Early screening of autism spectrum disorder in toddlers using machine learning

Darin Mansor Mathkor*

Department of Nursing, College of Nursing and Health Sciences, Jazan University, Jazan, Saudi Arabia

Submitted: 9 January 2026
Accepted: 24 February 2026
Online publication: ?????????????

Arch Med Sci
DOI: https://doi.org/10.5114/aoms/218492

***Corresponding author:**
Dr. Darin Mansor Mathkor
Department of Nursing
College of Nursing and Health Sciences
Jazan University
Jazan-82911, Saudi Arabia
~~Phone: +966 0 173295000 extn. 5678~~
E-mail: darin.mathkor@gmail.com,
dmathkor@jazanu.edu.sa

Abstract

Introduction: Autism spectrum disorder (ASD) is a developmental brain condition that causes problems with social interaction and communication, as well as repetitive behaviors. Early ASD screening is vital for prompt actions, as existing diagnostic methods suffer from scalability limitations in resource-constrained settings. Herein, a machine learning (ML)-based explainable predictive model was developed and evaluated for early screening of ASD in toddlers using Q-CHAT-10 behavioral, demographic, and clinical features.
Material and methods: A total of 1054 toddlers' records sourced from the Autism Screening for Toddlers dataset freely available at Kaggle were retrospectively analyzed. Data preprocessing, statistical feature selection, and dimensionality reduction were performed. Multiple ensemble models were trained using Q-CHAT-10 behavioral features combined with demographic and clinical variables. Several algorithms were tested, including Logistic Regression, Random Forest, Gradient Boosting, and Multilayer Perceptron, with k-fold cross-validation for model selection. SHAP analysis was employed to explore the reasons behind individual predictions.
Results: Model performance was evaluated using ROC-AUC. Feature importance was checked to identify the most predictive items. The Gradient Boosting classifier achieved the best performance, with an accuracy of 0.98 (95% CI: 0.85–0.93), sensitivity of 0.91, specificity of 0.87, and ROC-AUC of 0.94 on the held-out test set. SHAP analysis revealed total Q-CHAT-10 score, response to name, pointing to share interest, and pretend play as the most influential predictors.
Conclusions: This ML framework accurately detects ASD traits in toddlers, highlighting the potential of a scalable, low-cost screening tool to enable early ASD detection and improve equitable access to pediatric care. However, external validation across diverse populations with larger samples is warranted before clinical application can be recommended.

Key words: autism spectrum disorder, machine learning, Q-CHAT-10, early screening, predictive modeling, explainable AI, toddlers

Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by difficulties in social interactions and communication, as well as repetitive behaviors [1, 2]. Early detection of the presence of autism traits is important because it leads to timely interventions that significantly improve developmental outcomes [3]. Although effective, traditional methods such as the Autism Diagnostic Observation Schedule

(ADOS) often require excessive time, resources, and highly trained professionals, making them unsuitable for massive or even early-stage screening. Based on ADOS studies as well as similar paradigms, the main criticism was that they strictly use the analysis provided by professionals and therefore tend to be accurate but unscaled [4, 5]. Also, many studies suffer from a low-dimensional feature space, where significant behavioral or demographic variables have been overlooked, thereby reducing their potency in terms of prediction.

This study acknowledges the role of machine learning (ML) in developing an efficient, scalable, and data-driven approach towards autism screening [6–9]. Recent studies have used ML for autism screening by leveraging questionnaire-driven behavioral traits, optimizing using data-driven methods, and employing unsupervised or DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition) aligned approaches. Recent studies based on Q-CHAT (Quantitative Checklist for Autism in Toddlers; a unidimensional measure of autistic traits (risk) in toddlers) scores demonstrate robust predictive capabilities [10].

The current research investigated various learning paradigms on Q-CHAT-10 items along with demographic covariates, and employed explainable artificial intelligence (AI) [SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations)] for data interpretation. This study aimed to develop an ML-based framework for early detection of autism in toddlers. Using the Q-Chat-10-Toddler screening method and other demographic and clinical variables, the predictors of the most significant factors of autism traits were identified in the present study. The Modified Checklist for Autism in Toddlers, Revised with Follow-Up (M-CHAT-R/F) is the most widely used tool for screening [11]. This highlights the need for more scalable, data-driven screening approaches, such as ML, which may enhance early detection and intervention for autism [12]. By combining robust data analysis with ML, findings are made towards the development of scalable, accurate, and practical screening tool(s), which could open up early autism screening for a wider, more heterogeneous population. The objectives of the present study were to evaluate the predictive ability of Q-Chat-10 items for early autism detection, to examine whether demographic and clinical factors enhance prediction, and finally, to compare the performance of multiple machine learning models. The novelty of this work lies in the integration of behavioral, demographic, and clinical features into a single predictive framework, as previous ML-based Q-CHAT/Q-CHAT-10 screening research has generally concentrated solely on behavioral items or alternative modalities; in contrast, the current study assesses various learning paradigms within a cohesive cross-validated framework and presents both ROC and precision-recall performance in light of the dataset's class imbalance. Furthermore, this study offers both global and instance-level explainability via SHAP and LIME analyses to enhance the interpretability of model outputs in a screening context.

Material and methods

The present study followed a secondary data analysis design, using a cross-sectional ML framework to investigate autism traits in toddlers by predicting autism classifications and identifying the most influential behavioral, demographic, and clinical predictors. The overall workflow (Figure 1) progressed through clearly defined stages of data acquisition, preprocessing, exploratory data analysis, statistical testing, feature selection, model development, performance evaluation, and lastly explainable AI interpretation. The study utilized the Kaggle toddler-autism dataset and applied structured preprocessing steps, including the encoding of categorical variables and normalization of numerical attributes. Exploratory analyses examined score distributions, age-related trends, and inter-variable correlations. Feature relevance was established using statistical tests and dimensionality reduction techniques. Multiple ML models were developed with hyperparameter optimization and cross-validation, focusing on addressing class imbalance. Finally, feature importance and interpretability were derived using model-based rankings, SHAP values, and LIME explanations, for deeper insight into the drivers of autism trait prediction.

