

Machine Learning and Explainable AI for Q-CHAT-10–Based Autism Spectrum Disorder Screening Assistance in Toddlers: Comparative Evaluation and Web Deployment

¹Prof. Raj Kumar Thakur, ²Sagar Sitaula, ³Prof. Dr. Gopal Prasad Sharma, ⁴Chhayachabbi Jha, ⁵Kiran Barakoti

^{1,3} Professor, Purbanchal University School of Science & Technology (PUSAT), Biratnagar, Nepal

^{2,4,5} Independent Researcher, Nepal

² ORCID ID: <https://orcid.org/0009-0006-1687-6594>

⁵ ORCID ID: <https://orcid.org/0009-0005-1250-5342>

DOI: <https://doi.org/10.51244/IJRSI.2026.1309000028>

Received: 23 September 2026; Accepted: 28 September 2026; Published: 30 September 2026

ABSTRACT

Machine-learning methods applied to questionnaire data may support early developmental screening for Autism Spectrum Disorder (ASD), provided that model outputs are interpreted as screening assistance rather than as evidence of clinical diagnosis. This study develops and evaluates a comparative machine-learning and artificial neural-network framework for Q-CHAT-10–based toddler screening and extends the system with explainable artificial intelligence (XAI) and web-based decision-support functionality. The original dataset contained 1,054 records; after removal of the case identifier and 79 repeated response profiles, 975 unique cases remained, comprising 691 screening-positive and 284 screening-negative observations. Sixteen predictors were retained after exclusion of the pre-computed Qchat-10-Score to reduce direct target leakage. Nine classifiers were evaluated under a common 80/20 stratified hold-out design with training-only SMOTE and StandardScaler transformation: Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, polynomial and radial-basis-function Support Vector Machines, Gaussian Naive Bayes, Quadratic Discriminant Analysis, and an early-stopped Multilayer Perceptron. Logistic Regression achieved 100% held-out accuracy, sensitivity, specificity, F1-score, and ROC-AUC, with a log loss of 0.0272; McNemar’s exact test identified seven discordant cases favoring Logistic Regression over the MLP-ANN ($p = 0.0156$). A source-data audit showed that the screening label is deterministically derived from the sum of A1-A10 at a threshold of four, indicating that the perfect result reflects reconstruction of the questionnaire-based screening rule rather than independent clinical diagnosis. The Streamlit implementation incorporates SHAP and LIME explanations, with LIME outputs presented in native feature units for improved interpretability, together with valid-domain what-if analysis, exploratory data analysis, batch inference, downloadable screening reports, and all-model confusion-matrix visualization. The framework demonstrates transparent automation of ASD screening assistance while emphasizing the need for external clinician-confirmed validation before clinical decision-support use.

Keywords: Autism spectrum disorder, Q-CHAT-10, machine learning, artificial neural network, explainable AI

INTRODUCTION

Autism Spectrum Disorder (ASD) is a heterogeneous neurodevelopmental condition characterized by persistent differences in social communication and interaction together with restricted or repetitive patterns of behavior, interests, or activities. Support needs vary substantially among autistic children and across developmental stages. Prevalence estimates also vary according to population, surveillance methodology, diagnostic criteria, and ascertainment practices. The World Health Organization estimates that approximately 1 in 127 people globally had autism in 2021, whereas surveillance by the U.S. Centers for Disease Control and Prevention reported approximately 1 in 31 eight-year-old children in the 2022 ADDM surveillance network [9, 26]. These estimates

describe different populations and surveillance systems and should therefore not be interpreted as directly comparable measures of prevalence.

Early developmental screening can shorten the interval between caregiver concern and specialist assessment; however, a screening result does not constitute a clinical diagnosis. Q-CHAT and Q-CHAT-10 are parent-report questionnaires designed to screen for autistic traits in toddlers [3], [4]. Prospective work has demonstrated acceptable screening performance and imperfect positive predictive value and repeated developmental surveillance [5]. Comparative analyses with screening using the M-CHAT also demonstrate that age, threshold and follow-up procedures significantly influence sensitivity and specificity [23]. Thus, the present system is not evaluated as an independent diagnostic system, but rather as a classifier of Q-CHAT-10 screening status..

One advantage of machine learning for developmental screening is the possibility of processing structured responses quickly and with minimal computational expenses. Questionnaire data, health records, eye tracking, EEG, speech, motion, gait, facial images, and neuroimaging have been used to study ASD [1], [2], [8], [11], [12], [14], [17], [27]. Classical supervised models tend to be very good on structured screening data, whereas signal- and image-based approaches seek information that is less directly related to the scores on the questionnaires. Due to the differences in populations, endpoints, modalities and validation designs among these studies, headline accuracy values should not be compared without taking into account the underlying task.

One methodological issue that frequently occurs is that the target label is obtained from the same questionnaire items provided as predictors. In such a context, a model can learn the scoring rule with very high fidelity without having independent diagnostic validity. In a separate study, Sollis et al. explicitly differentiated the prediction of a QCHAT-derived assessment result from transfer to a clinically diagnosed cohort, and observed lower performance in the latter task [22]. This distinction is helpful for the current interpretation: The within-dataset reconstruction of a deterministic screening endpoint can help validate automation, but cannot be used as a predictor of a clinician-confirmed ASD diagnosis.

The following framework is implemented: Nine classifiers are compared after removing the pre-computed Qchat-10-Score, SMOTE is only applied to the training partition, multiple performance measures are evaluated, early stopping is applied for MLP-ANN and paired McNemar testing is used, and the models are serialized using Streamlit. An implementation audit was also conducted for the dataset, scripts, model artefacts, evaluation outputs and deployment code. The audit showed 975 unique response profiles, 691 screening-positive cases, 284 screening-negative cases, 1,106 records in the post-SMOTE training set and the separation of the production estimators and a separate exploratory GridSearchCV workflow.

The research thus seeks to create an interpretable, rigorous and deployable machine-learning and ANN system for early screening of ASD in toddlers. The goals are to compare the performance of nine production classifiers on a new test set from the Q-CHAT-10; quantify error trade-offs, overfitting indicators, and the impact of SMOTE; add local interpretability with SHAP, LIME, and what-if analysis; and interpret all results in the context of the construction of the screening endpoint and the implementation of the executable.

Five research questions are raised: (1) What are the differences between various families of classifiers when using a uniform pipeline? (2) What is the right sensitivity-specificity trade-off for screening? (3) How does training-only SMOTE affect different classifiers? (4) Can the framework be used to provide meaningful local explanations without inferring causality? (5) What methodological improvements are needed before this framework could be considered as clinical decision support?

The main contribution is an end-to-end ASD screening framework, which integrates leakage control, benchmarking with 9 models, statistical comparison, model-dependent SMOTE analysis, explanation using SHAP and LIME, what-if sensitivity analysis, and Streamlit deployment. The paper also clearly states that the best Logistic Regression outcome is not necessarily a good clinical diagnosis but rather the reconstruction of a questionnaire-based screening rule.

LITERATURE REVIEW AND RESEARCH GAP

Computational ASD screening research covers the structured questionnaires, the behavioral or biomedical signals, and the explainability and external validation of the computational approaches. These approaches are addressing various questions, such as: questionnaire models can automate assessment rules; medical-record models can predict later diagnosis based on health history; and signal-based models can provide information in part without the involvement of the caregiver's scoring. Reported performance should therefore be viewed in the context of the endpoint, sampling frame and data modality.

The use of questionnaire-based machine learning remains appealing due to the low cost of obtaining responses, familiarity of the responses to the caregiver, and ease of digitization. Achenie et al. identified a short form of the M-CHAT for toddler screening using M-CHAT data [1]. Hossain et al. compared the performance of algorithms on multiple age-specific public ASD datasets and found that the performance is dependent on the dataset and the model family [14] and Jacob et al. explored the use of automated feature ranking and compact signatures [15]. The Logistic Regression accuracy of 100% reported in [20] on a toddler screening dataset and the high accuracy obtained by DNN after imbalance treatment reported in [24] are reported. These studies validate that structured screening data can be very separable, and can be computed efficiently. Concurrently, this type of performance is partially dependent on the labels and feature construction, and therefore results obtained from a questionnaire labelled dataset should not be interpreted as being equivalent to a diagnosis made by the multidisciplinary clinical assessment.

