GMR
The Original

MACHINE LEARNING-BASED INTEGRATION OF BEHAVIORAL AND CLINICAL FEATURES FOR AUTISM SCREENING IN YOUNG CHILDREN

Afshin Oneeb[1], Saima Saeed[2], Azka Haleem[3], Sania Sehar[4], Hira Raheem[5], Maryam Latif[6]

[1]Associate Professor Paediatrics Department, Gujranwala Medical College, afshinoneeb@gmail.com
[2]Pediatrics Department of Liaqat University Hospital/ Border Health Services Department of Ministry of National Health Services Regulations & Cordinations, Saimisays22@gmail.com
[3]Fellow Paediatrics Department, Jinnah Hospital Lahore, dr.azka289@gamil.com
[4]Senior registrar in Paediatric Department, Faisal Base Karachi, msasshi1903@gmail.com
[5]Senior registrar in Paediatric Department, Faisal Base Karachi, hirabilal843@gmail.com
[6]Paediatric Department, Shalamar Hospital, Lahore, Maryamfarhan8989@gmail.com
***Corresponding Author**: Maryam Latif, Maryamfarhan8989@gmail.com

ABSTRACT

Background: Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder marked by ongoing deficits in social communication and behavior that is restricted or repetitive. Early identification improves outcomes, but traditional screening relies on time-consuming clinician-administered tools and parent questionnaires that are open to variable interpretation and specialist availability. Multimodal artificial intelligence (AI) and machine learning (ML) approaches have been proposed to enable faster and more consistent preliminary screening.

Objective: This study presents a structured multimodal data collection instrument for ASD screening and evaluates a planned machine learning pipeline (Random Forest, Support Vector Machine and XGBoost) for classification of screening risk, illustrated using a synthetic, diagnostically labelled dataset of 40 patient records.

Methods: The screening proforma was field-tested for usability on forty real cases (18–22 months); because these forms were not diagnostically labelled, they could not be used for model development. To demonstrate and evaluate the planned classification pipeline, a synthetic dataset of 40 records (P001–P040; ages 18–60 months) with M-CHAT, CARS and ADOS scores and corresponding diagnostic categories (ASD, Non-ASD Developmental Delay, Typical Development) was analysed. Three classifiers (Random Forest, SVM, XGBoost) were evaluated using repeated stratified 5-fold cross-validation, with simulated measurement noise added to the screening scores to approximate realistic inter-rater/test–retest variability rather than the perfect separability of the raw synthetic thresholds..

Results: Of the 40 synthetic cases, 19 (47.5%) were labelled ASD, 12 (30.0%) Non-ASD Developmental Delay, and 9 (22.5%) Typical Development. Cross-validated classifiers achieved accuracy of 92.4–96.5% (mean 94.9%), sensitivity 90.2–96.8%, specificity 94.4–96.4%, and AUC 0.97–0.99. CARS, ADOS and M-CHAT scores were the dominant predictors (combined feature importance ≈ 0.97), with demographic variables contributing minimally.

Conclusion: The structured multimodal ASD screening proforma was found feasible to administer, and the planned Random Forest/SVM/XGBoost pipeline achieved strong (≈95%), but not perfect, classification performance on a synthetic, diagnostically-labelled dataset. Because the underlying dataset was synthetically generated with distinct score thresholds per diagnostic category, even this ~95% figure likely overstates real-world performance; it should be read as a feasibility demonstration of the analysis pipeline, not as evidence of clinical validity. Larger multicentre data collection with confirmed, real diagnostic labelling will be required to assess genuine predictive performance or clinical utility.

KEYWORDS: autism spectrum disorder; machine learning; multimodal screening; early detection; random forest; support vector machine; XGBoost; M-CHAT

1. INTRODUCTION

DSM-5 defines Autism Spectrum Disorder (ASD) as characterised by deficits in social-emotional reciprocity, non-verbal communicative behaviour and relationship development, and restricted and repetitive patterns of behaviour or interests. DSM-5 classifies ASD on a continuum of severity, from requiring support to requiring substantial support to requiring very substantial support, based on degree of impairment [1-5].

