

AI-Assisted Early Prediction of Autism Spectrum Disorder in Children Using Clinical and Developmental Features

Dr Sumera Kanwal¹, Dr. Hafiz Shafique Ahmad², Dr. Fareeha F. Khan^{3*}, Dr. Zaid Waqar⁴, Sher Muhammad⁵, Dr Ayesha Rafique⁶, Fahad Asim⁷

¹ Designation, Affiliation/Institute name & City*: Assistant Professor, Faculty of Health and Medical Sciences, Metropolitan University, Karachi, Contact no.: 0335-3789240

² Associate Professor of Psychiatry, DGKMC DGK

^{3*}Designation: Physician -Researcher/ Senior medical officer ICU, Affiliation: The Kidney Center, Karachi
<https://orcid.org/0009-0004-3391-1326>

⁴ Assistant Professor of Neurology, Department of clinical neurosciences, college of medicine, king Faisal University, Al Ahsa, KSA

⁵ Advanced Research Center, European University of Lefke, Lefke, Northern Cyprus, TR-10 Mersin, Turkey
00923008435768, <https://orcid.org/0000-0002-5235-2193>

⁶ Designation: SMO, Department: Pediatrics, Institute: Faisal Hospital, Faisalabad

⁷ Designation: Lecturer in Pharmacology & Therapeutics, Faculty of Pharmacy, The University of Lahore, Lahore, Pakistan

sumerakanwal_ku@yahoo.com ¹, Drhshafiq@gmail.com², drfareeha42@gmail.com^{3*}, drzaidwaqar@gmail.com⁴, sher1325@hotmail.com⁵, m_a_tabassum@hotmail.com⁶, fahadasim6@gmail.com ⁷

Corresponding Author: **Dr. Fareeha F. Khan^{3*}**

HOW TO CITE: | ABSTRACT:

Sumera Kanwal, Hafiz Shafique Ahmad, Fareeha F. Khan, Zaid Waqar, Sher Muhammad, Ayesha Rafique, Fahad Asim (2026). AI-Assisted Early Prediction of Autism Spectrum Disorder in Children Using Clinical and Developmental Features. International Journal of Special Education, 41(24s), 1288 - 1300

Background: Machine-learning models are increasingly applied to autism spectrum disorder (ASD) screening and frequently report exceptionally high classification effectiveness. However, such results may be inflated when recorded screening labels are derived from questionnaire responses that also serve as model predictors. This study considers the influence of questionnaire-derived features on predictive performance and examines whether demographic and contextual characteristics provide independent discriminatory information in toddler and child autism screening datasets.

Methods: Two publicly available datasets, comprising 1,054 toddler and 292 child screening records, were analyzed using five machine-learning algorithms: logistic regression, random forest, support vector machines, gradient boosting, and XGBoost. Models were evaluated using stratified 75/25 training-testing splits, five-fold training cross-validation, and 400-resample bootstrap confidence intervals. Three predictor configurations were compared: complete features, exclusion of aggregate screening scores, and exclusion of both aggregate scores and

individual questionnaire responses. Predictive discrimination was assessed primarily using the area under the receiver operating characteristic curve (ROC-AUC).

Results: Aggregate screening-score thresholds reproduced 100% of recorded classification labels in both datasets, showing a deterministic relationship between screening scores and outcomes. For toddlers, random-forest held-out ROC-AUC was 1.000 with all features, 0.995 after aggregate-score removal, and 0.670 (95% CI: 0.598–0.747) using contextual features alone. Corresponding values in children were 1.000, 0.985, and 0.567 (95% CI: 0.428–0.685), respectively. Across all five algorithms, context-only models showed weak discrimination in children (ROC-AUC: 0.526–0.580). The substantial reduction in performance following questionnaire exclusion suggests that near-perfect classification primarily reflects reconstruction of questionnaire-derived labels rather than independent identification of autism-related clinical characteristics.

Conclusions: Exceptionally high predictive validity in questionnaire-based autism screening datasets may largely reflect target leakage due to overlap between predictors and recorded outcomes. Removing aggregate screening scores alone does not resolve this issue if the underlying questionnaire items remain available. Independent clinician-adjudicated diagnostic results, external validation, and rigorous feature-selection procedures are essential before such models can be considered reliable for clinical applications.

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the
terms and conditions
of the Creative Commons Attribution
(CC BY)
license (<http://creativecommons.org/licenses/by/4.0/>).

Keywords: autism spectrum disorder; developmental screening; machine learning; target leakage; ROC-AUC; cross-validation; clinical validation.

1. Introduction

Autism spectrum disorder (ASD) is a heterogeneous neurodevelopmental disorder defined by differences in social communication and interaction and by restricted or repetitive behaviors and interests. Presentation and developmental course vary substantially among children, complicating early recognition and referral. Some children show developmental differences during the first years of life, while others come to clinical attention later as social and social and communication requirements increase. The clinical assessment of ASD therefore draws on developmental history, observation, caregiver reports, and professional

judgment rather than a single questionnaire score [1,2].