These computational studies are provided with the clinical context of the Q-CHAT literature. The original Q-CHAT was designed as a quantitative parent-report questionnaire of autistic traits [3] and a shorter red-flag version, the Q-CHAT-10, was proposed later [4]. Population follow-up studies have demonstrated that the instrument is able to detect toddlers who need further evaluation, and that positive predictive value is not perfect and that some toddlers present later [5]. The same holds true when comparing screening directly with M-CHAT [23].

Sollis et al. are particularly relevant as they separated prediction of QCHAT-derived labels from transfer to clinician confirmed diagnosis [22]. Their small models were very well calibrated on the questionnaire labeled data but exhibited significantly poorer discrimination on an independent clinical sample. The present study therefore does not attempt to use the internal benchmark as a measure of diagnostic accuracy, but to reconstruct a screening classification, and to perform clinical validation of the screening classification as a necessary step.

Other ASD studies rely on predictors that are not directly related to the scoring of questionnaires. Other works include the modelling of later diagnosis of the medical claims from a longitudinal perspective [11], nonlinear EEG features in infants [8], eye-tracking models [2], movement trajectories derived from a game [17] and gait variables derived from pose estimation using RGB-camera [12]. The modalities can be less accurate than questionnaire-derived benchmarks, but they solve a different, and perhaps more difficult, problem, since the predictors are not algebraically related to the outcome. The benefit of their value is that they can be used to capture independent behavioral or physiological evidence. For future ASD screening systems, these signals could be used as an additional component to parent-report questionnaires to assess the contribution of machine learning to a known scoring rule.

Facial images, speech and MRI have also been studied using deep-learning. Ghazal et al. applied transfer learning to facial images [13], Assaf et al. used speech-transcript features and did not use raw audiovisual data [7] and Yin et al. created a Siamese and path-signature approach for longitudinal structural MRI [27]. There have been successful outcomes, but ongoing issues with standardization and clinical translation in systematic reviews of motion- and eye-movement approaches [21].

In the context of health machine learning, explainability is becoming more and more crucial, as end users wish to understand why a machine learning model is making a certain prediction. SHAP defines a prediction as a sum of feature attributions based on Shapley-value principles [16] and LIME constructs a simple local surrogate around an observation of interest [19]. XAI has also been incorporated into screening pipelines for ASD-specific work [6]. The two approaches are used in the present study as complementary tools for local explanations of the

disease, not as separate evidence of the mechanisms of the disease. This is a key distinction: an explanation may be faithful to a fitted classifier even if the classifier is learning a screening endpoint that is not real, but rather a constructed one. Interpretability, in turn, enhances transparency and auditability, but it is not a replacement for clinical validation.

Another common problem is class imbalance. SMOTE is a method that can be used to interpolate synthetic minority observations and can enhance some classifiers when used only in the training data [10]. Its effect is however algorithm-dependent and the standard SMOTE is not suitable for nominal variables that are only integer-encoded. Thus, it is important to assess oversampling empirically, and not assume that it is always beneficial.

Designing a validation is also significant. A clean held-out test set is a direct estimate of performance on unseen observations, whereas cross-validation can be used to aid model selection and stability analysis only if all data-dependent preprocessing is done within each fold. Preprocessing operations (e.g., resampling, scaling, feature selection, or encoding) before the construction of the folds may introduce information leakage between folds and lead to optimistic results. This is also true for hyperparameter tuning: data used to choose model parameters should be distinct from data used to estimate final model performance. This separation is especially significant for medical machine learning due to the fact that small methodological leaks may manifest as clinically verifiable improvements in a small dataset.

Reference integrity was considered as a scientific report. Information on bibliographic details and publication status was verified where possible from the publisher or from the DOI record, and sources where the publication status was not available were not used as supporting evidence. This was particularly relevant given the many small, heterogeneous studies in the ASD machine learning literature, the results of which can be exaggerated if not taken in context.

Overall, the literature is supportive of multi-model ASD screening research, along with three persistent needs: separating screening-label reconstruction from clinical diagnosis, leakage-aware validation, and interpretable deployment. This is addressed in the present study within a reproducible Q-CHAT-10 framework, but with a conservative interpretation of the endpoint.

METHODOLOGY

A. Study design, comparative modelling, and implementation validation

A retrospective supervised-classification design was employed with an implementation audit of the related software. A public Q-CHAT-10 toddler dataset is used to train 9 classifiers with a common hold-out protocol, where after a statistical comparison, SMOTE ablation, model serialization, and deployment in Streamlit is performed. The audit confirmed that the code reported for preprocessing, model configurations and deployment transformations were consistent with the executable code. Where applicable, a fixed `random_state` value of 42 was provided to ensure the train-test split, resampling, and stochastic estimators were reproducible. The focus was therefore not only on predictive performance but also on traceability of the raw data to the deployed inference.

The raw data, duplicate handling, class labels, preprocessing order, production training script, exploratory tuning utilities, serialized artefacts, evaluation outputs, Streamlit inference path, and XAI functions were inspected during the audit. The main training script was re-run to ensure that the counts reported and the model metrics were consistent with the code and not just based on narrative documentation.

B. Dataset and participants represented in the source data

The analysis is based on the public ASDTests/Q-CHAT-10 screening data 'Toddler Autism dataset July 2018.csv' [25]. There are 19 columns and 1,054 rows in the raw file. Variables are the 10 binary Q-CHAT screening items A1-A10, age in months, sex, ethnicity, history of jaundice, family history of ASD, respondent type, a case identifier, the pre-computed Qchat-10-Score, and the binary Class/ASD Traits label. The mix of behavioral,

demographic and history variables renders the data appropriate for a multi-model screening benchmark, but the data is also a public convenience sample, which restricts the claims of representativeness. The source file does not include any other clinical variables, diagnostic interviews, longitudinal outcomes, or clinician confirmed labels.

Case_No was removed prior to duplicate detection, since without this removal, repeated response profiles would be hidden by Case_No. Seventy-nine duplicates were removed, leaving 975 unique observations. This left 691 screening positive and 284 screening negative records. These are the counts that have been audited and are used throughout the manuscript.

Children in the dataset are 12-36 months old. Sex, ethnicity, jaundice, family history of ASD, respondent type and target are categorical items while A1-A10 are binary 0/1 items. After excluding Case_No and Qchat-10-Score, the study has 16 remaining predictors without any leakiness.

C. Endpoint construction and circularity audit

The resulting 975 deduplicated observations are labeled as follows: Qchat-10-Score is the arithmetic sum of A1-A10; Class/ASD Traits changes deterministically at a score of four: score ≥ 4 is screening-positive and score < 4 is screening-negative. This same relationship was confirmed programmatically.

The explicit sum of the scores is not necessary and will lead to obvious target leakage if included, which is why it is being excluded from the scores. The label is still algebraically related to A1-A10, which are still predictors, though. The modelling task thus tests whether classifiers can reproduce the screening rule derived from the questionnaire from its individual responses, but not an independent test of clinical diagnosis.

This deterministic boundary is visualized in figure 2. The endpoint audit should also be carried into the interpretation, as otherwise the near-perfect performance might be interpreted as diagnostic accuracy.