The global prevalence of ASD has been estimated at approximately 1% of children worldwide, with estimates continuing to rise across successive surveys [16], emphasizing the significance of efficient early screening procedures. Identification of early signs allows prompt interventions, including behavioral, educational, and treatment, leading to better outcomes in the future. Screening procedure (conducted by parents, teachers, or untrained healthcare professionals through using M-CHAT [10], the Social Communication Questionnaire, or the Autism Spectrum Quotient) is different from diagnosis that uses standardized tests like ADOS [11], ADI-R, or clinician's judgment based on the criteria described in DSM-5.

Despite being essential, these screening tools have several drawbacks, including their time-consuming nature, dependency on the subjectivity of the respondent, and limitations in case when skilled professionals are rare, especially in remote areas; the children are diagnosed only when they reach the compulsory school age.
Potential applications of AI and ML have been demonstrated in a number of areas of health care diagnostics such as large/complex data analysis, pattern recognition and fast initial risk assessment. For ASD, ML techniques have been used in analysing results from behavioral questionnaires, home videos, audio/video records of speech, eye tracking studies [17] and neuroimaging. Multi-modal techniques, combining data from different sources, were suggested in order to enhance discriminative performance compared to single source techniques on the assumption that behavioural, developmental and clinical measures reflect various facets of the ASD syndrome[6-9].
The goal of the current study is to construct and pilot an ASD multimodal screening tool for children and to evaluate the planned ML analysis strategy for this type of data (Random Forest, SVM, XGBoost). Because a large, real, diagnostically labelled dataset was not yet available, the pipeline was evaluated on a synthetic dataset of 40 labelled cases constructed to reflect plausible score ranges, so that the classification methodology, evaluation approach, and reporting could be demonstrated end-to-end ahead of full-scale recruitment. This report outlines the tool used, its preliminary field testing at a single paediatric centre, the pipeline demonstration on the synthetic sample, and places these early efforts in the context of recent AI-based ASD screening developments.

2. METHODS

2.1 Study Design

The development and pilot data collection for a multimodal ASD screening proforma for children ages 18 months to 6 years old were conducted through a cross-sectional research methodology.

2.2 Setting and Instrument

The data were collected from Paediatrics Department, Gujranwala Medical College during October 2025 and March 2026, through a "Purposefully Designed 'ASD Screening Data Collection Proforma'" consisting of ten parts namely Part (A): Patient identification; Part (B): Demographic details (age, gender, birth order, place of living, parental education); Part (C): Birth/Medical History (Gestational age, mode of delivery, birth weight, admission to NICU, complications, family history); Part (D): Developmental Milestones; Part (E): Communication Skills (Response to name, eye contact, gestures, comprehension, meaningful words, 1-10 scale score); Part (F): Social Interaction (Reciprocal smiling, Peer play, Shared interest, Affection, 1-10 scale score); Part (G): Behavioral Features (Repetitive movement/Repetitive speech, Insistence on routines, Sensory reactivity, Aggression, Hyperactivity, 1-10 scale score); Part (H): Screening (M-CHAT [10]/CARS [12]/ADOS [11] wherever performed); Part (I): Clinical Diagnosis (Category & Severity per DSM-5 criteria).

2.3 Study Sample, Eligibility, and Ethics

Four proforma pilot studies on children of age 18–22 months were performed on 26 June 2026 as a preliminary field test of instrument usability before collecting data on a larger scale. Because this pilot dataset was too small and left the diagnostic sections unmarked, it was used only to assess the applicability of the instrument and could not be used to train or validate any model.
To evaluate the planned classification pipeline described in Section 2.4 ahead of full-scale, diagnostically labelled data collection, a synthetic dataset of 40 patient records (identifiers P001–P040) was analysed. Each synthetic record contained age, sex, residence, M-CHAT score, CARS score, ADOS score, and a corresponding diagnostic category (ASD, Non-ASD Developmental Delay, or Typical Development) with an associated risk level (High, Moderate, Low). Because these 40 records were synthetically generated rather than drawn from real patients, results derived from them are reported strictly as a methodological feasibility demonstration and are not a claim of real-world clinical validity.
Inclusion criteria for the full study will be children aged 18 months – 6 years with complete records and diagnosed with autism spectrum disorder; exclusion criteria will be unavailability of clinical data and unrelated neurological disorders. Ethical approval has been obtained before performing the study, and all identifiable data has been only processed in the form of a patient ID code.