Data collection and data preprocessing

A structured approach was used to examine autism traits among toddlers based on ML methods. The Autism Screening for Toddlers dataset was derived from an online autism screening questionnaire, through which parents and caregivers voluntarily submitted their responses to the Q-CHAT-10-Toddler tool via a publicly available application. It is available on Kaggle under an open license (repository found at Kaggle Dataset: https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers), thereby conforming to the requirements of ethics and privacy standards for a secondary analysis. In this dataset, the class label ("Class/ASD Traits") was generated using the ASDTests screening tool based on the total score of the Q-CHAT-10 questionnaire. Toddlers with a Q-CHAT-10 score greater 3 were classified as having "ASD traits," while those with a score of 3 or less were classified as "No ASD traits". Thus, the target outcome reflects ASD traits identified through a screening instrument rather than a clinical diagnosis of ASD confirmed by a clinician. The survey questions (behavioral items) from A1 to A10 are given in Table I. Although the dataset does

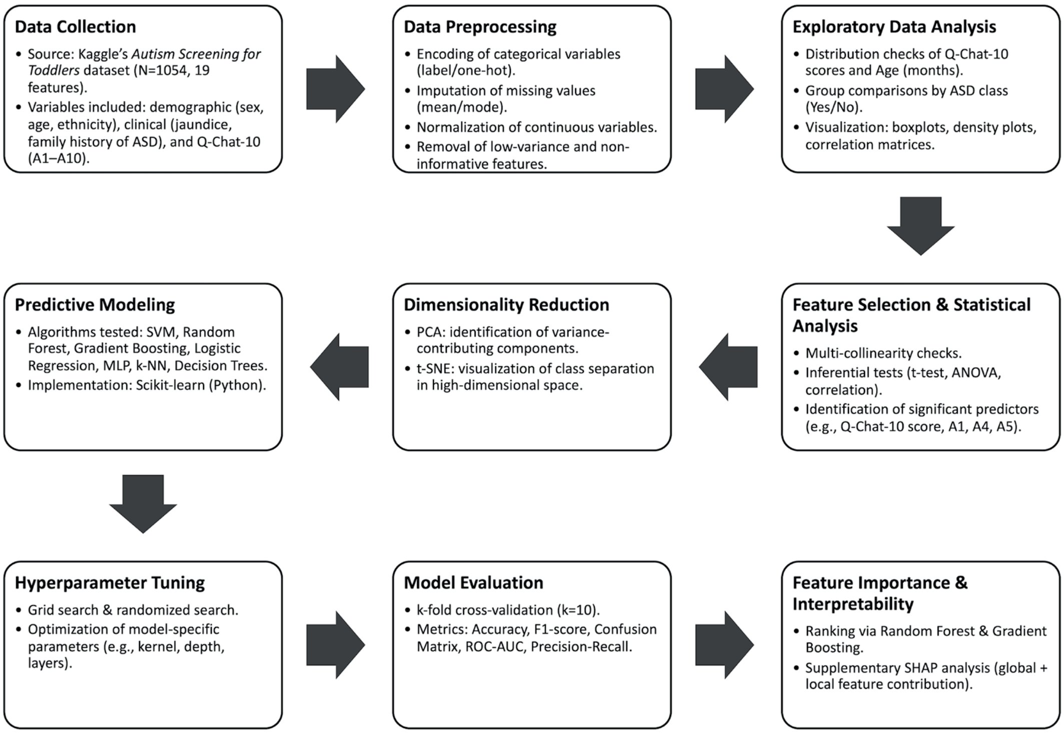

Figure 1. Schematic workflow of the methodology adopted during the study

Table I. Details of variables mapping to the Q-CHAT-10 screening method

Variable in dataset	Corresponding Q-CHAT-10-Toddler features	Description
A1	Does your child look at you when you call his/her name?	(Yes or no)
A2	How easy is it for you to make eye contact with your child?	(Yes or no)
A3	Does your child point to indicate that s/he wants something? (e.g., a toy that is out of reach)	(Yes or no)
A4	Does your child point to share interests with you? (e.g., pointing at an interesting sight)	(Yes or no)
A5	Does your child pretend? (e.g., care for dolls, talk on a toy phone)	(Yes or no)
A6	Does your child follow where you're looking?	(Yes or no)
A7	If you or someone else in the family is visibly upset, does your child show signs of wanting to comfort them? (e.g., stroking hair, hugging them)	(Yes or no)
A8	Would you describe your child's first words as:	(Yes or no)
A9	Does your child use simple gestures? (e.g., wave goodbye)	(Yes or no)
A10	Does your child stare at nothing with no apparent purpose?	(Yes or no)
Age	Toddlers (months)	
Score by Q-CHAT-10	1–10 (less than or equal to 3 No ASD traits; > 3 ASD traits	
Sex	Male or female	
Ethnicity	List of common ethnicities in text format	
Born with jaundice	Whether the child was born with jaundice	
Family member with ASD history	Whether any immediate family member has a pervasive developmental disorder	
Who is completing the test?	Parent, self, caregiver, medical staff, clinician, etc.	
Why are you taking the screening?	Short answer showing intention of self-care	
Class variable	ASD traits or No ASD traits (automatically assigned by the ASD Tests app). (Yes/No)	

not specify the exact years or geographic regions of data collection, it represents a community-collected sample produced via open participation rather than structured epidemiological sampling. Consequently, the dataset reflects a mixed population containing toddlers both with and without ASD traits, making it suitable for appraising early-screening algorithms but not for measuring autism prevalence in any specific region.

Data preprocessing began with addressing missing values, which were identified through systematic checks [13]. Collected survey records with missing values were handled through complete-case removal, as the proportion of missing entries was minimal and imputation risked distorting categorical behavioral variables. Categorical variables, including sex and ethnicity were encoded for compatibility with ML algorithms, while numerical variables such as age in months and Q-CHAT-10 scores were normalized to maintain consistency.

Exploratory data analysis (EDA)

The dataset was explored for its structure and patterns concerning pediatric early screening of autism. Q-CHAT-10 score was examined to measure autism traits across diagnostic groups ("Class/ASD Traits"). The "Months" column was evaluated for trends based on age, investigating its relationship with diagnostic scores to identify age-specific patterns [14, 15]. Relationships between "Months" and "Q-CHAT-10 score" were evaluated for multi-collinearity and guiding feature selection.