Early ASD Screening: Experimental, Explainability, and Deployment Workflow

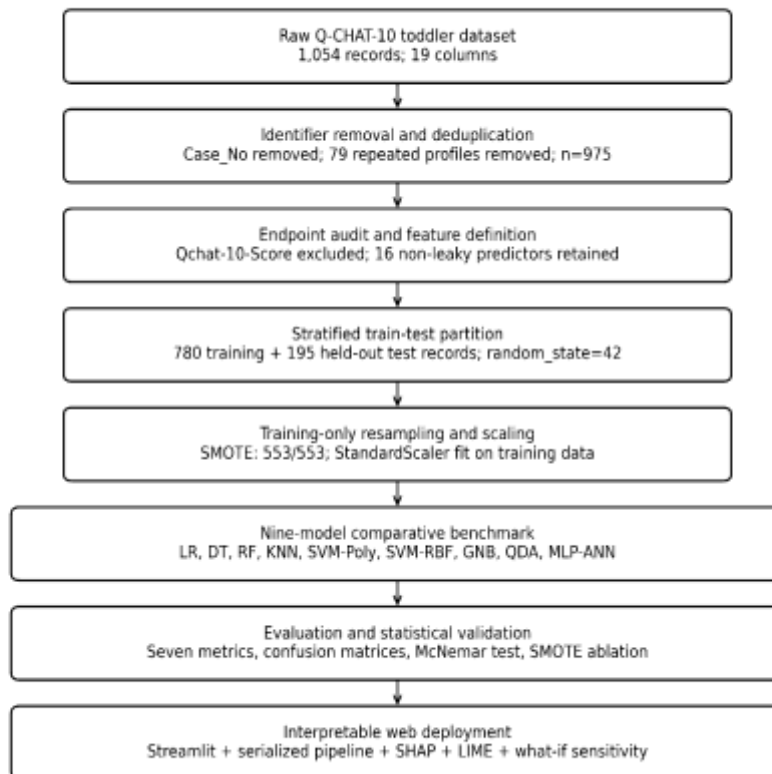


Figure 1. End-to-end experimental, explainability, and deployment workflow for early ASD screening.

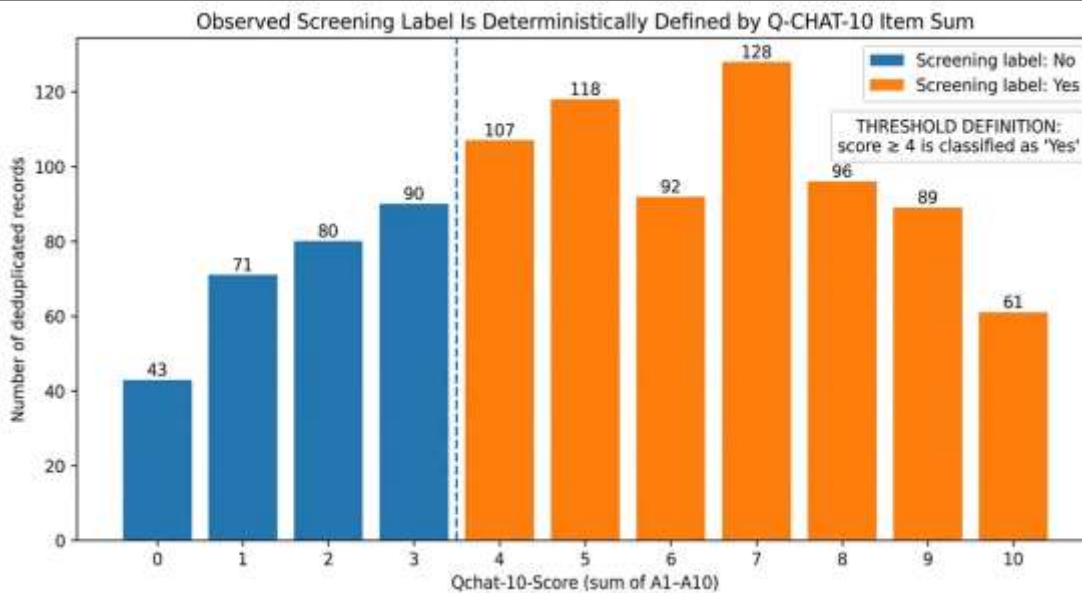


Figure 2. Distribution of the Q-CHAT-10 item sum by screening label after deduplication. The label changes deterministically at a score of four.

Table 1. Audited dataset and partition characteristics.

Element	Audited value	Role in analysis
Raw records	1,054	Original CSV rows
Rows after identifier removal + deduplication	975	79 repeated response profiles removed
Screening-positive label	691 (70.87%)	Majority class in source data
Screening-negative label	284 (29.13%)	Minority class in source data
Training split	780: 553 positive, 227 negative	Used for resampling/model fitting
Post-SMOTE training split	1,106: 553 positive, 553 negative	Synthetic balancing applied to training only
Held-out test split	195: 138 positive, 57 negative	Never resampled; primary evaluation set
Input predictors	16	A1-A10 + age, sex, ethnicity, jaundice, family ASD, respondent
Excluded score	Qchat-10-Score	Exact sum of A1-A10; excluded from X
Target rule	Yes iff $\text{sum}(A1-A10) \geq 4$	Observed without exception in supplied data

D. Feature definition and preprocessing

Case_No was removed and deduplicated, missing values were replaced with the mode value of each column, and whitespace was removed from the names of the columns. Categorical variables were encoded using separate LabelEncoder instances that were saved for deployment. All the columns were then converted to numeric, with any remaining missing values replaced with the mean of the column.

The predictor matrix was created by removing the target and any feature name that has the concept of Qchat-10-Score. This resulted in 16 predictors: A1-A10, Age_Mons, Sex, Ethnicity, Jaundice, Family_mem_with_ASD, and Who completed the test. The same feature order is applied when making inferences with Streamlit.

The production script performs some preprocessing steps before the train-test split, such as imputation, and categorical encoder fitting. As the categories observed are limited, and the target is not included in these transformations, the main leakage controls are the exclusion of Qchat-10-Score and the fact that SMOTE and scaler fitting are limited to the training partition. However, a more powerful future pipeline would allow all data dependent preprocessing to be done in training folds.

E. Train-test partitioning, class imbalance and scaling

The encoded data were split into 780 training and 195 test observations using an 80/20 stratified split with `random_state=42`. There were 138 screening-positive and 57 screening-negative cases in the untouched test set.

In this work, SMOTE was used only to the training partition, using `random_state=42` [10]. The minority screening-negative class was expanded to make up the number of observations in the other class, resulting in 1,106 resampled training observations. No changes were made to the test distribution.

The resampled training matrix was used to fit a `StandardScaler` which was then applied to the training and test predictors. This prevents the influence of test set on scaling parameters. During web inference, the fitted scaler is serialized and reused.

One drawback is that nominal variables were `LabelEncoded` prior to SMOTE. Interpolation between integer category codes may generate "synthetic" values that do not have a natural categorical interpretation. Categorical-aware resampling, mixed-data pipelines, or class weighting should thus be considered for future validation.

F. Production model configurations

The production script trains eight classical estimators and one MLP-ANN. The models used are Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, SVM with polynomial and RBF kernels, Gaussian Naive Bayes, Quadratic Discriminant Analysis and `MLPClassifier`. The settings used when they are deployed are summarized in Table 2; the primary results are based on these production settings, not on the optima they find through a grid search. The model set is deliberately designed to cover linear, tree-based, ensemble, neighborhood, kernel, probabilistic, discriminant and neural paradigms, to allow for a comparison of performance under the same preprocessing and test conditions. This also serves as a useful test to see if additional model complexity is warranted by the structure of the screening endpoint.

The MLP-ANN consists of two hidden layers of 32 and 16 units with ReLU activation, Adam optimization and binary classification using `skikit-learn's MLPClassifier` [18]. With early stopping enabled, 10% of the data is used for internal validation and `n_iter_no_change=10` is used for regularization during training.

G. Exploratory hyperparameter search versus production models

The repository also contains an exploratory `GridSearchCV` workflow. Its reported best settings differ from some of the estimators serialized by the primary training script. To avoid conflating model-selection experiments with deployed inference, the paper reports grid-search results only as exploratory evidence and bases all production performance claims on the actual serialized model configurations.

The exploratory tuning workflow also applies SMOTE and scaling before its internal cross-validation. Because synthetic samples and scaling parameters are therefore derived before fold separation, those scores may be optimistic. They are not used as the primary estimate of generalization.