This study concentrates on the extent to which the models reproduce recorded screening classifications and how their performance depends on the questionnaire items used to assign those classifications. The available records do not permit assessment of accuracy against an independent clinical diagnosis. Accordingly, these data should not be interpreted as evidence of diagnostic validity. The results instead describe agreement with questionnaire-derived screening labels, which is distinct from the ability to identify autism as established by clinical evaluation. This distinction matters to understand the study's translational implications and limitations.

The evaluation included cross-validation, confidence intervals, calibration, and feature-importance analyses, along with an examination of the relationship between recorded screening labels and questionnaire scores.

This study tackles these methodical issues through a secondary analysis of two publicly available, age-specific ASD screening datasets comprising 1,054 toddler records and 292 child records. We compare logistic regression, random forest, support vector machine, gradient boosting, and XGBoost under a unified train–test evaluation framework. We consider three predictor settings: the first feature set, a set excluding aggregate screening scores, and a context-only set excluding both aggregate scores and individual questionnaire items. These comparisons show how much apparent predictive performance depends on access to variables that may directly determine the recorded class.

This distinction matters when model effectiveness is reported using standard clinical metrics. Sensitivity, specificity, and area under the receiver operating characteristic curve (ROC-AUC) are always defined relative to the reference outcome used in evaluation. When the outcome is a screening-derived class, these metrics quantify agreement with the screening classification rather than the ability to identify clinician-confirmed autism. Similarly, feature-attribution methods may highlight the questions that most efficiently reconstruct a score without establishing that those questions are independently diagnostic. High discrimination does not necessarily ensure well-calibrated probabilities or practical usefulness in pediatric clinical settings [19–21,25,26,29].

A substantial methodological issue arises when the outcome label is derived from the same questionnaire responses provided to the model. For instance, a dataset may include ten screening-item responses, their aggregate score, and a binary class determined by applying a threshold to that score. A classifier trained to predict this class can achieve very high accuracy by replicating the existing scoring rule, even without learning

about independently confirmed ASD. Excluding only the aggregate score does not necessarily resolve this issue, as the individual item responses may still encode the rule. This constitutes target-information leakage, or circular outcome construction, which may lead to overstatements about clinical prediction [10].

Computational approaches have attracted interest because structured screening records contain patterns that can be analyzed systematically. Machine-learning classifiers can model nonlinear associations among questionnaire responses, demographic characteristics, and developmental indicators, and they may offer reproducible ways to compare alternative predictor sets. Earlier studies have examined abbreviated behavioral instruments and combinations of questionnaires with observationally or home-video data [5,6,11,12,15,16]. Yet these studies differ in recruitment, input modality, reference standards, and clinical context. Their reported performance cannot simply be transferred to publicly distributed screening datasets or interpreted as evidence of diagnostic effectiveness.

Early identification matters because it can facilitate timely developmental assessment, individualized support, and access to services for children and families. Pediatric practice typically combines developmental surveillance with age-appropriate standardized screening and referral when concerns develop. Screening, however, is not equivalent to diagnosis: a positive screening result signals a need for further evaluation, and a negative result does not rule out ASD in a child with persistent developmental concerns. The Modified Checklist for Autism in Toddlers, Revised with Follow-Up (M-CHAT-R/F), illustrates the importance of follow-up assessment and of interpreting test performance within a defined clinical pathway [3,4,13,39,40].

2. Literature and research gap

2.1 Clinical screening versus diagnosis

ASD is defined through developmental and behavioral assessment rather than a single short questionnaire [1,2]. Pediatric guidance

emphasizes surveillance, validated screening, and appropriate referral [3]. The M-CHAT-R/F literature demonstrates the importance of follow-up evaluation and of considering the screening pathway as a whole [4]. Neither uploaded dataset supplies independently adjudicated diagnostic results in the analysis performed here.

2.2 AI and methodological concerns

Earlier behavioral classification studies show that machine learning can distinguish questionnaire patterns [5,6]. The open methodological question is whether such models add information beyond the screening rule used to generate their targets. Here, a third, context-only feature set provides a more demanding test than removal of aggregate scores alone. Small samples, convenience recruitment, and population shift further limit generalizability. Interpretability reveals what predicts the recorded labels; it does not establish etiologic or clinical validity.

2.3 Evidence from clinical screening and AI studies

Large pediatric-network evaluations emphasize that real-world screening performance depends on follow-up and the reference standard [13,14]. Contemporary reviews additionally emphasize

differences across age, language, cultural background, and clinical setting [14,18,24]. Earlier machine-learning studies examined shortened behavioral assessments and combinations of questionnaire and video features [11,12,15,16]. Their research settings and outcome ascertainment differ materially from the two public screening-label datasets analyzed here. Recent systematic reviews summarize increasing algorithmic interest but cannot resolve dataset-specific label circularity [17,23,24].

2.4 Methodological reporting, leakage and clinical utility

TRIPOD+AI and PROBAST+AI emphasize transparent outcome definitions, appropriate participant recruitment, and reproducible evaluation [8,9]. Leakage can lead to optimistic performance when information used to construct the target is also available to the predictor [10]. Discrimination alone does not establish the clinical value of a screening model; the intended clinical population requires evaluation of calibration, decision thresholds, and net benefit [19–21]. In this study, the primary target is recorded screening classification, not a blinded specialist diagnosis.