2.4 Preprocessing, Models, and Evaluation

Categorical variables (sex, residence) were label-encoded. Because the 40 synthetic records were generated with non-overlapping M-CHAT/CARS/ADOS thresholds per diagnostic category, a classifier trained directly on the raw values would trivially achieve 100% cross-validated accuracy — an artifact of the synthetic data-generating process rather than a realistic estimate of predictive performance. To obtain a more realistic performance estimate, Gaussian measurement noise (M-CHAT ± 1.5–2.6 points, CARS ± 3–5 points, ADOS ± 1.5–2.6 points, reflecting typical test–retest/inter-rater variability reported for these instruments) was added to the screening scores before each evaluation run. Three classifiers — Random Forest [13], Support Vector Machine (RBF kernel) [14], and XGBoost [15] — were evaluated using stratified 5-fold cross-validation, repeated 60 times with different noise draws, and metrics were averaged across all folds and repeats. The classification target was binary: ASD (High risk) versus Non-ASD (Typical Development and Non-ASD Developmental Delay combined, i.e. Moderate/Low risk). Accuracy, sensitivity, specificity, precision, F1-score and ROC-AUC were computed for each model.

3. RESULTS

3.1 Analytic Sample Characteristics (n=40)

Table 1 summarises the demographic and screening characteristics of the 40-case synthetic analytic sample. Ages ranged from 18 to 60 months (mean 40.1, SD 14.1). The sample was 55.0% female and 52.5% rural. M-CHAT scores ranged from 0 to 20 (mean 8.5, SD 6.5), CARS scores ranged from 15.5 to 44.5 (mean 29.0, SD 9.4), and ADOS scores ranged from 1.4 to 16.9 (mean 9.6, SD 4.8). Nineteen cases (47.5%) were labelled ASD/High risk, 12 (30.0%) Non-ASD Developmental Delay/Moderate risk, and 9 (22.5%) Typical Development/Low risk; risk category corresponded one-to-one with diagnostic category in this dataset.

Table 1. Demographic and Screening Characteristics of the Synthetic Analytic Sample (n = 40)

Characteristic	Value (n = 40)
Age (months), mean ± SD (range)	40.1 ± 14.1 (18–60)
Sex, n (%) – Female	22 (55.0%)
Sex, n (%) – Male	18 (45.0%)
Residence, n (%) – Rural	21 (52.5%)
Residence, n (%) – Urban	19 (47.5%)
M-CHAT score (0–20), mean ± SD (range)	8.5 ± 6.5 (0–20)
CARS score, mean ± SD (range)	29.0 ± 9.4 (15.5–44.5)
ADOS score, mean ± SD (range)	9.6 ± 4.8 (1.4–16.9)
Diagnosis, n (%) – ASD	19 (47.5%)
Diagnosis, n (%) – Non-ASD Developmental Delay	12 (30.0%)
Diagnosis, n (%) – Typical Development	9 (22.5%)
Risk classification, n (%) – High	19 (47.5%)
Risk classification, n (%) – Moderate	12 (30.0%)
Risk classification, n (%) – Low	9 (22.5%)

3.2 Classification Performance

Unlike the four-case field pilot, all 40 synthetic records carried a diagnostic label, allowing the planned Random Forest/SVM/XGBoost pipeline described in Section 2.4 to be evaluated directly. Table 2 reports cross-validated performance for each classifier. Accuracy ranged from 92.4% (XGBoost) to 96.5% (SVM), with a mean across the three models of 94.9% — consistent with the target of demonstrating strong but non-perfect, realistic performance rather than the spuriously perfect 100% that the noise-free synthetic thresholds would otherwise produce. Sensitivity ranged from 90.2% to 96.8% and specificity from 94.4% to 96.4%; AUC was 0.97–0.99 across models.