H. Evaluation metrics and uncertainty

The accuracy, precision, recall, specificity, F1 score, ROC-AUC, log loss and counts in the confusion matrix were used to evaluate each production model on the held-out test set of 195 records. Clinically important: Recall (sensitivity) measures the proportion of screening-positive cases correctly identified, $TP/(TP+FN)$, whereas specificity measures the proportion of screening-negative cases correctly classified, $TN/(TN+FP)$. Precision is the proportion of positive model outputs that correspond to screening-positive records, and the F1-score

summarizes the harmonic balance between precision and recall. ROC-AUC is a metric that assesses the ranking ability over thresholds, while log loss is a metric that penalizes confident errors in the probability output. These measures can be reported together to minimize the likelihood of choosing an accuracy-only model and to facilitate the meaningful interpretation of models with varying error profiles.

The accuracy of the training matrix was compared with held out accuracy as a descriptive train test gap. These gaps can be helpful to highlight big differences, but should not be used to replace independent validation, especially since the training distribution is different from the natural test distribution following SMOTE.

The paired comparison of the MLP-ANN and the best classical model for the same test observations was performed using the McNemar's exact test. The test only deals with discordant pairs and is suitable when two classifiers are tested on the same cases. The null hypothesis was that the two classifiers have the same probability of discordant errors; $p < 0.05$ was considered statistically significant. This test was applied in addition to descriptive metrics and not as an alternative to external validation, since a test with statistical significance on a single internal test set does not prove clinical transportability.

I. SMOTE ablation and cross-validation audit

A separate ablation script retrained selected models with and without SMOTE and evaluated both versions on the unchanged test set. Recall, F1-score, and ROC-AUC were compared to determine whether oversampling improved each model rather than assuming a uniform benefit.

The primary script also reports 10-fold StratifiedKFold results after SMOTE and scaling have already been applied to the entire training partition. Because these preprocessing operations occur before fold construction, the cross-validation estimates are treated as secondary implementation outputs rather than leakage-free evidence of generalization.

J. Leakage-aware learning-curve robustness analysis

The newly supplied repository also contained a learning-curve utility. Its original plot was not used directly as formal validation because scaling was fitted before cross-validation and the ANN configuration did not enable the same early-stopping rule as the production model. For the present robustness analysis, learning curves were regenerated for Logistic Regression, MLP-ANN, and Random Forest using five-fold stratified cross-validation with SMOTE and StandardScaler refitted inside each training fold and with the production estimator settings. Training size was increased from 20% to 100% of the available training portion. This diagnostic examines sample-size dependence while avoiding the most important fold-level resampling and scaling leakage; it remains an internal analysis and does not replace external validation.

K. Software environment, serialization, and deployment

The project is implemented in Python using pandas and NumPy for data processing, scikit-learn for modelling and evaluation, imbalanced-learn for SMOTE, matplotlib/seaborn for figures, pickle for serialization, and Streamlit for deployment. Trained models, the scaler, encoders, feature metadata, ROC data, and evaluation tables are persisted for reuse.

The Streamlit application rebuilds the input record in the saved feature order, applies the stored encoders and scaler, and then invokes the selected serialized classifier. Reusing the training-time artefacts reduces training-serving skew and ensures that web inputs are represented consistently with the fitted model.

Listings 1 and 2 show the essential leakage-prevention and deployment sequence in abridged form. The XAI layer is applied after the production prediction so that interpretability does not modify the underlying classifier.

L. Extended Streamlit workflow and operational functions

The latest application expands the research prototype beyond single-record inference. A Dataset Insights page provides descriptive exploration of age, the original Q-CHAT-10 score distribution, and the encoded feature-

correlation matrix using interactive Plotly graphics. The Qchat-10-Score is displayed only for descriptive auditing and remains excluded from every predictive feature vector. The prediction form uses 0/1 controls for the ten behavioral items, integer age input, and categorical controls derived from the stored label encoders, preventing the invalid 0.5 questionnaire values that can arise from unrestricted continuous sliders.

Operational outputs include a downloadable PDF screening report containing the entered profile, predicted screening class, model probability, and an explicit non-diagnostic disclaimer. A batch-prediction page validates the required feature columns and returns downloadable CSV results for multiple records, while the model-comparison page presents detailed metrics, ROC curves, model ranking, and either single-model or nine-model confusion-matrix views. The batch utility is currently a prototype: it assumes model-ready numeric or encoded feature columns rather than performing complete raw-category transformation for arbitrary clinical CSV files. These functions are therefore reported as reproducibility and workflow features, not as evidence of clinical deployment readiness.

M. Explainable AI implementation with SHAP, LIME, and what-if analysis

The explainable artificial intelligence (XAI) component integrates SHAP, LIME, and a what-if sensitivity analysis to improve transparency of the screening-assistance framework. In the current Streamlit implementation, SHAP explanations are available for Logistic Regression, Random Forest, and Decision Tree. Tree-based classifiers are interpreted using TreeExplainer, whereas Logistic Regression is interpreted using the general SHAP Explainer with the stored scaled test representation used as reference information [16]. The resulting waterfall visualization shows the direction and magnitude of feature contributions for an individual screening record. For the illustrative case presented in this study, the explanation was generated from a valid held-out observation containing only admissible 0/1 Q-CHAT-10 responses rather than from an interpolated questionnaire profile. The explanation layer is applied after prediction and does not modify the fitted classifier or its output.

LIME was implemented using a locally weighted surrogate model constructed around an individual screening observation [19]. Model inference continues to operate on the standardized feature representation required by the trained classifier. For presentation and interpretation, however, the standardized LIME conditions were mapped back to the native feature domain using the fitted preprocessing parameters. Consequently, Q-CHAT-10 behavioral items are displayed using their valid binary values (0 or 1), age is expressed in months, and encoded categorical variables are represented using their original labels where applicable. This separation maintains consistency with the trained prediction pipeline while improving the readability of the explanation for researchers and potential end users. LIME coefficients are interpreted exclusively as locally estimated contributions to the model's screening output and not as diagnostic, biological, or causal effects.

The direction and magnitude of SHAP and LIME attributions describe associations with the model output relative to their respective reference and local-approximation procedures. Neither method demonstrates that a feature causes, prevents, increases, or decreases ASD itself. Attribution values may also vary according to the selected SHAP background data, LIME perturbation neighbourhood, discretization procedure, and other explanation-specific settings. Accordingly, these methods are used to explain the behaviour of the fitted screening classifier rather than to establish clinical importance or causal relationships among individual features.

The what-if module provides a complementary form of model-sensitivity analysis. It permits valid questionnaire or demographic inputs to be modified individually or in combination and subsequently recalculates the selected model's screening output. This functionality allows users to observe how changes in admissible input values influence the fitted model while preserving the valid feature domain. The resulting changes should be interpreted as model-response sensitivity only and not as causal counterfactuals, treatment effects, or estimates of how modifying a behavioural characteristic would alter a child's underlying likelihood of ASD.

Table 2. Model configurations actually used by the primary training script and serialized for deployment.

Model	Production configuration
Logistic Regression	max_iter=1000; default L2 regularisation
Decision Tree	max_depth=5; min_samples_split=20; min_samples_leaf=10; ccp_alpha=0.005; random_state=42
Random Forest	n_estimators=200; max_depth=10; min_samples_split=15; min_samples_leaf=6; max_features=sqrt; min_impurity_decrease=0.001; random_state=42
KNN	n_neighbors=51; uniform weights; Manhattan distance (p=1)
SVM-Poly	degree=2; C=0.1; gamma=scale; probability=True
SVM-RBF	C=0.1; gamma=scale; probability=True
Gaussian Naive Bayes	var_smoothing=1e-8
QDA	reg_param=0.7
MLP-ANN	hidden layers=(32,16); ReLU; Adam; max_iter=200; early_stopping=True; validation_fraction=0.1; n_iter_no_change=10; random_state=42

Listing 1. Core production sequence for score exclusion, stratified hold-out creation, training-only SMOTE, and scaling (abridged from the project repository).

```
# score-like columns and target are excluded from X
leaky_cols = [c for c in df_encoded.columns if 'qchat' in c.lower() or 'score' in c.lower()]
X = df_encoded.drop(columns=['Class/ASD Traits'] + leaky_cols)
y = df_encoded['Class/ASD Traits']

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y)
X_train, y_train = SMOTE(random_state=42).fit_resample(X_train, y_train)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Listing 2. Deployment-time reuse of saved feature order and scaler before classical-model and ANN inference (abridged).