Table 1. Comparison of selected published evidence.

Study	Main contribution	Limitation for this study
[13] Guthrie et al. (2019)	Pediatric network screening accuracy	Clinical follow-up differs from public screening-label files
[11] Wall et al. (2012)	Abbreviated observation-based autism classification	Research samples and reference standards differ
[15] Abbas et al. (2018)	Questionnaire plus video-based modeling	Multimodal input unavailable in current datasets
[16] Tariq et al. (2018)	Prospective video-based evaluation	Different data modality and clinical pathway
[17] Ding et al. (2024)	Deep-learning evidence synthesis	Heterogeneity limits direct numerical comparison
[18] Huda et al. (2024)	Cultural/linguistic screening differences	Transportability not established here

[9] Moons et al. (2025)	PROBAST+AI framework	bias/applicability	Guidance, not ASD outcome evidence
[10] Kapoor and Narayanan (2023)	Leakage and reproducibility	taxonomy	General ML methods, not ASD-specific

Table 1 contrasts the present screening-label analysis with studies using other recruitment procedures, data modalities, and outcome definitions. Their accuracy estimates are not directly comparable.

3. Data and methods

This retrospective secondary analysis examined the toddler screening CSV and child screening ARFF records as separate age-specific cohorts. The recorded outcomes were questionnaire-derived screening classifications, rather than independently confirmed diagnoses. We did not recruit or clinically assess participants for this analysis.

Table 2. Cohort sizes, recorded screening classifications and duplicate records.

Dataset	N	Positive label	Negative label	Exact rows	duplicate
Toddler	1054	728	326	0	
Child	292	141	151	2	

Feature settings: Original excluded identifiers; No aggregate score additionally excluded Qchat-10-Score (toddlers) or result (children); Context only additionally excluded all A1–A10 questionnaire items. Context-only variables may still encode screening-related information and are not necessarily prospectively available. We excluded the age_desc child field as a descriptor.

The same stratified 75/25 split (seed 42) was used for each feature setting. Within the training partition, a five-fold stratified shuffled cross-validation searched small prespecified grids of logistic regression regularization, random-forest tree/leaf settings, SVM C, and boosting tree settings. All imputation, scaling, and categorical encoding occurred inside training-fold pipelines. We selected hyperparameters by ROC-AUC; we did not use the test set for tuning. We used a 0.50 probability threshold throughout.

For uncertainty, 400 nonparametric bootstrap resamples of held-out predictions generated percentile 95% ROC-AUC intervals; single-class replicates were omitted. Such intervals quantify

sampling variability for the fixed test sample and do not replace independent validation. No paired significance claims were made. We descriptively audited dataset-specific aggregate-score thresholds to reveal potential deterministic target derivation.

We analyzed the datasets without pooling age groups. We obtained all predictors and outcomes from existing records; we collected no additional clinical or hospital data.

3.1 Dataset sources and screening-label definitions

The analysis included 1,054 toddler records and 292 child records. The child dataset is cataloged by the UCI Machine Learning Repository (Thabtah, 2017; DOI: 10.24432/C5659W; CC BY 4.0) [7]. The toddler file is associated with the publicly circulated Q-CHAT-10/ASDTests collection, although the supplied metadata could not establish its precise version or redistribution conditions. Both outcome variables denote screening classifications, not independently confirmed diagnoses.

In the supplied toddler file, all 1,054 labels coincide with the rule that the sum of A1–A10 is at least four. The child dataset contains A1_Score–A10_Score, a result score, and Class/ASD. Removing only the total-score predictor leaves the individual questionnaire responses from which the outcome may be derived; this does not remove target leakage. The context-only analysis therefore excludes the questionnaire items and the aggregate score.

Source identifiers and computational materials are described in the supplementary reproducibility materials. Redistribution of the toddler source file is subject to confirmation of the applicable terms.

3.2 Preprocessing & replication scope

The toddler CSV contains 19 columns, including an identifier and the class field; we observed no pandas-recognized null cells, and the preprocessing pipeline handled categorical responses. The child ARFF contains literal? missing-value placeholders. We median-imputed and standardized numeric variables, and mode-imputed and one-hot encoded categorical variables within the training pipelines. The reported analysis uses a stratified 75/25 split (seed 42), five-fold stratified cross-validation on the training partition, and a limited two-candidate hyperparameter search per classifier.

Reproducibility: The accompanying analysis script defines RandomizedSearchCV with two

sampled hyperparameter combinations, five-fold shuffled stratified cross-validation (random_state=42), and selection by ROC-AUC. The 25% held-out partitions were not used for tuning. The results CSV records the selected settings, fold-averaged AUC, test metrics, and 400-resample percentile ROC-AUC intervals. The first toddler CSV version and redistribution license remain uncertain; the child dataset is cataloged at UCI under DOI 10.24432/C5659W. Neither dataset contains independently clinician-adjudicated outcomes. These issues should be considered under TRIPOD+AI and PROBAST+AI [8,9].