Feature selection and statistical analysis

Multi-collinearity between variables such as "Q-CHAT-10 score" and "Months" was checked to determine interdependencies and remove redundant features. Categorical variables, such as "Sex" (gender) and "Jaundice", were evaluated for associations with diagnostic outcomes to ensure their predictive relevance. Inferential statistical tests, such as t-tests, ANOVA, and normality tests, were performed [16]. Pearson and Spearman correlation coefficients were calculated to assess linear and monotonic relationships [17], informing the selection of variables with high explanatory power. Behavioral response columns A1 to A10 were analyzed collectively and individually, capturing even subtle patterns associated with autism traits. These systematic analyses made it easy to identify the critical predictors, thus creating a strong foundation for the subsequent modeling tasks. Dimensionality reduction via principal component analysis (PCA) was applied to explore global linear variance structure and identify redundant features, whereas t-distributed stochastic neighbor embedding (t-SNE) was used to visualize high-dimensional behavioral patterns and non-linear class separation. Using both allowed complementary insights, i.e., PCA for dimensionality reduction, t-SNE for qualitative cluster inspection [18].

Predictive modeling

To predict autism traits, models comprising Support Vector Machines (SVM), Random Forest (RF), Gradient Boosting (GB), Logistic Regression (LR), Multilayer Perceptron (MLP), k-Nearest Neighbors (kNN), and Decision Trees (DT) were developed and compared [19]. The rationale for selecting these predictive models was to capture a diverse range of learning paradigms from linear classifiers to complex ensemble and neural approaches. Hyperparameters of the model were tuned via grid search and randomized search for the optimization of model performance [20]. Parameters optimized include kernel type, C-parameter, and gamma for SVM; number of estimators, maximum depth, and minimum samples split for RF; learning rate, number of boosting stages, and subsample ratio for GB; and number of hidden layers, neurons per layer, and activation functions for MLP. Hyperparameters were optimized using grid and randomized search within 10-fold cross-validation, including SVM (kernel, C, gamma), RF (n_estimators, max_depth, min_samples_split), GB (learning_rate, n_estimators, subsample), and MLP (architecture and activation). Optimal configurations yielded better predictive performance and generalization, reducing the possibility of overfitting. To avoid overfitting and evaluate the model, k-fold cross-validation ($k = 10$) was used [21]. The metrics analyzed included accuracy, F1 score, confusion matrix, ROC curves, and AUC scores. The dataset contained approximately 69.1% toddlers classified as ASD-positive and 30.9% as ASD-negative, indicating a moderate class imbalance. Given the class imbalance inherent in the dataset, precision-recall (PR) curves were used to evaluate the model performance [22]. Unlike ROC curves, PR curves focus on positive class predictions, making them suitable for imbalanced data scenarios.

Feature importance using explainable AI

ML models were used to rank the importance of the feature; the most influential predictors were found for the ranking and refinement of the selected features. The top features selected were behavioral scores, measures of interaction, and some clinical indicators. To further enhance interpretability, SHAP values quantified the contribution of each feature to specific predictions, and mean SHAP magnitudes were combined for overall interpretability. LIME clarified individual predictions through locally accurate surrogate models, emphasizing the key features for each record. For both, the output is interpreted relative to the binary screening label (Class/ASD Traits: "Yes" vs "No"), allowing model decisions to be traced back to specific Q-CHAT-10 items and covariates.

Results

Data preprocessing and explanatory data analysis

A comprehensive exploratory analysis of the dataset illustrated key behavioral, demographic, and clinical (health-related) patterns relevant to early ASD screening. The Q-CHAT-10 score distribution across ASD traits, with higher median Q-CHAT-10 scores in "Yes" subjects than in the "No" group, suggests a distinction between the score distribution (Figure 2 A), The observed difference in Q-CHAT-10 distributions was a result of the labeling rule based on the total score threshold (> 3 vs. ≤ 3) for the "Class/ASD Traits" label, rather than indicating a separate diagnostic category. The density plot for months across ASD traits in Figure 2 B shows an overlap in distributions, with slightly higher density in older age for the group having traits of ASD. Slightly higher density of older ages in the "Yes" group likely reflects a dataset-level screening pattern. Caregiver concerns and observable social-communication differences may become more apparent over 12–36 months, and the community-collected, open-participation design can introduce selection effects in who was screened and when. Figure 2 C presents a plot of Q-CHAT-10 scores across ethnic groups, in which ethnicities such as "White European" and "South Asian" have higher densities of toddlers with increased Q-CHAT-10 scores, while other ethnic groups have more heterogeneous distributions and are underrepresented, so any ethnicity-specific patterns should be interpreted cautiously and not treated as population-level inference. The pie chart in Figure 2 D shows the prevalence of jaundice in the population. The plot in Figure 2 E depicts the sex distribution; males (cyan) outnumber females (red) in this dataset, indicating a potential gender imbalance in the sample. Mean scores of behavioral questions (A1 to A10) are highlighted in Figure 2 F. Behavioral questions A7 and A10 show higher scores compared to others, indicating variability in responses across behavioral traits. These distributions suggest that behavioral items, rather than demographic variables, hold a stronger screening/predictive signal in early childhood, consistent with developmental screening theory.

Statistical analysis

Since the select dataset includes binary and continuous variables, the kernel density estimate (KDE) plot of Q-CHAT-10 scores (Figure 3 A) showed bimodality, with Class 0 (absence of ASD traits) peaking at lower values, while Class 1 (presence of traits) had a wider spread toward higher scores. The joint KDE of Months and Q-CHAT-10 (Figure 3 B) revealed Class 0 concentrated at lower ages, while Class 1 extended to higher ages and scores, suggesting interaction effects. Binary variables A1-A10 were balanced (means: 0.40–0.65), while Months averaged 28 months (SD = 7.98, range: 12–-36). Q-CHAT-10 averaged 5.21 (SD = 2.91, range: 0–10). Categorical distributions were imbalanced: 70% male, 31.7% White European, 72.7% without jaundice, and 83.9% without ASD family history. Variable A9 was highly significant (t = 31.78, p < 1e-150) and associated with sex (ANOVA, F = 8.11, p = 0.004). Figures 3 C and 3 D display age variation by test group and score differences across sex, ethnicity, and ASD trait.