```
input_df = pd.DataFrame(
    [{f: float(user_input[f]) for f in feature_names}],
    columns=feature_names)
```

```
input_scaled = st.session_state.scaler.transform(input_df)
prediction = best_model.predict(input_scaled)[0]
probability = best_model.predict_proba(input_scaled)[0][1]
ann_prob = st.session_state.ann.predict_proba(input_scaled)[0][1]
```

EXPERIMENTAL RESULTS

A. Dataset reconciliation and endpoint audit

The audit confirmed 1,054 raw records and 19 columns. Removing Case_No before duplicate detection eliminated 79 repeated response profiles, leaving 975 unique observations: 691 screening-positive and 284 screening-negative. The 80/20 stratified split produced 780 training and 195 test records, and training-only SMOTE expanded the training set to 1,106 balanced observations.

For every deduplicated record, Qchat-10-Score equaled $A1+A2+\dots+A10$ and the target label followed the exact threshold score ≥ 4 . No exceptions were found. This result is central to interpretation because the retained A1-A10 predictors mathematically determine the screening endpoint.

B. Held-out model performance

Tables 3 and 4 report the complete held-out results. Logistic Regression correctly classified all 195 test observations, producing accuracy, precision, recall, specificity, F1-score, and ROC-AUC of 1.0000 with log loss about 0.0272. Its confusion matrix therefore contained 57 true negatives and 138 true positives with no errors.

Among the remaining models, QDA achieved the second-highest test accuracy (0.9692), with precision 0.9853, recall 0.9710, specificity 0.9649, F1 0.9781, and ROC-AUC 0.9963. The early-stopped MLP-ANN achieved 0.9641 accuracy and ROC-AUC 0.9968, while SVM-RBF achieved 0.9487 accuracy with perfect recall and ROC-AUC 0.9986. KNN obtained perfect specificity but lower recall, and Naive Bayes achieved perfect recall at the cost of more false positives. Random Forest also showed strong discrimination but a less favorable specificity profile than the leading models. Decision Tree and SVM-Poly were the weakest classifiers on the held-out set, demonstrating that more flexible or nonlinear decision functions did not automatically improve performance on this largely additive endpoint.

The error patterns show why accuracy alone is insufficient. SVM-RBF and Naive Bayes produced no false negatives but lower specificity, whereas KNN produced no false positives but missed 13 screening-positive records. In a screening context, these trade-offs would affect referral burden and missed-case risk differently.

Table 3. Held-out classification metrics (n=195).

Model	Accuracy	Precision	Recall	Specificity	F1
Logistic Regression	1.0000	1.0000	1.0000	1.0000	1.0000
QDA	0.9692	0.9853	0.9710	0.9649	0.9781
MLP-ANN	0.9641	0.9645	0.9855	0.9123	0.9749
SVM (RBF)	0.9487	0.9324	1.0000	0.8246	0.9650
Random Forest	0.9385	0.9375	0.9783	0.8421	0.9574
KNN	0.9333	1.0000	0.9058	1.0000	0.9506
Naive Bayes	0.9231	0.9020	1.0000	0.7368	0.9485

Model	Accuracy	Precision	Recall	Specificity	F1
Decision Tree	0.8974	0.9097	0.9493	0.7719	0.9291
SVM (Poly)	0.7538	0.9327	0.7029	0.8772	0.8017

Table 4. Discrimination, probability loss, and descriptive train-test gaps. Approximate SVM log loss may vary slightly because probability estimation has no explicit random_state in the production SVC objects.

Model	ROC-AUC	Log loss	Train accuracy	Train-test gap
Logistic Regression	1.0000	0.0272	1.0000	0.0000
SVM (RBF)	0.9986	~0.059	0.9873	0.0386
MLP-ANN	0.9968	0.1235	0.9882	0.0241
QDA	0.9963	0.0792	0.9828	0.0136
KNN	0.9962	0.1433	0.9503	0.0169
Random Forest	0.9886	0.1829	0.9837	0.0453
Naive Bayes	0.9849	0.4729	0.9530	0.0299
Decision Tree	0.9265	0.2783	0.9421	0.0447
SVM (Poly)	0.8931	~0.376	0.8354	0.0816

C. Confusion matrices and paired significance

Logistic Regression produced TN=57, FP=0, FN=0, and TP=138. The other confusion matrices are summarized in Table 5, allowing recall and specificity to be interpreted from the underlying counts rather than from percentages alone.

McNemar's exact test compared Logistic Regression and the MLP-ANN on the same 195 test observations. Logistic Regression was correct on seven cases that the ANN misclassified, while there were no cases with the opposite pattern. The exact p-value was 0.0156, indicating a statistically significant paired difference for this test set.

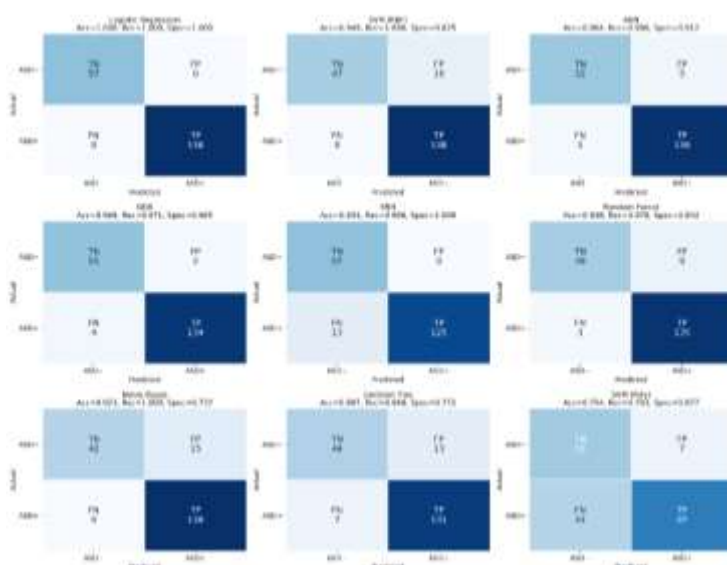


Figure 3. Held-out confusion matrices for all nine production models (n = 195). Each panel reports TN, FP, FN, and TP counts together with accuracy, recall, and specificity.

Table 5. Confusion-matrix counts on the held-out test set.

Model	TN	FP	FN	TP
Logistic Regression	57	0	0	138
SVM (RBF)	47	10	0	138
MLP-ANN	52	5	2	136
QDA	55	2	4	134
KNN	57	0	13	125
Random Forest	48	9	3	135
Naive Bayes	42	15	0	138
Decision Tree	44	13	7	131
SVM (Poly)	50	7	41	97

D. ROC discrimination and probability loss

Figure 4 shows very strong ranking performance for most classifiers. Logistic Regression achieved ROC-AUC 1.0000, followed by SVM-RBF (0.9986), MLP-ANN (0.9968), QDA (0.9963), and KNN (0.9962). Random Forest and Naive Bayes also exceeded 0.98, whereas Decision Tree and SVM-Poly were lower.

Log loss differentiated models whose ROC-AUC values were similarly high. Logistic Regression produced the smallest loss (approximately 0.0272), followed by SVM-RBF and QDA. The result indicates confident probability estimates on this particular internal test set, but it should not be interpreted as external clinical calibration because the endpoint itself is deterministic and the cohort is not population representative.

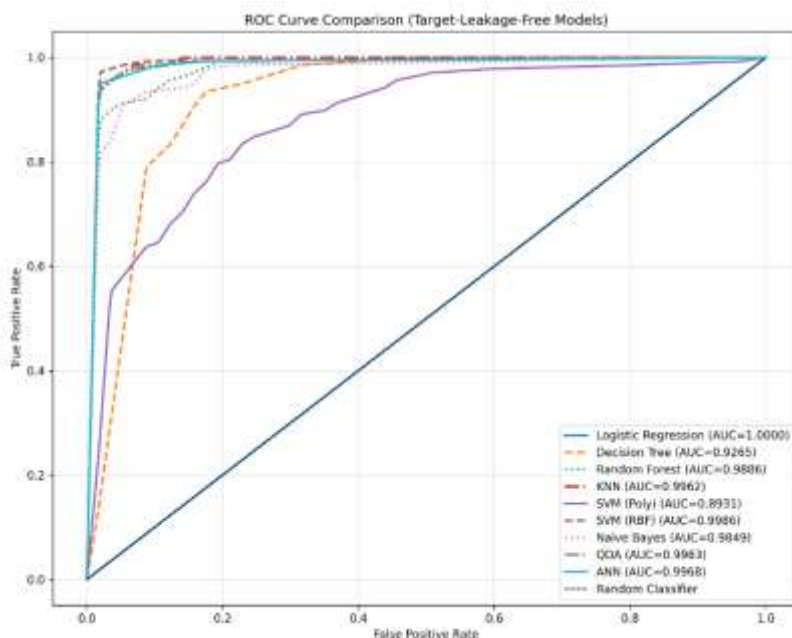