Scope of validation. The analysis uses the downloaded records as provided and internally partitions them for cross-validation and held-out evaluation. No additional participant recruitment, independently adjudicated diagnoses, prospective clinical validation, or external test cohort was available. Accordingly, performance estimates measure agreement with the supplied screening labels and should not be interpreted as clinical diagnostic sensitivity or accuracy.

4. Results

We compared the recorded labels with threshold classifications calculated from the aggregate questionnaire scores. This establishes their correspondence in the analyzed files without assuming access to the original data-collection procedures.

Table 3. Agreement between questionnaire-score thresholds and recorded screening classifications.

Dataset	Aggregate score	Best threshold $\geq$	Agreement with class
Toddler	Qchat-10-Score	4	1.000
Child	result	7.0	1.000

4.1 Toddler cohort

Table 4. Toddler cohort: training cross-validation and held-out model effectiveness.
 Supplementary results contain precision, F1, PR-AUC, Brier scores, and tuned hyperparameters.

Feature set	Model	CV AUC	Test AUC [95% CI]	Sensitivity	Specificity	FN
Original	Logistic regression	1.000	1.000 [1.000, 1.000]	1.000	1.000	0
Original	Random forest	1.000	1.000 [1.000, 1.000]	1.000	1.000	0
Original	SVM	1.000	1.000 [1.000, 1.000]	0.995	1.000	1
Original	Gradient boosting	1.000	1.000 [1.000, 1.000]	1.000	1.000	0
Original	XGBoost	1.000	1.000 [1.000, 1.000]	1.000	1.000	0
No aggregate score	Logistic regression	1.000	1.000 [1.000, 1.000]	1.000	1.000	0
No aggregate score	Random forest	0.995	0.995 [0.989, 0.999]	0.978	0.902	4
No aggregate score	SVM	0.999	1.000 [0.999, 1.000]	0.995	0.988	1
No aggregate score	Gradient boosting	0.996	0.997 [0.994, 1.000]	0.978	0.976	4
No aggregate score	XGBoost	1.000	0.999 [0.998, 1.000]	0.989	0.988	2
Context only	Logistic regression	0.614	0.598 [0.520, 0.680]	0.978	0.134	4
Context only	Random forest	0.618	0.670 [0.598, 0.747]	0.956	0.195	8
Context only	SVM	0.628	0.631 [0.560, 0.709]	0.978	0.146	4
Context only	Gradient boosting	0.606	0.698 [0.625, 0.775]	0.962	0.171	7
Context only	XGBoost	0.616	0.691 [0.616, 0.767]	0.945	0.183	10

4.2 Child cohort

Table 5. Child cohort: training cross-validation and held-out model results. Supplementary results contain precision, F1, PR-AUC, Brier scores, and tuned hyperparameters.

Feature set	Model	CV AUC	Test AUC [95% CI]	Sensitivity	Specificity	FN
Original	Logistic regression	1.000	1.000 [1.000, 1.000]	1.000	1.000	0
Original	Random forest	1.000	1.000 [1.000, 1.000]	1.000	1.000	0
Original	SVM	0.999	0.999 [0.995, 1.000]	1.000	0.947	0
Original	Gradient boosting	1.000	1.000 [1.000, 1.000]	1.000	1.000	0
Original	XGBoost	1.000	1.000 [1.000, 1.000]	1.000	1.000	0
No aggregate score	Logistic regression	1.000	1.000 [1.000, 1.000]	1.000	0.974	0
No aggregate score	Random forest	0.983	0.985 [0.961, 1.000]	0.971	0.921	1
No aggregate score	SVM	0.995	0.997 [0.988, 1.000]	1.000	0.921	0
No aggregate score	Gradient boosting	0.970	0.964 [0.927, 0.990]	0.914	0.842	3
No aggregate score	XGBoost	0.993	0.992 [0.974, 1.000]	1.000	0.921	0
Context only	Logistic regression	0.619	0.566 [0.430, 0.696]	0.543	0.368	16
Context only	Random forest	0.583	0.567 [0.428, 0.685]	0.457	0.605	19
Context only	SVM	0.589	0.580 [0.442, 0.710]	0.543	0.579	16
Context only	Gradient boosting	0.654	0.526 [0.390, 0.658]	0.343	0.789	23
Context only	XGBoost	0.578	0.538 [0.399, 0.677]	0.571	0.474	15

Figure 1 shows the marked decline in ROC-AUC after excluding all questionnaire items.

Figure 1. Effect of feature ablation on held-out ROC-AUC for random forest.

