Q. et al. Mobile detection of autism through machine learning on home video: a development and prospective validation study. *PLoS Med.* **15**, e1002705 (2018).
- Tariq, Q. et al. Detecting developmental delay and autism through machine learning models using home videos of Bangladeshi children: development and validation study. *J. Med. Internet Res.* **21**, e13822 (2019).
- Abbas, H., Garberson, F., Glover, E. & Wall, D. P. Machine learning approach for early detection of autism by combining questionnaire and home video screening. *J. Am. Med. Inform. Assoc.* **25**, 1000–1007 (2018).
- He, S. & Liu, R. Developing a new autism diagnosis process based on a hybrid deep learning architecture through analyzing home videos. In *Proceedings of the International Conference on Artificial Intelligence*

- and Machine Learning for Healthcare Applications (ICAIMLHA 2021) 1–11 (IEEE, 2021).
28. Wu, C. et al. In *2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM)* 1–6 (2021).
 29. S., J. B., Pandian, D., Rajagopalan, S. S. & Jayagopi, D. Detecting a child's stimming behaviours for autism spectrum disorder diagnosis using Rbpose-Slowfast network. In *2022 IEEE International Conference on Image Processing (ICIP)* 3356–3360 (2022).
 30. Krishnappa Babu, P. R. et al. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* 1–7 (2024).
 31. Li, Q. et al. Prevalence and trends of developmental disabilities among US children and adolescents aged 3 to 17 years, 2018–2021. *Sci. Rep.* **13**, 17254 (2023).
 32. Ke, G. et al. LightGBM: a highly efficient gradient boosting decision tree. In: *31st Conference on Neural Information Processing Systems (NIPS 2017)* (Curran Associates, Inc., 2017).
 33. LaValley, M. P. Logistic regression. *Circulation* **117**, 2395–2399 (2008).
 34. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*. Vol. 30 (2017).
 35. Ha, E. H. Discriminant validity of the CBCL 1.5–5 in the screening of developmental delayed infants. *Korean J. Clin. Psychol.* **30**, 137–158 (2011).
 36. Leblanc, E. et al. Feature replacement methods enable reliable home video analysis for machine learning detection of autism. *Sci. Rep.* **10**, 21245 (2020).
 37. Ozonoff, S. et al. Diagnosis of autism spectrum disorder after age 5 in children evaluated longitudinally since infancy. *J. Am. Acad. Child Adolesc. Psychiatry* **57**, 849–857.e842 (2018).
 38. Yoo, H. J. et al. *Korean Version of Autism Diagnostic Observation Schedule (ADOS)* (Hakjisa, 2023).
 39. Bacon, E. C. et al. Rethinking the idea of late autism spectrum disorder onset. *Dev. Psychopathol.* **30**, 553–569 (2018).
 40. Radford, A. et al. In *International Conference on Machine Learning* 28492–28518 (2023).
 41. Xiao, B., Wu, H. & Wei, Y. Simple baselines for human pose estimation and tracking. In *Proceedings of the 2018 European Conference on Computer Vision (ECCV)* 466–481 (2018).
 42. Varghese, R. & Sambath, M. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)* 1–6 (2024).
 43. Campbell, K. et al. Computer vision analysis captures atypical attention in toddlers with autism. *Autism* **23**, 619–628 (2019).
 44. Yim, C.-H., Kim, G.-H. & Eun, B.-L. Usefulness of the Korean Developmental Screening Test for infants and children for the evaluation of developmental delay in Korean infants and children: a single-center study. *Korean J. Pediatr.* **60**, 312–319 (2017).
 45. Bong, G. et al. The feasibility and validity of autism spectrum disorder screening instrument: Behavior Development Screening for Toddlers (BeDevel)—a pilot study. *Autism Res.* **12**, 1112–1128 (2019).
 46. Park, H. H., Hong, K. H., Hong, S. M. & Kim, S. J. The validity of the Korean version of Modified-Checklist for Autism in Toddlers-Revised (KM-CHAT-R). *Korean J. Early Child. Spec. Educ.* **15**, 1–20 (2015).
 47. Park, S. et al. Reliability and validity of the Korean translation of quantitative checklist for autism in toddlers: a preliminary study. *J. Korean Acad. Child Adolesc. Psychiatry* **29**, 80–85 (2018).
 48. Kim, Y. T. Content and reliability analyses of the Sequenced Language Scale for Infants (SELSI). *Commun. Sci. Disord.* **7**, 1–23 (2002).
 49. Hwang, S.-T., Kim, J.-H., Hong, S.-H., Bae, S.-H. & Jo, S.-W. Standardization study of the Korean Vineland Adaptive Behavior Scales-II (K-Vineland-II). *Korean J. Clin. Psychol.* **34**, 851–876 (2015).
 50. Kim, J.-H. et al. A validation study of the Korean version of social communication questionnaire. *J. Korean Acad. Child Adolesc. Psychiatry* **26**, 197–208 (2015).
 51. Chun, J., Bong, G., Han, J. H., Oh, M. & Yoo, H. J. Validation of Social Responsiveness Scale for Korean preschool children with autism. *Psychiatry Investig.* **18**, 831–840 (2021).
 52. Kim, Y.-T. Content and reliability analyses of the preschool receptive-expressive language scale (PRES). *Commun. Sci. Disord.* **5**, 1–25 (2000).
 53. Lee, S., Yoon, S.-A. & Shin, M.-S. Validation of the Korean childhood autism rating. <https://doi.org/10.1016/j.rasd.2023.102128>. (2023).
 54. Mundy, P. et al. *Early Social Communication Scales (ESCS)* (MIND Institute, University of California at Davis, 2003).
 55. Lin, T.-Y. et al. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V* 13 740–755 (2014).
 56. Li, J. et al. Appearance-based gaze estimation for ASD diagnosis. *IEEE Trans. Cyber* **52**, 6504–6517 (2022).
 57. Dong, L. et al. A comparative study on fundamental movement skills among children with autism spectrum disorder and typically developing children aged 7–10. *Front. Psychol.* **15**, 1287752 (2024).
 58. Minissi, M. E. et al. The whole-body motor skills of children with autism spectrum disorder taking goal-directed actions in virtual reality. *Front. Psychol.* **14**, 1140731 (2023).
 59. Yoshida, H., Cirino, P., Mire, S. S., Burling, J. M. & Lee, S. Parents' gesture adaptations to children with autism spectrum disorder. *J. Child Lang.* **47**, 205–224 (2020).
 60. Cox, D. R. The regression analysis of binary sequences. *J. R. Stat. Soc. Ser. B: Stat. Methodol.* **20**, 215–232 (1958).
 61. 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*. pp. 785–794. (2016).
 62. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* 6639–6649 (Curran Associates Inc., Montréal, Canada, 2018).
 63. Ho, T. K. In *Proceedings of 3rd International Conference on Document Analysis and Recognition* 278–282 (1995).
 64. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
 65. Schapire, R. E. In *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik* 37–52 (Springer, 2013).
 66. Cortes, C. & Vapnik, V. Support-Vector Networks. *Mach. Learn.* **20**, 273–297 (1995).
 67. Peterson, L. E. K-nearest neighbor. *Scholarpedia* **4**, 1883 (2009).
 68. Haykin, S. *Neural Networks: A Comprehensive Foundation* (Prentice Hall PTR, 1998).