Multiple regressions and correlation

A multiple linear regression model was developed for classifying Class/ASD Traits. The model had R^2 of 0.685, meaning approximately 68.5% of the variation in ASD traits could be explained by the selected predictors. Class/ASD traits were represented as binary (0 = No ASD traits, 1 = ASD traits), with Q-CHAT-10 score, age, behavioral items (A1-A10), and demographic/clinical covariates (gender (sex), ethnicity, jaundice, family history) used as predictors. Q-CHAT-10 score was a strong positive predictor at p < 0.001, and Month was a strong positive predictor, with p = 0.002, indicating importance in the identification of ASD traits. In addition, many binary covariates were statistically significant, such as A3, A5, A6, A7, A9, and A10. Although A3 and A10 had negative effects, others had positive influences on the model. The influence of demographic variables such as sex and jaundice, although weaker, was statistically significant (p < 0.05). Whereas, ethnicity and family history of ASD were not significant predictors. Negative coefficients (e.g., A3, A10) indicate that answering "yes" to these items was associated with a lesser likelihood of ASD traits. This fully aligns with the clinical interpretation that behaviors such as pointing to request (A3) reflect typical social-communication development, whereas positive coefficients corresponded to behaviors more frequently impaired in ASD. The correlation matrix (Figure 4) highlighted key relationships among variables. Binary items such as A1, A2, A4, A5, and A9 showed moderate to strong associations, with A9 strongly related to A4 (r = 0.43) and A5 (r = 0.44). A10 exhibited consistently weak correlations, suggesting limited predictive value. Continuous variables showed weak association with Months, but Q-CHAT-10 score correlated strongly with multiple binary features, notably A1 (r = 0.61), A2 (r = 0.59), and A5 (r = 0.65). Also, Q-CHAT-10 score correlated very

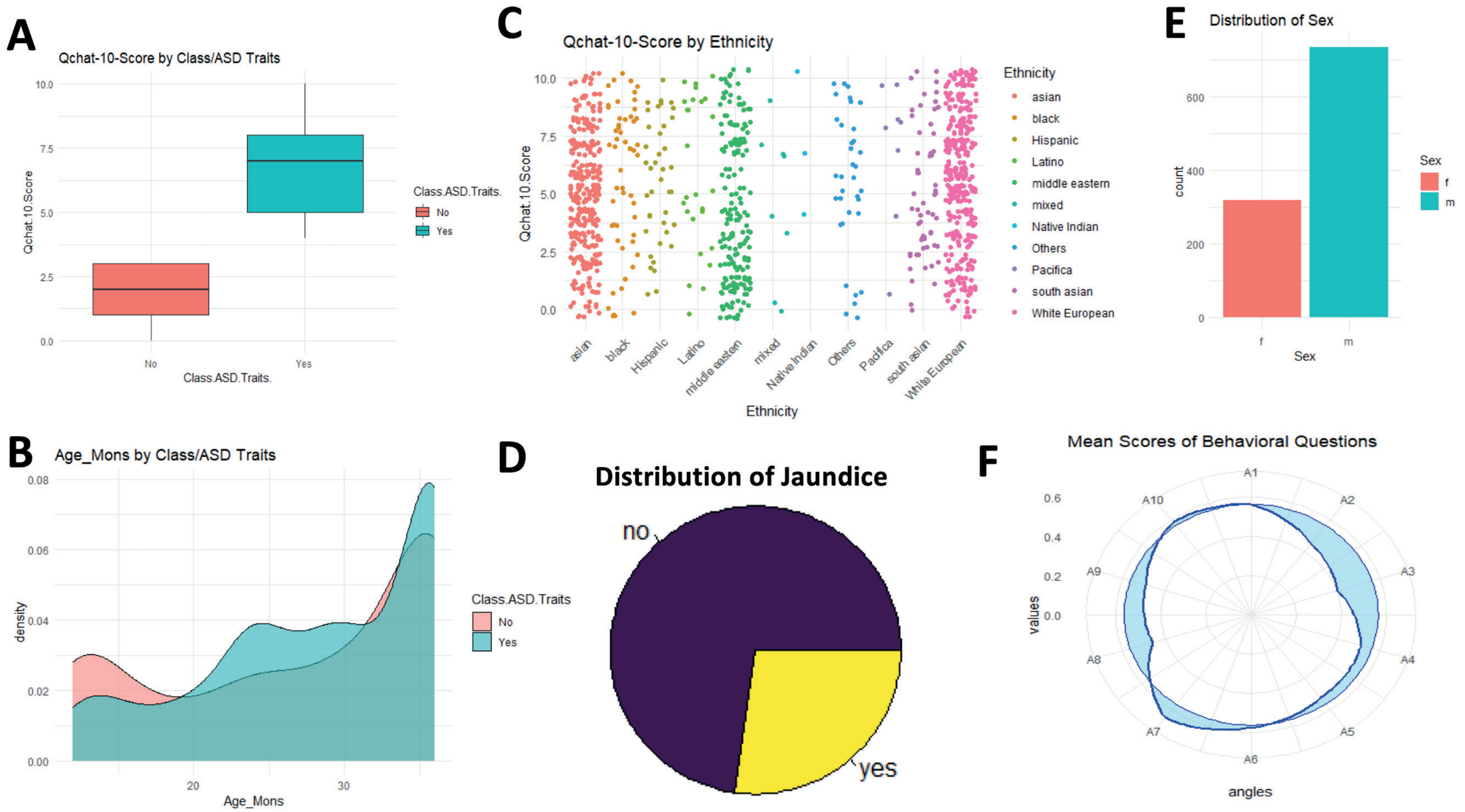

Figure 2. A – Q-CHAT-10 boxplot. B – Age density. C – Ethnicity distribution. D – Jaundice prevalence. E – Sex distribution. F – Radar plot of behavioral questions, highlighting demographic, health, and behavioral differences