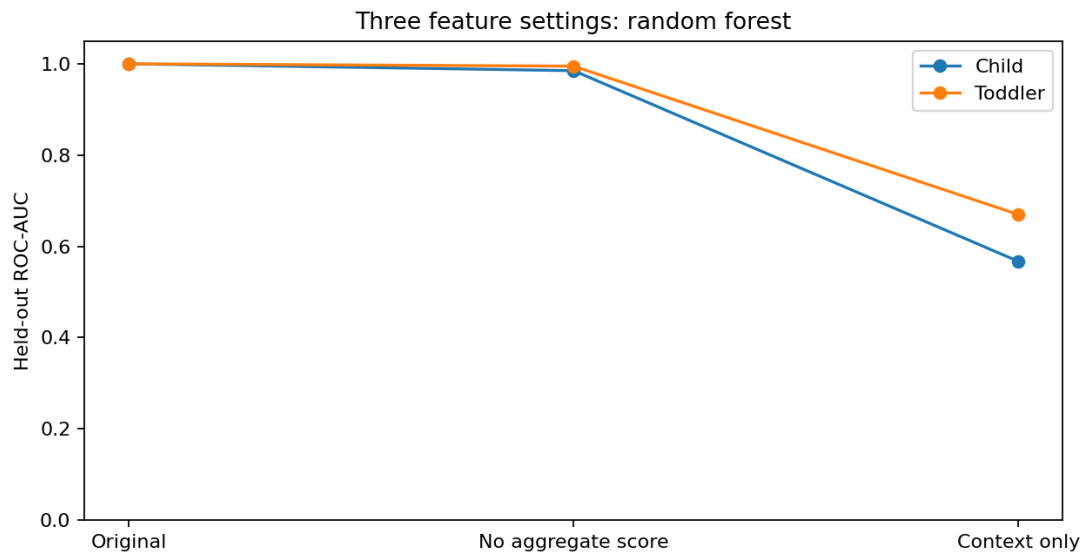


Figure 2 displays exploratory probability calibration against screening labels, not clinical diagnoses.

Figure 2. Exploratory held-out calibration plots; small bins should not be overinterpreted.

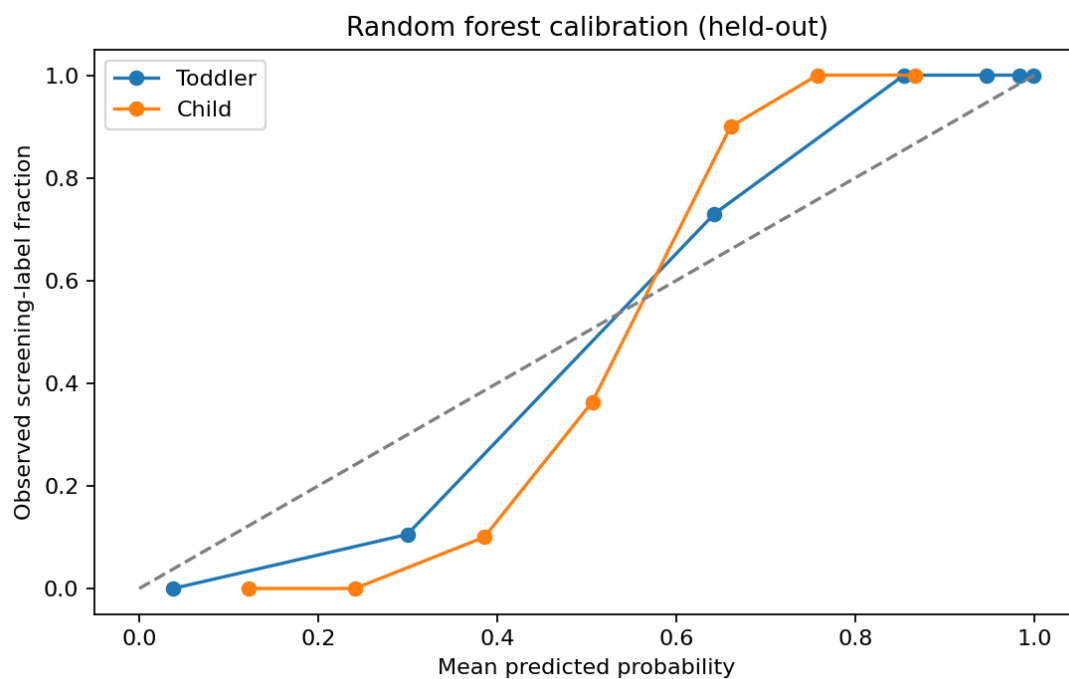
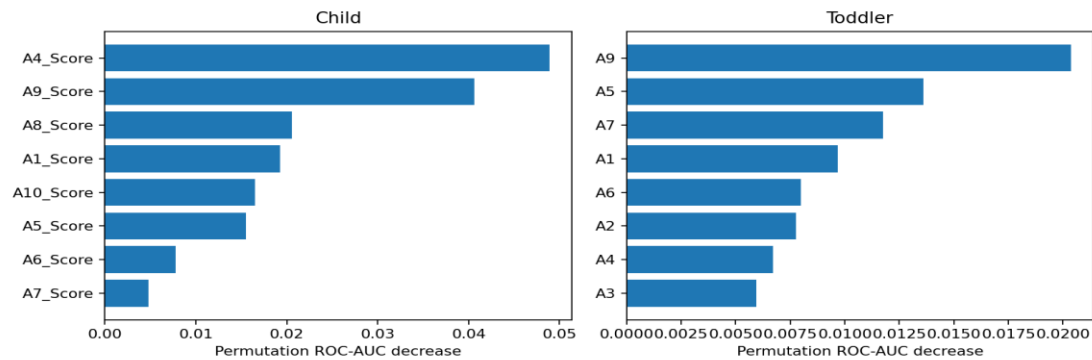


Figure 3 shows that questionnaire items dominate permutation importance in the score-removed random forests.