Acknowledgements

This study was supported by a research fund from the National Center for Mental Health, Ministry of Health & Welfare, Republic of Korea (grant number: DMHR25E01).

Author contributions

K.A.C., Y.G.K., and B.N.K. conceptualized and designed the study. H.L., S.H., J.Y., H.C., J.L., M.H.P., J.L., A.C., C.M.Y., D.L., H.Y., Y.L., G.B., J.I.K., H.S., H.W.K., E.J., S.C., J.W.S., J.H.Y., Y.B.L., J.C., W.C., S.L., S.P., J.A., C.R.L., and S.J. contributed to data acquisition and quality control. D.Y.K., R.D., Y.S., G.L., S.P., B.J., and J.J. contributed to the analysis. H.L., H.S., and H.K. contributed to data interpretation. D.Y.K. and R.D. drafted the manuscript. All authors had full access to the study design information and all data and approved the final version to be published. Agreement to be accountable for all aspects of the work to ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41746-025-01993-5>.

Correspondence and requests for materials should be addressed to Keun-Ah Cheon, Young-Gon Kim or Bung-Nyun Kim.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025

¹Interdisciplinary Program in Bioengineering, Seoul National University College of Engineering, Seoul, Republic of Korea. ²Department of Transdisciplinary Medicine, Seoul National University Hospital, Seoul, Republic of Korea. ³Biomedical Research Institute, Seoul National University Hospital, Seoul, Republic of Korea. ⁴Interdisciplinary Program in Medical Informatics, Seoul National University College of Medicine, Seoul, Republic of Korea. ⁵Department of Medical Device Development, Seoul National University College of Medicine, Seoul, Republic of Korea. ⁶Department of Psychiatry, Institute of Behavioral Science in Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea. ⁷Department of Child and Adolescent Psychiatry, Yonsei University Severance Hospital, Seoul, Republic of Korea. ⁸Department of Psychiatry, Eunpyeong St. Mary's Hospital, Seoul, Republic of Korea. ⁹Department of Psychiatry, Wonkwang University Hospital, Iksan, Republic of Korea. ¹⁰Department of Psychiatry, Wonkwang University College of Medicine, Iksan, Republic of Korea. ¹¹Department of Psychiatry, Seoul National University Bundang Hospital, Seongnam, Republic of Korea. ¹²Department of Psychiatry, Seoul National University College of Medicine, Seoul, Republic of Korea. ¹³Department of Psychiatry, Hanyang University College of Medicine, Seoul, Republic of Korea. ¹⁴Institute of Mental Health, Hanyang University Industry-University Cooperation Foundation, Seoul, Republic of Korea. ¹⁵Department of Psychiatry, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea. ¹⁶Yonsei Jaram Psychiatry Clinic, Seoul, Republic of Korea. ¹⁷Department of Psychiatry, Chungbuk National University Hospital, Cheongju, Republic of Korea. ¹⁸Department of Psychiatry, Seoul St. Mary's Hospital, The Catholic University of Korea College of Medicine, Seoul, Republic of Korea. ¹⁹Vision R&D, SK Telecom, Seongnam, Republic of Korea. ²⁰Department of Child and Adolescent Psychiatry, Seoul National University Hospital, Seoul, Republic of Korea. ²¹Department of Psychiatry and the Institute of Behavioral Science in Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea. ²²Department of Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea. ²³Innovative Medical Technology Research Institute, Seoul National University Hospital, Seoul, Republic of Korea. ²⁴These authors contributed equally: Dong Yeong Kim, Ryemi Do. ²⁵These authors jointly supervised this work: Keun-Ah Cheon, Young-Gon Kim, Bung-Nyun Kim. ✉ e-mail: kacheon@yuhs.ac; younggon2.kim@gmail.com; kbn1@snu.ac.kr