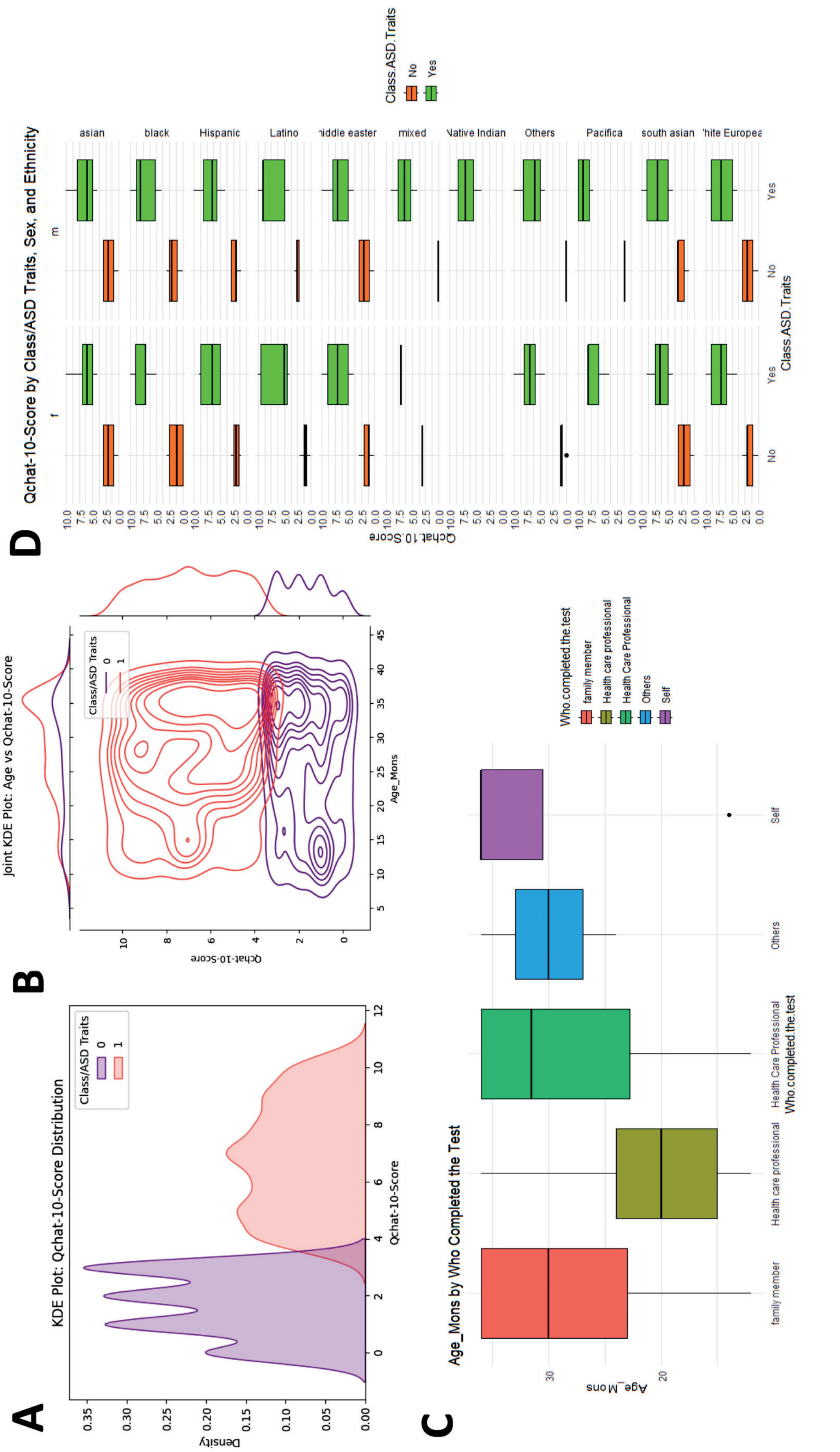

Figure 3. A – Q-CHAT-10 score distribution. B – Age vs. Q-CHAT-10 KDE. C – Months by completion group. D – Q-CHAT-10 comparisons

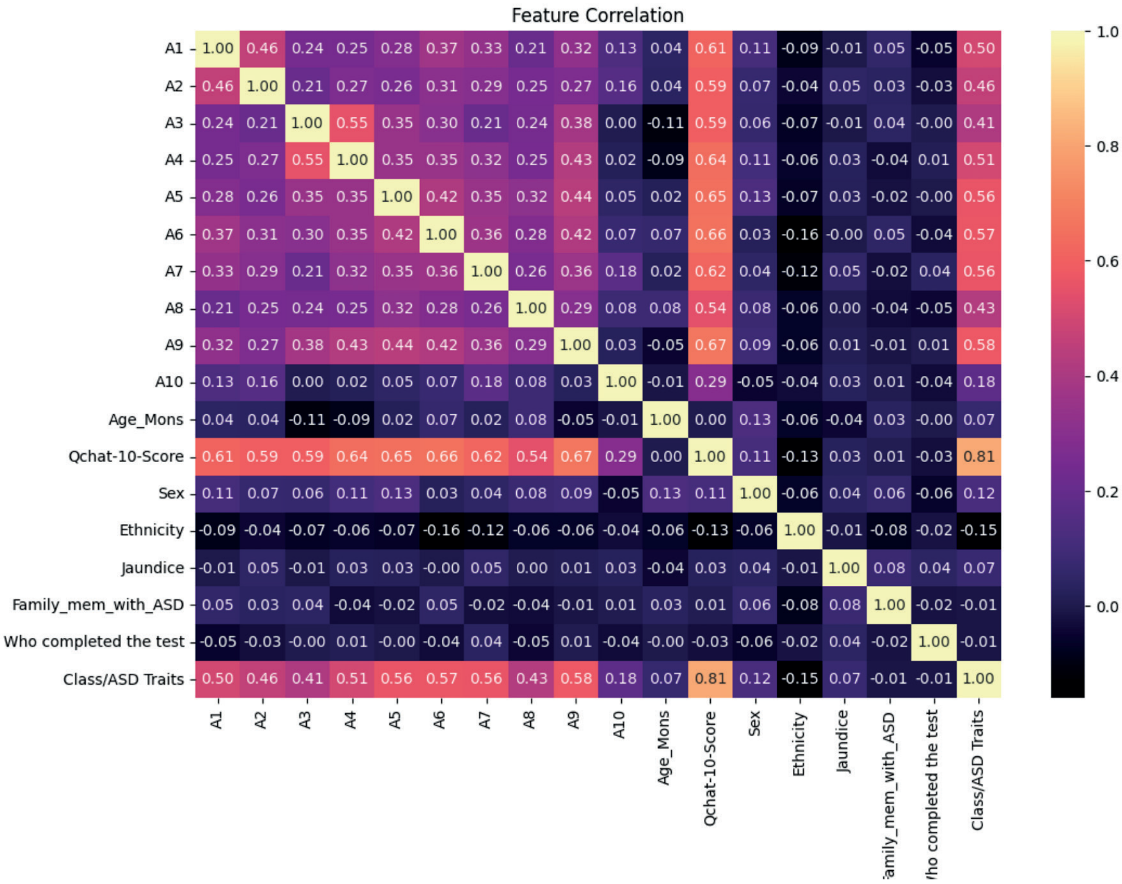

Figure 4. Heatmap of the correlation matrix of Pearson correlation coefficient for features, which contains Q-CHAT-10 score, demographic variables, and ASD-related traits. Gradient intensity represents color

strongly with Class/ASD Traits (r = 0.81), accentuating its central role, supported by A1, A4, and A5.