Figure 3. Test-set permutation importance for no-aggregate-score random forests (descriptive only; not clinical feature validation).



4.3 Subgroup analysis

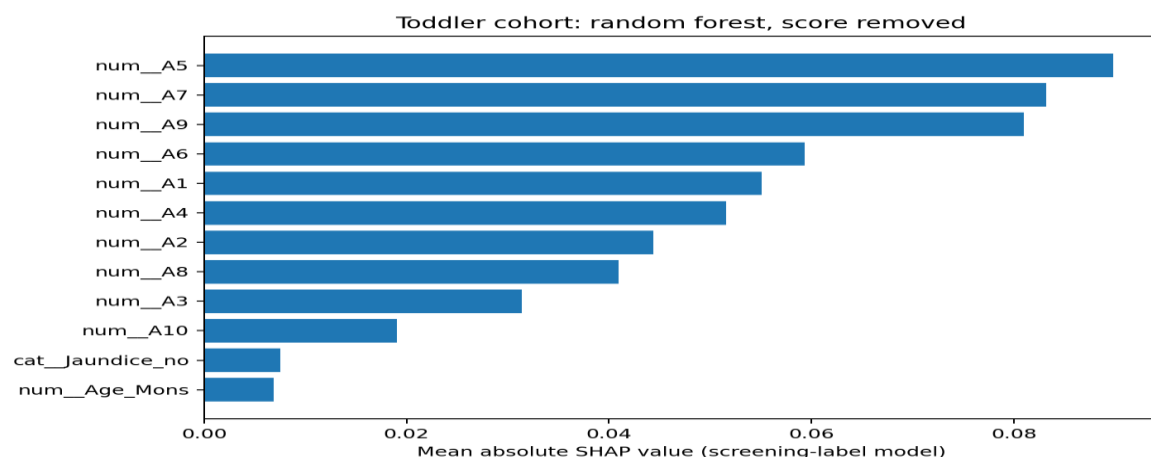
Table 6. Exploratory subgroups for no-aggregate-score logistic regression. Subgroups with fewer than ten records were omitted; results are not for fairness validation.

Dataset	Variable	Subgroup	N	Positive labels	Sensitivity
Toddler	Sex	m	169	123	1.000
Toddler	Sex	f	95	59	1.000
Child	age	(7.5, 11.0]	23	11	1.000
Child	age	(3.993, 7.5]	50	24	1.000
Child	gender	m	53	26	1.000
Child	gender	f	20	9	1.000

4.4 SHAP explanation of toddler screening-label classification

Figure 4 shows mean absolute SHAP values for a random forest trained on the toddler training partition after removing Qchat-10-Score. The plot uses 100 held-out records. These attributions describe the fitted screening-label classifier and do not establish clinical diagnostic importance.

Figure 4. SHAP feature contributions for the toddler no-aggregate-score random forest.



Published screening studies that used clinical follow-up as well as video observations [13,15,16] cannot be directly compared with accuracy for labels reconstructed from questionnaire responses. The relevant clinical question is whether a model can identify children in need of additional assessment beyond what a screening instrument already indicates. The available datasets do not answer that question. Calibration and decision-curve analysis are important for later clinical evaluation, but their interpretation here remains restricted to recorded screening outcomes [19–21,25,26].

5. Discussion

The principal result is the decline in discrimination after removing questionnaire responses. Toddler random-forest held-out ROC-AUC fell from 1.000 (all features) to 0.995 (aggregate score removed) and 0.670 (context only). The corresponding child results were 1.000, 0.985, and 0.567 (Figure 1; Tables 4–5). This suggests that much of the apparent performance comes from recovering the recorded screening classification, rather than extracting independent diagnostic information.

The two cohorts differed in context-only performance. Toddler models reached ROC-AUCs of 0.598–0.698, while child models ranged from 0.526 to 0.580. The smaller child sample, differing recruitment, and available contextual variables may contribute, but these datasets cannot isolate the cause. Demographic and historical variables are not substitutes for clinical developmental assessment. Poor performance in reconstructing screening labels from context alone does not establish that contextual information has no clinical value.

Earlier studies using shortened observation instruments or questionnaire-plus-video assessments [11,12,15,16] employed different input modalities and reference outcomes, so their reported accuracies are not directly comparable. Pediatric-network screening studies [13,14] likewise point out the importance of follow-up and clinical pathways. In this analysis, the simple aggregate-score thresholds reproduce the labels

exactly (Table 3), providing a transparent benchmark for judging complex classifiers.