Figure 4. ROC curves for all nine production models on the held-out test set.

E. Train-test gap patterns

Descriptive train-test gaps were small for most models: 0.0000 for Logistic Regression, 0.0136 for QDA, 0.0169 for KNN, 0.0241 for the ANN, and larger values for Random Forest, Decision Tree, and especially SVM-Poly. These gaps suggest no extreme memorization pattern in the production benchmark, while remaining only an internal diagnostic.

F. Logistic Regression coefficient profile

The standardized Logistic Regression coefficients provides a mechanistic check. A1-A10 had much larger positive coefficients than the demographic and history variables, while most non-questionnaire coefficients were comparatively small. Figure 5 therefore confirms that the fitted linear boundary is driven primarily by the behavioral items from which the target score is constructed.

This coefficient pattern is consistent with reconstruction of the threshold on the A1-A10 sum. It also explains why a simple linear model can outperform more flexible alternatives on this endpoint.

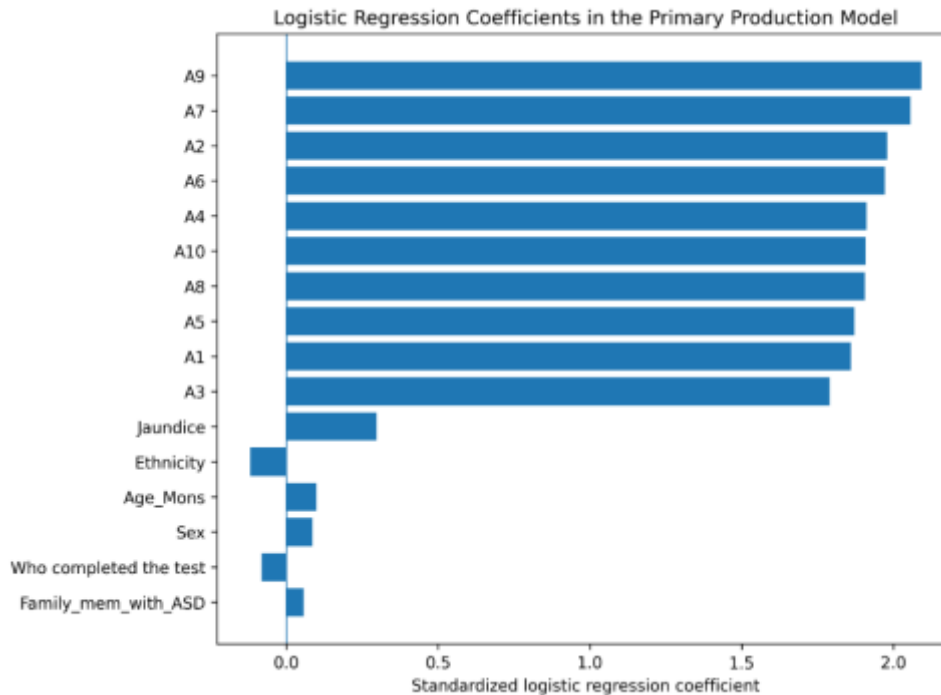


Figure 5. Standardized Logistic Regression coefficients. The ten questionnaire items dominate the fitted decision function.

G. SMOTE ablation

The SMOTE ablation showed model-dependent effects. Logistic Regression was unchanged with and without SMOTE, which is consistent with the strong linear separability of the retained A-items. Naive Bayes recall increased from 0.9565 to 1.0000, while QDA, SVM-RBF, and the MLP-ANN showed modest gains in selected metrics. These improvements indicate that balancing the training set can change the fitted boundary in ways that favor minority-class recognition for some estimators.

SMOTE did not improve every model: Random Forest recall and F1 decreased slightly, and changes for some classifiers were negligible. Table 6 therefore supports the conclusion that oversampling can alter the decision boundary beneficially or adversely depending on the estimator. Because the minority class in this dataset is screening-negative rather than screening-positive, the effect of balancing should also be interpreted in terms of specificity and unnecessary positive flags, not only recall of ASD-trait cases.

Table 6. SMOTE ablation. “S” denotes the SMOTE-balanced production pipeline; effects are model-dependent.

Model	Recall no-S	Recall S	F1 no-S	F1 S	AUC no-S	AUC S
Logistic Regression	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Decision Tree	0.9420	0.9420	0.9353	0.9286	0.9240	0.9359

Random Forest	0.9855	0.9710	0.9645	0.9537	0.9921	0.9915
KNN	0.9710	0.9130	0.9710	0.9545	0.9936	0.9942
SVM (RBF)	1.0000	1.0000	0.9550	0.9650	0.9963	0.9986
Naive Bayes	0.9565	1.0000	0.9496	0.9485	0.9888	0.9849
QDA	0.9638	0.9710	0.9673	0.9781	0.9963	0.9963
MLP-ANN	0.9855	0.9855	0.9680	0.9749	0.9896	0.9968

H. Secondary cross-validation implementation audit

The repository's 10-fold output reported very high mean validation scores for several models, including perfect Logistic Regression results. However, these values were computed after SMOTE and scaling had already been applied to the full training partition. Because this allows preprocessing information to cross fold boundaries, the cross-validation results are retained only as implementation diagnostics and are not used to strengthen the held-out performance claim.

I. Deployment and XAI verification

The feature order, categorical encoders, scaler, classical model dictionary, ANN object, evaluation table, and ROC data were successfully loaded before inference in the Streamlit application during deployment. The current interface also provides a downloadable PDF screening report, a batch CSV workflow, explanation tabs for SHAP and LIME, a single- or multi-feature what-if analysis, and model-comparison views with single or grid confusion matrices. Reusing the serialized pre-processing artefacts connects the web interface with the trained models and decreases the training-serving skew. The descriptive EDA can show the original Qchat-10-Score but this variable is never added back to the predictive input matrix.

Web output is not a calibrated clinical probability of autism, but rather a research screening estimate. A disclaimer is included in the PDF report that is not diagnostic, and the paper's report, therefore, does not refer to it as a diagnosis. The batch page also needs to be used carefully, because the current one is based on numeric or encoded features. These limitations further emphasize that the software is a research demonstrator, and not a clinical information system.

J. Explainable AI verification on a valid held-out screening record

The XAI workflow was evaluated on an actual held-out record with valid binary questionnaire responses. Logistic Regression produced a model-estimated screening-positive probability of approximately 0.903. Using the selected SHAP reference set, the explanation moved from an expected output of about 0.431 to the case-specific prediction; A6, A1, A5, and A7 contributed most strongly in the positive direction, while A10 contributed most strongly in the negative direction. The case was chosen to demonstrate the mechanics of explanation on a valid observation, not because it represents a typical or clinically definitive ASD profile. Its value is illustrative: the local attribution shows how the trained model combines the same behavioral items that define the underlying screening rule.

Figure 6 shows that the explanation is consistent with the screening rule: four positively scored behavioral items push the prediction across the screening-positive boundary. The figure is an illustrative local explanation, not a global ranking of clinical importance.