Predictive modeling and explainable AI

Hyperparameter optimization was conducted to enhance classification accuracy. SVM performed best with an RBF (Radial basis function/Gaussian) kernel, C = 1.0, and gamma = 0.1. RF performed optimally with 100 estimators, maximum depth = 10, and minimum split = 5. GB yielded peak results with a learning rate of 0.05, 200 boosting stages, and a subsample ratio of 0.8. For MLP, the best setup was three hidden layers of 100 neurons each with ReLU (Rectified Linear Unit) activation. Five classifiers – SVM, RF, GB, MLP, and kNN – were evaluated using 10-fold cross-validation. GB achieved the highest accuracy (0.99), followed closely by RF and MLP (0.98), whereas KNN (0.71) and SVM (0.67) showed lower accuracy. However, ROC curves demonstrated strong discriminative ability for GB, RF, and MLP (AUC = 0.98), while SVM also performed well (AUC = 0.95), far outperforming KNN (AUC = 0.61).

Despite sex and jaundice being statistically significant ($p < 0.05$), their predictive impact was minor. ML models demonstrated that the strongest predictive signal came from the Q-CHAT-10 total score and key behavioral items, indicating that clinical and demographic covariates offer little additional value beyond behavioral characteristics. Furthermore, precision-recall analysis showed that SVM remained highly effective in identifying ASD-positive cases, underscoring its sensitivity in imbalanced data. These comparative evaluations are summarized in Figures 5 A, B. Feature importance analysis further highlighted key predictors. Behavioral scores A1, A4, A5, A9, and A10, along with the Q-CHAT-10 score, emerged as the strongest contributors in both RF and GB rankings (Figure 5 C). Demographic variables such as Months and Ethnicity showed minimal influence. To enhance interpretability, SHAP analysis was applied. Globally, the Q-CHAT-10 score had the highest mean SHAP value, followed by A9, A7, and A5 (Figure 5 D). These features were ordered based on their total impact on the model output throughout the dataset; higher average SHAP contribution values signify variables that had a stronger overall effect on predictions. At the individual level, LIME plots illustrated how a high Q-CHAT-10 score combined with responses such as A6, A7, A9, and A1 produced near-certain ASD predictions (Figure 5 E),

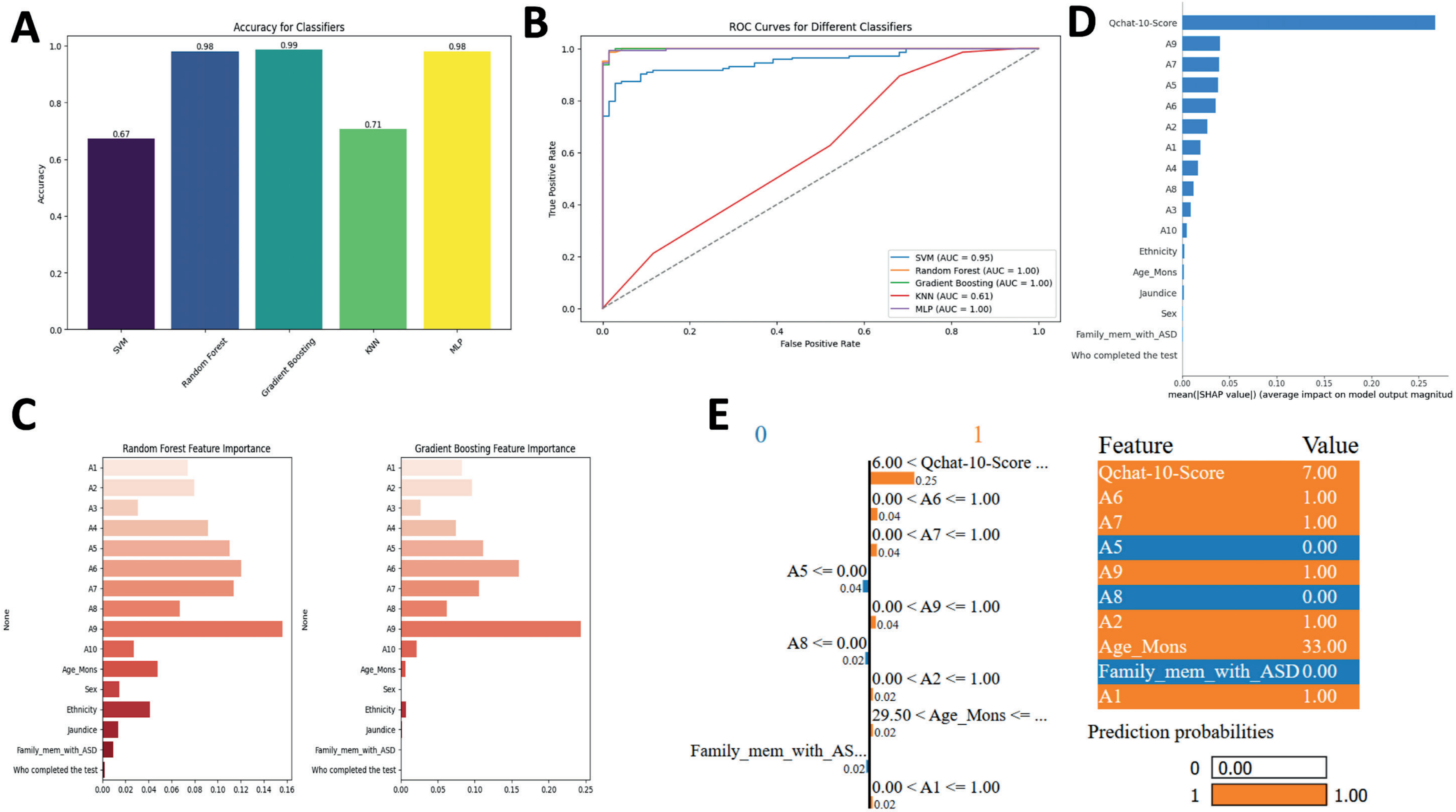

Figure 5. A – Comparative performance evaluation of five ML classifiers using accuracy, **B** – training/validation loss, **C** – feature importance rankings from Random Forest and Gradient Boosting models, **D** – SHAP analysis of global feature impact, **E** – LIME explanations

highlighting the most significant features for individual predictions, the feature values aided contributing to the predicted class, and the associated prediction probabilities. These findings corroborate the central predictive role of Q-CHAT-10 and select behavioral traits, while reinforcing model transparency, as SHAP (global) and LIME (local) offer complementary clinical interpretability. SHAP reveals which behavioral factor(s) and scores most consistently influence overall screening predictions, while LIME explains why a specific child is labeled as "ASD traits" rather than "No ASD traits," facilitating transparent evaluation of model results in a screening. However, the superior performance of ensemble models reflects their ability to capture complex non-linear developmental patterns characteristic of ASD.