Figure 2 shows exploratory calibration against screening-derived outcomes; its small held-out bins cannot establish calibration for clinical ASD. Figure 3 ranks predictors for the score-removed random forests, and Figure 4 presents toddler SHAP contributions. These explanations describe the fitted screening-label models and cannot establish clinical or causal feature significance [29,30].

A false negative for a questionnaire-derived label is not necessarily a missed clinical diagnosis. Real-world assessment would require independently adjudicated outcomes, prospective predictor availability, external validation, clinically selected decision thresholds, and evaluation of net benefit [3,8,9,19–21,25–28].

The high sensitivity of the context-only toddler classifiers should be considered alongside their low specificity. At the prespecified 0.50 threshold, they flagged most screening-positive records yet also misclassified a substantial proportion of screening-negative records. This trade-off limits the usefulness of accuracy or sensitivity considered in isolation. The bootstrap confidence intervals summarize the held-out test partitions and do not establish performance in new clinics or populations.

Study limitations. Neither cohort contains an independently adjudicated clinical ASD diagnosis, so the results quantify agreement with recorded screening classifications. In the context-only toddler models, sensitivity was high (0.945–0.978), but specificity was low (0.134–0.195), indicating that many screening-negative records were classified as positive at the 0.50 threshold. The smaller child cohort, possible selection effects, and the uncertain prospective availability of contextual predictors further limit generalizability. Bootstrap intervals reflect uncertainty within the held-out partitions, not variation across clinical settings. The exact toddler dataset version and redistribution terms remain unconfirmed. Prospective evaluation with independent clinical results is necessary before clinical application.

Computational reproducibility. Re-execution of the analysis script on the downloaded files reproduced the 30 model-result rows, including ROC-AUC estimates and bootstrap intervals. Random seeds and software requirements are provided within the supplementary materials. This establishes repeatability for these files, not external clinical validity. The supplementary audit documents bibliographic source verification and the toddler file's redistribution terms separately.

6. Conclusion

This secondary analysis of 1,054 toddler and 292 child screening records demonstrates that the apparent accuracy of machine-learning classifiers depends strongly on the relationship between the predictors and the recorded outcome. Across both cohorts, models achieved almost perfect discrimination when aggregate screening scores or their constituent questionnaire responses were included. For the random-forest classifier, held-out ROC-AUC decreased from 1.000 using the original feature set to 0.670 in toddlers and 0.567 in children when both aggregate scores and individual screening items were excluded. These findings show that much of the initially high performance reflected reconstructing existing questionnaire-based classifications rather than identifying independent diagnostic information.

The feature-set comparison also shows why excluding a total screening score is not adequate to address target-information leakage. Individual

questionnaire items can retain the information used to construct the outcome, allowing complex algorithms to reproduce a simple scoring rule. The exact correspondence between aggregate-score thresholds and recorded labels in both datasets reinforces this interpretation. Although contextual characteristics provided modest discrimination in the toddler cohort, they performed worse and less consistently in the child cohort. The findings therefore do not support using these models as stand-alone tools to diagnose autism or replace established developmental screening and clinical assessment.

The principal contribution of this study is methodological: it demonstrates how feature-ablation analysis and explicit examination of outcome definitions can distinguish genuine predictive information from circular classification. Subsequent research should evaluate models compared with independently confirmed clinical outcomes, establish whether additional developmental or observational features improve performance beyond existing screening instruments, and assess calibration and clinical usefulness in representative external populations. Until such evidence is available, report machine-learning performance on questionnaire-derived labels as screening-label agreement, and clearly separate its clinical interpretation from diagnostic performance.

References

- American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision. 2022.
- Lord C, Elsabbagh M, Baird G, Veenstra-Vanderweele J. Autism spectrum disorder. *Lancet*. 2018;392:508–520. doi:10.1016/S0140-6736(18)31129-2.
- Hyman SL, Levy SE, Myers SM. Identification, evaluation, and management of children with autism spectrum disorder. *Pediatrics*. 2020;145:e20193447. doi:10.1542/peds.2019-3447.
- Robins DL, et al. Validation of the Modified Checklist for Autism in Toddlers, Revised with Follow-Up. *Pediatrics*. 2014;133:37–45. doi:10.1542/peds.2013-1813.
- Duda M, Ma R, Haber N, Wall DP. Use of machine learning for behavioral distinction of autism and ADHD. *Transl Psychiatry*. 2016;6:e732. doi:10.1038/tp.2015.221.
- Thabtah F. Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment. In: *Proceedings of the 1st International Conference on Medical and Health Informatics (ICMHI 2017)*. ACM; 2017:1–6. doi:10.1145/3107514.3107515.
- Thabtah F. Autistic Spectrum Disorder Screening Data for Children. UCI Machine Learning Repository. 2017. doi:10.24432/C5659W. <https://archive.ics.uci.edu/dataset/419/autistic+spectrum+disorder+screening+data+for+children>
- Collins GS, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. doi:10.1136/bmj-2023-078378.