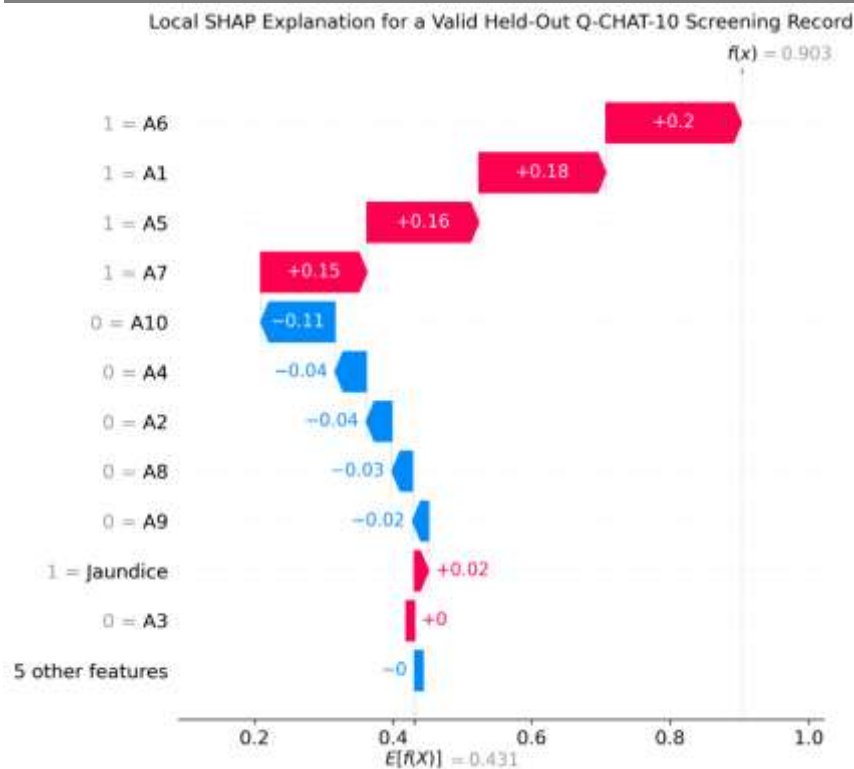


Figure 6. Local SHAP waterfall for a valid held-out Q-CHAT-10 record classified by Logistic Regression. The expected model output (0.431) is shifted to a model-estimated screening-positive probability of 0.903; attributions describe model behavior for this record and are not causal effects.

K. LIME explanation with native-unit presentation

To complement the SHAP attribution analysis, LIME was used to provide a model-agnostic local explanation of an individual screening output. The underlying classifier continues to operate on the standardized feature representation used during model training; however, the LIME conditions presented to the user were mapped back to the original feature domain using the fitted preprocessing parameters. Consequently, the ten Q-CHAT-10 behavioral items are expressed using their valid binary values (0 or 1), age is reported in months, and categorical variables are presented using their original labels rather than standardized numerical values.

In the illustrative LIME explanation, an age greater than approximately 33.1 months contributed positively to the local surrogate for the screening-positive output. Several questionnaire responses with values of zero, including A4, A5, A3, A6, A8, A10, A7, and A9, contributed in the opposite direction, while the recorded history of jaundice also had a negative local contribution. These values describe the behavior of the locally fitted LIME surrogate around this particular observation and should not be interpreted as estimates of clinical risk, symptom importance, protective effects, or causal relationships with ASD.

The native-unit representation improves interpretability because the explanation can be read in the same measurement domain in which the questionnaire is completed. Nevertheless, LIME remains a local approximation of the fitted model, and its feature weights may vary according to the sampled neighborhood, discretization strategy, and perturbation settings.

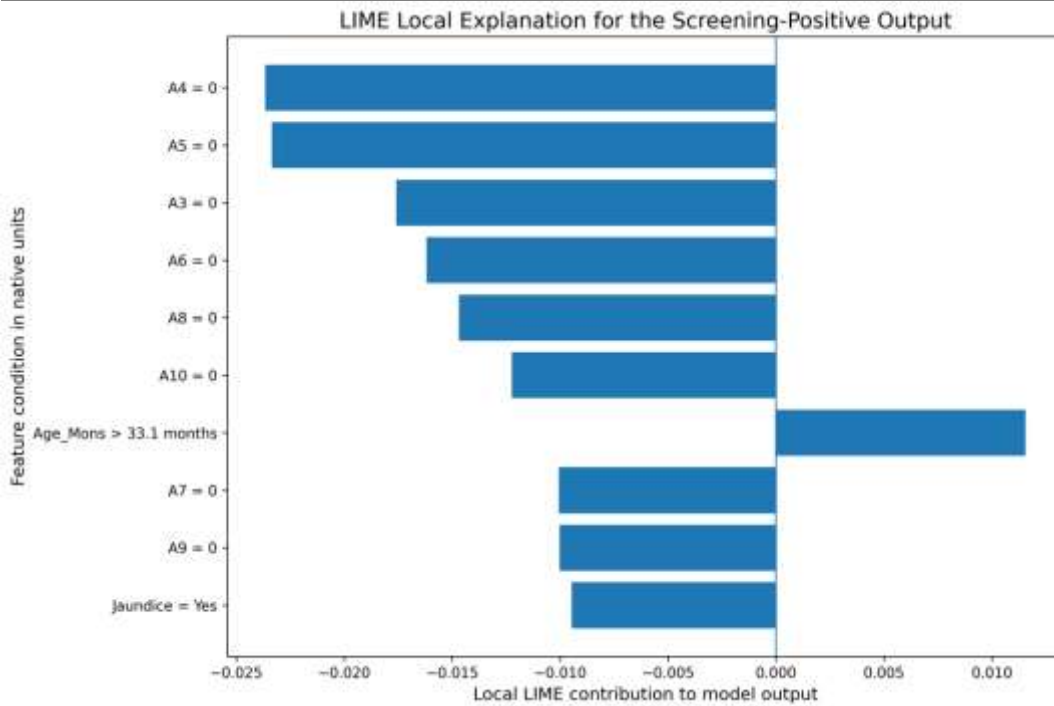


Figure 7. LIME local explanation of a Q-CHAT-10 screening output presented in native feature units. Q-CHAT-10 behavioral items are displayed using their valid binary values, age is expressed in months, and categorical information is mapped to its original label.

L. Learning-curve robustness analysis

Figure 8 shows the fold-local resampling/scaling learning curves. Logistic Regression reached perfect mean cross-validation accuracy once the training size approached roughly 360 observations and remained stable thereafter, which is consistent with the deterministic additive endpoint. MLP-ANN improved progressively as sample size increased, with mean cross-validation accuracy rising to approximately 0.97 at the largest training size. Random Forest remained comparatively stable around the mid-0.95 range and retained a small train-validation gap. These patterns support the interpretation that the linear rule is readily learnable, while the more flexible models require more observations and do not gain a clear advantage on this endpoint.

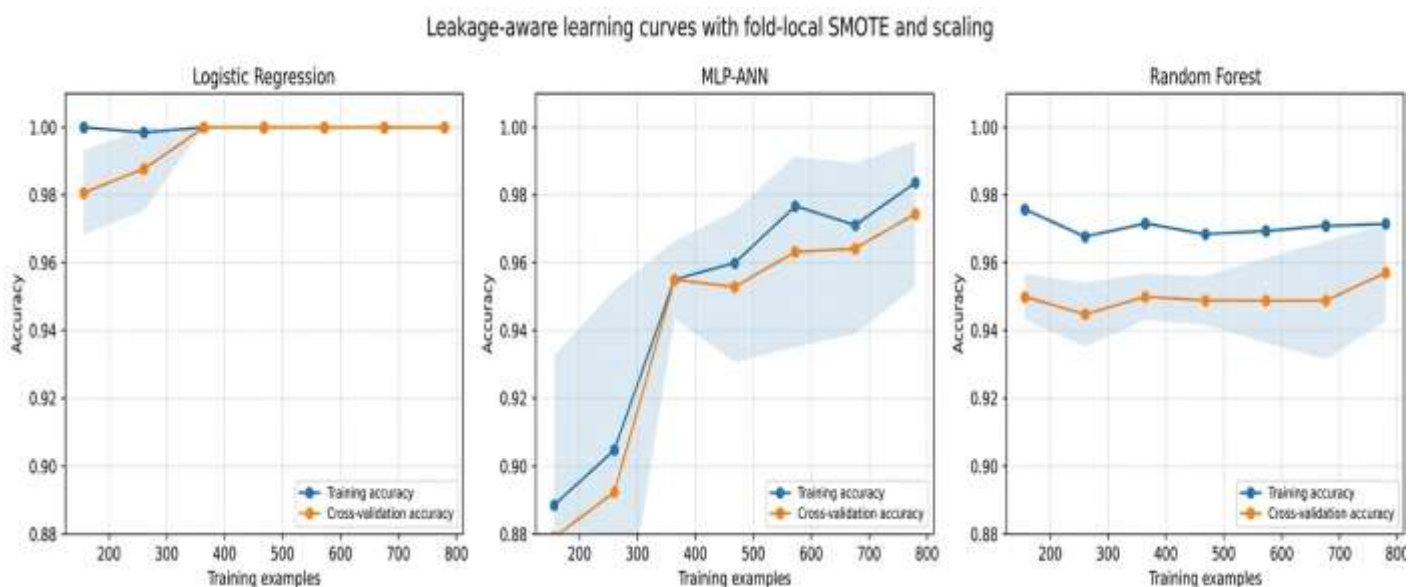


Figure 8. Leakage-aware learning curves for Logistic Regression, MLP-ANN, and Random Forest using five-fold stratified cross-validation with fold-local SMOTE and scaling. Shaded bands indicate between-fold variability.

DISCUSSION

A. Principal interpretation

The 100% accuracy number should be interpreted in the context of the endpoint audit. The problem Logistic Regression is trying to solve is a rule-reconstruction problem that is mostly linear, since Class/ASD Traits is just the thresholded sum of A1-A10. It is not a perfect autism detection tool, but rather a tool that reproduces the screening rule of the dataset to the degree that is possible without error. This distinction also accounts for the fact that the simplest model performed better than several more flexible models: the structure to be learned is rather simple after the explicit total score is removed, since the component items still carry the same additive information that was used to construct the label.

This interpretation is consistent with the previous questionnaire-based studies that reported very high internal accuracy [20] and with the more informative external validation studies of Sollis et al. that demonstrated poorer performance for transfer of QCHAT-derived models to clinician-determined diagnoses [22]. The contrast highlights the importance of endpoint provenance in the assessment of medical machine learning, alongside model choice.

In this context, words like 'prediction' and 'early detection' are to be understood in the screening sense. The system is able to predict the ASD-trait screening classification of the dataset, and can be used to assist with triage or research demonstration, but not to diagnose without separate clinical evaluation.

B. Why a simple linear model outperformed more complex alternatives

The model hierarchy also shows that the complexity should follow the structure of the problem. Logistic Regression is very appropriate for an additive threshold since a weighted linear combination can directly represent the underlying rule. The boundary is also well approximated by QDA, MLP-ANN and SVM-RBF, and these methods provide some extra flexibility that is not needed for this endpoint and may lead to a slight loss of finite-sample performance.