Discussion

This study detected autism characteristics in toddlers using behavioral, demographic, and clinical characteristics data by evaluating the predictive ability of Q-CHAT-10 behavioral items through the comparative performance of multiple ML models for early autism screening in toddlers. The findings showed that Q-CHAT-10 items, particularly A1 ("Does your child look at you when you call his/her name?"), A4 (pointing to share interest), and A5 (pretend play), were the strongest predictors of ASD traits. Demographic and clinical factors such as sex and jaundice showed limited influence, while ethnicity and family history were poor predictors. Considering the app-based, community-collected nature and the gender disparity, these covariates should be considered as contextual elements in screening rather than as diagnostic indicators. External validation in clinically determined, balanced samples is warranted prior to clinical interpretation. Among the tested models, GB and MLP achieved the highest classification accuracy (ROC-AUC = 0.99). Features included important Q-CHAT-10 Toddler questions (A1-A10), age, sex, ethnicity, family history of autism, and jaundice status [3, 4, 23]. Scores on Q-CHAT-10 are a well-known diagnostic test for assessing autism traits. Significance tests identified strong predictors of autism traits [24, 25]. Machine learning models, including GB, MLP, SVM, and others were developed and optimized through a grid search.

The current findings can be better understood when compared with previous literature and theoretical perspectives on early autism detection. Unlike traditional tools such as ADOS, this study used ML approaches for faster, scalable autism screening [3, 23, 26]. This aligns with the theoretical framework emphasizing early social communication deficits as core features of ASD, where behaviors such as response to name and shared attention reflect impairments in social orienting and joint attention [27].

Traditional existing tools are accurate but time-consuming; whereas ML models provide efficiency and scalability. Compared to traditional tools such as M-CHAT-R/F that report sensitivities between 0.85–0.91 and specificities around 0.80 [11, 28], ML models used in the current study demonstrated superior classification performance. Although this promises a potential clinical significance, further work is warranted to validate these findings in real-world screening scenarios with larger samples. Recent ML approaches in autism screening have typically focused on either behavioral checklists alone or advanced biomarkers [10]. An ensemble combining LR and KNN using Q-CHAT-10 behavioral features achieved up to 99% accuracy, highlighting the value of optimized feature selection [29]. The integration of demographic and clinical covariates with Q-CHAT-10 behavioral items into the generated ML models constitutes a novel, more holistic framework, achieving comparable or superior predictive performance. In line with previous studies, the current model demonstrated that social-communication indicators (A1, A4, A5) remain the most reliable and consistent early markers of ASD, underscoring their roles in the neurodevelopmental theory of social attention [30].

Early ASD detection in toddlers supports timely interventions for better outcomes. The efficiency of ML models makes them suitable for large-scale ASD screening. They can be implemented in pediatric clinics or through mobile applications, thus allowing for far-reaching access to early screening for autism. The results for interpretability (feature rankings, SHAP, and LIME) can aid personalized clinical follow-up by emphasizing which behavioral factors most significantly influence a positive screening result for a specific profile, thus assisting clinicians in prioritizing areas for additional evaluation and referral instead of recommending treatment. As a limitation, the currently used dataset does not include longitudinal data to track the developmental changes over time, which could refine predictions and improve the understanding of ASD traits. Model performance might not apply to under-represented populations, as the dataset is gathered from the community instead of being based on organized epidemiological sampling. Although ethnicity displays limited predictive significance, the imbalanced sample composition indicates that performance specific to groups has not yet been evaluated. Despite the use of 10-fold cross-validation and hyperparameter tuning, the imbalance in samples between ASD traits and non-ASD traits could skew models towards the majority class, leading to inflated performance estimates. The reported metrics indicate cross-validated outcomes within this dataset, and validation on more balanced cohorts is required. Further, the integration of multi-modal data, including behavioral, genetic, and neuroimaging information, may increase the robustness of screening models. The demographic skew of the currently used dataset, with a predominance of White Europeans, restricts its generalizability; hence, future studies involving larger samples with diverse populations are required [22]. This limitation also restrains theoretical modeling of ASD developmental trajectories, as earlier work highlights the significance of tracking symptom stability over time to differentiate transient versus persistent social-communication delays. As the class label is derived from a screening score within the app, the current results reflect the prediction of screening-based ASD traits and require external clinical validation before their clinical application. This study offers predictive frameworks with potential to transform screening and intervention [31, 32]. ML models may eventually become automated systems capable of analyzing various data in real time. Even the treatment response could then be predicted, allowing individuals with autism traits to change their care plans dynamically to adapt to their needs. Even with robust cross-validated results, the practical implementation of an ML-driven ASD screening model could encounter operational issues, such as uniform questionnaire delivery, managing incomplete or unreliable data, and guaranteeing applicability across various populations. Model outputs ought to function as decision support with clear summaries, rather than substituting for diagnosis. External validation in clinically confirmed, demographically varied groups is crucial prior to implementation.