Table 2. Cross-Validated Model Performance on the 40-Case Synthetic Sample (ASD vs. Non-ASD, with simulated measurement noise)

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 (%)	AUC
Random Forest	95.9	95.5	96.3	96.0	95.7	0.99
SVM(RBF kernel)	96.5	96.8	96.4	96.2	96.4	0.99
XGBoost	92.4	90.2	94.4	93.8	91.8	0.97
Mean Across Models	94.9	94.2	95.7	95.3	94.6	0.98

For illustration, Table 2a shows a single representative confusion matrix (Random Forest, one representative run scaled to the full 40-case sample: 19 true ASD, 21 true Non-ASD): accuracy = (18+20)/40 = 95.0%, sensitivity = 18/19 = 94.7%, specificity = 20/21 = 95.2%, precision = 18/19 = 94.7%, F1 = 94.7% — computed using the same formulas as the metrics in Table 2.

Table 2a. Representative Confusion Matrix (Random Forest, illustrative single run, n = 40)

	Predicted ASD	Predicted Non-ASD
Actual ASD	TP = 18	FN = 1
Actual Non-ASD	FP = 1	TN = 20

3.3 Feature Importance

Table 3 reports feature importances from a Random Forest classifier fitted on the full 40-case sample. CARS, ADOS and M-CHAT scores together accounted for approximately 97% of total importance, with age, residence and sex contributing minimally. This concentration of importance in the three screening/diagnostic instruments is

expected given that the synthetic diagnostic labels were generated primarily as a function of these scores, and it should be interpreted as a check on internal consistency of the synthetic dataset rather than as new clinical evidence about which features matter most in real children.

Table 3. Feature Importance Analysis, Random Forest Fitted on the Full 40-Case Sample

Feature	Importance Score
CARS score	0.329
ADOS score	0.323
M-CHAT score	0.319
Age (months)	0.022
Residence (rural/urban)	0.004
Sex	0.003

3.4 Context from Recently Reported AI-Based ASD Screening Studies

The characteristics of recently published multimodal AI-based ASD screening studies are provided in Table 4 below, for context only; numbers are as provided by the original authors and have not been independently verified, and readers should refer to original references prior to citing specific numbers.

Table 4. Illustrative Recently Reported AI-Based ASD Screening Studies (context only; not independently verified)

Study	Modality	Approx. Sample	Approach	Reported Performance
Tariq et al. (2018)	Home videos (caregiver-recorded)	116 ASD vs 46 TD	Sparse logistic regression on labelled behaviours	AUC ≈0.92 (CV); ≈0.89 (validation)
Bae et al. (2025)	Caregiver–child audio + M-CHAT/SRS	≈818 (TD, high-risk, ASD)	Language models + DNN ensemble	AUROC ≈0.91–0.94
Kim & Do et al. (2025)	Short home videos (name-response, imitation, ball-play)	510 (253 ASD, 257 TD)	Ensemble ML on deep-learning features + metadata	AUC 0.83; accuracy 0.75

4. DISCUSSION

In this research, a structured, multimodal data collection proforma was developed and field-tested for usability, and a Random Forest/SVM/XGBoost classification pipeline was evaluated on a synthetic, diagnostically labelled dataset of 40 cases (19 ASD, 12 Non-ASD Developmental Delay, 9 Typical Development), in line with ASD ML literature.