- Moons KGM, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. *BMJ*. 2025;388:e082505. doi:10.1136/bmj-2024-082505.
- Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*. 2023;4(9):100804. doi:10.1016/j.patter.2023.100804.
- Wall DP, Kosmicki J, DeLuca TF, Harstad E, Fusaro VA. Use of machine learning to shorten observation-based screening and diagnosis of autism. *Transl Psychiatry*. 2012;2:e100. doi:10.1038/tp.2012.10.
- Wall DP, Dally R, Luyster R, Jung JY, DeLuca TF. Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PLoS One*. 2012;7(8):e43855. doi:10.1371/journal.pone.0043855.
- Guthrie W, et al. Accuracy of autism screening in a large pediatric network. *Pediatrics*. 2019;144(4):e20183963. doi:10.1542/peds.2018-3963.
- Wallis KE, Guthrie W. Screening for autism: a review of the current state, persisting challenges, and novel approaches on the horizon. *Pediatr Clin North Am*. 2024;71(2):127–155. doi:10.1016/j.pcl.2023.12.003.
- Abbas H, Garberson F, Glover E, Wall DP. Machine learning method for early detection of autism by combining questionnaire and home video screening. *J Am Med Inform Assoc*. 2018;25(8):1000–1007. doi:10.1093/jamia/ocy039.
- Tariq Q, Daniels J, Schwartz JN, Washington P, Kalantarian H, Wall DP. Mobile detection of autism through machine learning on home video: a development and prospective validation study. *PLoS Med*. 2018. doi:10.1371/journal.pmed.1002705.
- Ding Y, Zhang H, Qiu T. Deep learning approach to predict autism spectrum disorder: a systematic review and meta-analysis. *BMC Psychiatry*. 2024;24:739. doi:10.1186/s12888-024-06116-0.
- Huda E, Hawker P, Cibralic S, et al. Screening tools for autism in culturally and linguistically diverse child populations: a systematic review. *BMC Pediatr*. 2024;24:610. doi:10.1186/s12887-024-05067-5.
- Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6). doi:10.1177/0272989X06295361.
- Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res*. 2019;3:18. doi:10.1186/s41512-019-0064-7.
- Vickers AJ, van Calster B, Wynants L, et al. Decision curve analysis: confidence intervals and hypothesis testing for net benefit—Diagn Progn Res. 2023;7:11. doi:10.1186/s41512-023-00148-y.
- Wallis KE, Nekrasova E, Bennett AE, et al. Autism spectrum disorder screening during the COVID-19 pandemic in a large primary care network. *Acad Pediatr*. 2022;22(8):1384–1389. doi:10.1016/j.acap.2022.04.005.
- Huda S, Khan DM, Masroor K, et al. Progress in automated diagnosis of autism spectrum disorder through deep learning and resting-state functional MRI biomarkers: a systematic review. *Cogn Neurodyn*. 2024;18(6):3585–3601. doi:10.1007/s11571-024-10176-z.
- Screening and diagnostic tools for autism spectrum disorder: systematic review and meta-analysis. *Clinics*. 2024;79:100323. doi:10.1016/j.clinsp.2023.100323.
- Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21:128–138. doi:10.1097/EDE.0b013e3181c30fb2.
- Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. *BMC Medicine*. 2019;17:230. doi:10.1186/s12916-019-1466-7.
- Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441. doi:10.1136/bmj.m441.
- Wolff RF, Moons KGM, Riley RD, et al. PROBAST: A Tool to Assess the Risk of Bias and Usability of Prediction Model Studies. *Ann Intern Med*. 2019;170:51–58. doi:10.7326/M18-1376.
- Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Estimates. *Advances in Neural Information Processing Systems*. 2017;30.
- Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*. 2020;2:56–67. doi:10.1038/s42256-019-0138-9.
- Breiman L. Random Forests. *Machine Learning*. 2001;45:5–32. doi:10.1023/A:1010933404324.
- Cortes C, Vapnik V. Support-vector networks. *Machine Learning*. 1995;20:273–297. doi:10.1007/BF00994018.
- Friedman JH. Greedy function estimation: a gradient boosting machine. *Annals of Statistics*. 2001;29:1189–1232. doi:10.1214/aos/1013203451.
- Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proceedings of KDD*. 2016:785–794. doi:10.1145/2939672.2939785.
- Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
- Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Analysis*. 2nd ed. Springer; 2009. doi:10.1007/978-0-387-84858-7.
- Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall; 1993. doi:10.1007/978-1-4899-4541-9.
- Niculescu-Mizil A, Caruana R. Predicting Good Probabilities With Supervised Learning. *Proceedings of ICML*. 2005:625–632. doi:10.1145/1102351.1102430.
- Robins DL, Fein D, Barton ML, Green JA. The Modified Checklist for Autism in Toddlers: an initial study investigating the early detection of autism and pervasive developmental disorders. *J Autism Dev Disord*. 2001;31:131–144. doi:10.1023/A:1010738829569.
- Zwaigenbaum L, Bauman ML, Choueiri R, et al. Early Identification and Interventions for Autism Spectrum Disorder: Executive Summary. *Pediatrics*. 2015;136(Suppl 1):S1–S9. doi:10.1542/peds.2014-3667B.