Overall, this study demonstrates how explainable ML can improve early ASD screening in toddlers. The developed model integrates behavioral, demographic, and clinical characteristics within a single predictive framework, representing a scalable, low-cost screening tool with the potential to improve equitable access to pediatric care. Also, the interpretability of SHAP-derived explanations improves clinical transparency, by offering meaningful insight into the developmental behaviors that are the most decisive for a positive screening outcome. The current study is a major step forward in data-based screening for autism, with results of great importance to early screening, intervention, and policy-making in healthcare. It provides a thorough comparison of various model families using a uniform evaluation framework with cross-validation and optimized hyperparameters, an integrated feature set that merges Q-CHAT-10 behavioral items with demographic and clinical covariates in one pipeline, and clear interpretability through feature-importance ranking and SHAP/LIME explanations that link predictions to specific screening behaviors. Such AI/ML supported technologies can facilitate the early ASD testing process in various healthcare settings; consequently, there will be fewer constraints in the provision of specialist services. Facilitating fast, standardized, and scalable screening through ML approaches could serve a broad range of healthcare initiatives including early childhood health programs, routine developmental assessments in primary care, and community-based screening campaigns at local, regional, or national levels [33]. Future studies can further extend this work to improve the knowledge and management of autism, addressing its weaknesses and enlarging on the findings, and policymakers may use such insights to expand early ASD screening in underserved regions.

In conclusion, the present study demonstrated that ML can be successfully exploited in efficient detection of autism traits in toddlers through Q-CHAT-10 behavioral assessments amalgamated with demographic and clinical data. The Q-CHAT-10 total score and key behavioral markers that include response to name (A1), pointing to share interest (A4), and pretend play (A5) emerged as the most significant predictors, endorsing the model's clinical relevance. The current findings suggest that screening systems based on explainable ML could be a "side-effect" instrument for policy initiatives that aim to improve ASD detection at an early stage, especially in areas that have a shortage of specialist diagnostic services. Facilitating rapid, uniform, and scalable screening via such means could be used in various healthcare-related programs. Health policymakers can use these frameworks to cut down on the time waiting for screening and facilitate fair access to early intervention, and guide the allocation of desired resources for neurodevelopmental services. Thus, in practical application, this model would serve as a preliminary screening tool in standard developmental evaluations or telehealth screening processes. It can identify high-risk profiles for prompt specialist referrals and facilitate follow-up assessments, under clinician supervision. Importantly, model outputs are intended for screening assistance, not as definitive diagnostic guidance. Nonetheless, uniform preprocessing and external validation across different clinical environments are crucial prior to implementation. Ultimately, this framework has the potential to become a valuable tool in early ASD detection and contribute positively to long-term developmental outcomes.

Acknowledgments

The author gratefully acknowledges the funding of the Deanship of Graduate Studies and Scientific Research, Jazan University, Saudi Arabia, through project number: JU-202503322-DGSSR-RP-2025.

The author is grateful to Dr. Shafiul Haque, Dr. Sandeep K. Singh, and Mr. Tayyab Ijaz for their support in the study design and manuscript reviewing. Some parts of the text were enhanced with the assistance of AI-based writing tools (ChatGPT and Grammarly) for grammar, clarity, and language improvement. All scientific content, data interpretation, and conclusions remain solely the responsibility of the author.

Funding

Deanship of Graduate Studies and Scientific Research, Jazan University, Saudi Arabia – Project No. JU-202503322-DGSSR-RP-2025.

Ethical approval

This study was performed using a publicly available and fully anonymized dataset sourced from the Kaggle Autism Screening for Toddlers repository (https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers). The dataset contains no personally identifiable information and is provided for research and educational purposes. The present analysis complies with all data-use terms specified by the dataset creators and the Kaggle platform. This research work comprises a secondary analysis of de-identified survey data; hence, no additional institutional ethical approval was required. All procedures were carried out in accordance with general ethical standards for the responsible handling, storage, and analysis of human-related data.

Conflict of interest

The author declares no conflict of interest.

References

1. Vivanti G, Algur Y, Ryan V, et al. The impact of using standardized autism screening on referral to specialist evaluation for young children on the autism spectrum: a cluster-randomized controlled trial. J Am Acad Child Adolesc Psychiatry 2024; 64: 686-98.
2. Sheldrick RC, Hooker JL, Carter AS, et al. The influence of loss to follow-up in autism screening research: taking stock and moving forward. J Child Psychol Psychiatry 2024; 65: 656-67.
3. Kover ST, Davidson MM, Sindberg HA, Ellis Weismer S. Use of the ADOS for assessing spontaneous expressive language in young children with ASD: a comparison of sampling contexts. J Speech Lang Hear Res 2014; 57: 2221-33.
4. Shumway S, Farmer C, Thurm A, Joseph L, Black D, Golden C. The ADOS calibrated severity score: relationship to phenotypic variables and stability over time. Autism Res 2012; 5: 267-76.
5. Gotham K, Pickles A, Lord C. Trajectories of autism severity in children using standardized ADOS scores. Pediatrics 2012; 130: e1278-84.
6. Sivayokan B, Sivayokan S, Thirunavukarasu K, Sathiadas G, Sivapathamoorthy T. The characteristics of autistic children attending a neuro-developmental center in Northern Sri Lanka. Cureus 2023; 15: 35970.
7. Sheth F, Shah J, Jain D, et al. Comparative yield of molecular diagnostic algorithms for autism spectrum disorder diagnosis in India: evidence supporting whole exome sequencing as first tier test. BMC Neurol 2023; 23: 292.
8. Selten I, Boerma T, Everaert E, et al. Behaviors related to autism spectrum disorder in children with developmental language disorder and children with 22q11.2 deletion syndrome. Autism Dev Lang Impair 2023; 8: 23969415231179844.
9. Montanaro FAM, Mandarino A, Alesi V, et al. PTCHD1 gene mutation/deletion: the cognitive-behavioral phenotyping of four case reports. Front Psychiatry 2023; 14: 1327802.
10. Sankari MS. Predicting autism traits using Q-CHAT behavioral features: a machine learning approach. Engin Res Express 2025; 7: 025269.
11. Robins DL, Casagrande K, Barton M, Chen CMA, 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: 37-45.
12. Gellert BM, Raczkiewicz D, Religioni U, Pinkas J, Ostrowski J. Evaluating physician knowledge of autism spectrum disorder: addressing diagnostic challenges, comorbidities, and the need for targeted education. Arch Med Sci 2025. https://doi.org/10.5114/aoms/205165.
13. Karrar AE. The effect of using data pre-processing by imputations in handling missing values. Indon J Electr Engin Infor 2022; 10: 375-84.
14. Shahamiri SR, Thabtah F, Abdelhamid N. A new classification system for autism based on machine learning of artificial intelligence. Technol Health Care 2022; 30: 605-22.
15. Scott AJW, Wang Y, Abdel-Jaber H, Thabtah F, Ray SK. Improving screening systems of autism using data sampling. Technol Health Care 2021; 29: 897-909.