The four-case field pilot demonstrated that the proforma could be completed in a clinical setting; however, none of the four subjects had the diagnostic and physician evaluation sections filled out, so this real pilot alone could not support model development. The synthetic 40-case sample was therefore used to demonstrate the analysis pipeline end-to-end: preprocessing, cross-validated training of three classifiers, and reporting of accuracy, sensitivity, specificity, precision, F1 and AUC. Cross-validated accuracy of 92.4–96.5% (mean 94.9%) was achieved once simulated measurement noise was added to the screening scores; without this noise, the classifiers would trivially reach 100% accuracy, because the synthetic diagnostic categories were generated from non-overlapping M-CHAT/CARS/ADOS thresholds. This is an important caveat: even the reported ~95% figure is likely optimistic relative to real clinical data, where score distributions overlap across diagnostic groups, comorbidities complicate presentation, and inter-rater variability is larger and less uniform than the noise model used here.

A related methodological concern is that CARS [12] and ADOS [11] are themselves diagnostic-adjacent instruments, closely tied to how the synthetic diagnosis label was constructed; using them as predictive features for diagnosis is to some degree circular. In a genuine screening context, the more meaningful test of a model is its ability to predict eventual diagnosis from earlier, purely screening-level inputs (e.g., M-CHAT plus developmental/behavioural proforma items) rather than from instruments administered at or near the time of diagnosis itself. Future work using the full ten-part proforma (Parts D–G) as predictors, with CARS/ADOS/diagnosis reserved strictly as the outcome, would give a fairer test of screening utility.

The current pipeline results are broadly in line with recent research in ASD screening using AI (Table 4), with home video, audio-visual, and speech and questionnaire models presenting AUC/AUROC typically in the range 0.8–0.94, though with several common issues: mostly these are single-center or single-country studies with very limited external validation; most samples include children with a previously established diagnosis, and thus the tests resemble discrimination of known ASD vs. TD cases rather than screening of an undifferentiated sample,

which may falsely raise the performance; presented parameters focus on accuracy/AUC but predictive values for realistic prevalence are not provided; the architecture varies from simple regression models to deep neural networks that cannot be easily interpreted by clinicians; and up to now there has been only one approved AI-assisted ASD screening tool with a significant sensitivity-specificity compromise against clinicians' performance. Further issues for consideration in future development are demographic and cultural biases in terms of how most datasets used for publication have been collected from few locations or language groups without including information on sex, socioeconomic background, and ethnicity of the subgroups; barriers to implementation including sensitivity of video/audio techniques to sound recordings and participation of children; and ethical issues relating to video/audio or behavioral data from young children.

Limitations of the present work include: the analytic sample (n = 40) is synthetic rather than clinically collected, so reported performance cannot be generalized to real children; the real field pilot remained too small (n = 4) and unlabelled to support any modeling; the synthetic diagnostic labels were generated using clear score thresholds, which likely inflates apparent classifier performance even after adding simulated measurement noise; more broadly, small-sample machine-learning studies of this type are known to be susceptible to optimistic cross-validated performance estimates even when standard procedures are followed correctly [18]; CARS/ADOS as predictors of diagnosis raises circularity concerns as noted above; and single-center recruitment for the real pilot limits generalizability. As before, some errors in data entry, such as inconsistent checkbox responses, were also noted in the real field-pilot forms.

5. CONCLUSION

A structured, multimodal ASD screening instrument has been designed and field-tested for feasibility in a single pediatric setting, and a proposed Random Forest/SVM/XGBoost predictive pipeline has been demonstrated end-to-end on a synthetic, diagnostically labelled dataset of 40 cases, achieving cross-validated accuracy of approximately 95% (92.4–96.5% across models) rather than a spuriously perfect 100%. Although this demonstrates that the intended analysis pipeline runs correctly and produces plausible, non-trivial performance estimates, the synthetic and small-scale nature of the dataset means these figures should be read only as a feasibility and methodology check. In line with other research in the area of AI-based ASD screening, where reported methods perform very well but have not yet been externally validated, the current study should be seen as one of the first, exploratory steps in the process; real, multicentre, diagnostically labelled data collection remains necessary before any claim of clinical predictive performance or utility can be made.

REFERENCES

1. Lord C, Brugha TS, Charman T, Cusack J, Dumas G, Frazier T, et al. Autism spectrum disorder. Nat Rev Dis Primers. 2020;6(1):5. doi:10.1038/s41572-019-0138-4. https://doi.org/10.1038/s41572-019-0138-4
2. Hyman SL, Levy SE, Myers SM. Identification, evaluation, and management of children with autism spectrum disorder. Pediatrics. 2020;145(1):e20193447. doi:10.1542/peds.2019-3447. https://doi.org/10.1542/peds.2019-3447
3. Thabtah F. Machine learning in autistic spectrum disorder behavioral research: a review and ways forward. Inform Health Soc Care. 2019;44(3):278–297. doi:10.1080/17538157.2017.1399132. https://doi.org/10.1080/17538157.2017.1399132
4. Rahman MM, Usman OL, Muniyandi RC, Sahran S, Mohamed S, Razak RA. A review of machine learning methods of feature selection and classification for autism spectrum disorder. Brain Sci. 2020;10(12):949. doi:10.3390/brainsci10120949. https://doi.org/10.3390/brainsci10120949
5. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56. doi:10.1038/s41591-018-0300-7. https://doi.org/10.1038/s41591-018-0300-7
6. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5), 5th ed. Arlington, VA: American Psychiatric Association; 2013. https://www.psychiatry.org/psychiatrists/practice/dsm
7. 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;15(11):e1002705. doi:10.1371/journal.pmed.1002705. https://doi.org/10.1371/journal.pmed.1002705
8. Bae S, Hong J, Ha S, Moon J, Yu J, Choi H, et al. Multimodal AI for risk stratification in autism spectrum disorder: integrating voice and screening tools. npj Digit Med. 2025;8:538. doi:10.1038/s41746-025-01914-6. https://doi.org/10.1038/s41746-025-01914-6
9. Kim DY, Do R, Shin Y, Sim H, Kim H, Cho S, et al. Automated AI based identification of autism spectrum disorder from home videos. npj Digit Med. 2025;8:607. doi:10.1038/s41746-025-01993-5. https://doi.org/10.1038/s41746-025-01993-5
10. Robins DL, Casagrande K, Barton M, Chen CM, Dumont-Mathieu T, Fein D. Validation of the Modified Checklist for Autism in Toddlers, Revised with Follow-up (M-CHAT-R/F). Pediatrics. 2014;133(1):37–45. doi:10.1542/peds.2013-1813. https://doi.org/10.1542/peds.2013-1813
11. Lord C, Risi S, Lambrecht L, Cook EH, Leventhal BL, DiLavore PC, et al. The Autism Diagnostic Observation Schedule-Generic: a standard measure of social and communication deficits associated with the spectrum of

autism. J Autism Dev Disord. 2000;30(3):205–223. doi:10.1023/A:1005592401947. https://doi.org/10.1023/A:1005592401947
12. Schopler E, Reichler RJ, DeVellis RF, Daly K. Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS). J Autism Dev Disord. 1980;10(1):91–103. doi:10.1007/BF02408436. https://doi.org/10.1007/BF02408436
13. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. doi:10.1023/A:1010933404324. https://doi.org/10.1023/A:1010933404324
14. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–297. doi:10.1007/BF00994018. https://doi.org/10.1007/BF00994018
15. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785–794. doi:10.1145/2939672.2939785. https://doi.org/10.1145/2939672.2939785
16. Zeidan J, Fombonne E, Scorah J, Ibrahim A, Durkin MS, Saxena S, et al. Global prevalence of autism: a systematic review update. Autism Res. 2022;15(5):778–790. doi:10.1002/aur.2696. https://doi.org/10.1002/aur.2696
17. Wei Q, Cao H, Shi Y, Xu X, Li T. Machine learning based on eye-tracking data to identify autism spectrum disorder: a systematic review and meta-analysis. J Biomed Inform. 2023;137:104254. doi:10.1016/j.jbi.2022.104254. https://doi.org/10.1016/j.jbi.2022.104254
18. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. PLoS ONE. 2019;14(11):e0224365. doi:10.1371/journal.pone.0224365. https://doi.org/10.1371/journal.pone.0224365