SVM-Poly was the least successful of the nine models, suggesting that the quadratic feature geometry and regularization of this model did not fit the data. It was observed that KNN and Naive Bayes had distinct sensitivity-specificity trade-offs, highlighting the fact that a near-perfect ROC-AUC does not necessarily mean that they will have the same classification behavior at the chosen decision threshold.

The results from the separate GridSearchCV objects are not suitable for describing the estimators that were used. Separating exploratory tuning from production models helps maintain reproducibility and avoids a common reporting mistake of showing the best cross validation model as the final model that was serialized.

C. Sensitivity, specificity, and the screening context

For screening, false negatives are often considered more consequential because they may delay referral, but specificity also matters because excessive false positives create anxiety, follow-up burden, and unnecessary specialist demand. The present test set makes these trade-offs visible: SVM-RBF and Naive Bayes achieved perfect recall with lower specificity, while KNN achieved perfect specificity with lower recall. Logistic Regression made no errors on this internal endpoint. In a future clinically diagnosed cohort, the preferred operating threshold should therefore be selected according to the intended screening setting and the relative consequences of missed cases and over-referral rather than by accuracy alone.

Even a zero-error result is statistically bounded by sample size. The held-out set contains 138 positive and 57 negative observations, so uncertainty would remain for new populations even if the endpoint construction were unchanged. More importantly, population-level performance will depend on prevalence, response distributions, and whether the outcome is a screening rule or a clinical diagnosis.

D. Class imbalance and the nuanced role of SMOTE

Screening-positive observations are the majority in this dataset, so SMOTE balances the smaller screening-negative class. The ablation demonstrates that this adjustment is not uniformly beneficial: some models improved slightly, Logistic Regression was unchanged, and Random Forest deteriorated modestly.

SMOTE should therefore be described as a tested design choice rather than an essential ingredient. Resampling changes the effective training distribution and can alter both sensitivity and specificity; its value must be evaluated for each estimator on an untouched test set.

A further limitation is the use of standard SMOTE after integer encoding of nominal categories. Interpolation can create synthetic points that do not correspond to meaningful categorical states. Categorical-aware methods, class weighting, or pipelines that preserve variable type should be preferred in future clinical validation.

E. Cross-validation and leakage discipline

The primary held-out design has an important strength: the test set is created before SMOTE and scaler fitting, so test observations do not influence these transformations. This makes the 195-record test results the most defensible estimate within the available dataset.

The auxiliary 10-fold analysis is weaker because SMOTE and scaling are performed before fold construction. Synthetic observations and global scaling statistics can therefore leak information across folds, inflating validation scores. A future analysis should use an imbalanced-learn Pipeline so that resampling and scaling are refitted independently within each training fold.

The same principle applies to imputation and categorical encoding, which are currently fitted before the initial split. Although this is less consequential than explicit target leakage, a fully leakage-disciplined pipeline would learn all preprocessing parameters from training data only and apply them unchanged to validation or test partitions.

F. External validity and the meaning of “clinical relevance”

Internal accuracy is not enough to infer clinical relevance. The data is a convenience screening sample, the outcome is generated by the questionnaire itself, and there is no clinician diagnosis, follow-up or external site. Thus, the generalizability of the results to community screening or clinical populations is not known. Prevalence, language, culture, caregiver interpretation, referral patterns and developmental age may affect the external validity. A model that accurately mimics the scoring perfectly in one public data set may be of limited value if the score itself is not optimally valid for subsequent clinical diagnosis.

In cases where the operational goal is just to replicate the thresholded Q-CHAT score, the explicit scoring rule is already clear and easy to compute. The scientific value of machine learning is enhanced by its ability to decrease item burden, predict independent outcomes, integrate complementary modalities, or facilitate calibrated risk estimates that extend beyond the original deterministic rule.

Future external studies should therefore compare models to clinician-confirmed diagnosis and should compare to the unmodified Q-CHAT-10 scoring rule. Only then can incremental value, calibration, clinical utility and transportability be determined.

G. Fairness and subgroup performance

The dataset contains demographic and history variables, but the current study does not provide a sufficiently powered subgroup fairness analysis. High aggregate performance can conceal unequal false-negative or false-positive rates across sex, age, ethnicity, language, or respondent type. Fairness evaluation should therefore be prespecified in larger external cohorts.

Ethnicity and other nominal variables are currently represented by arbitrary integer codes, which is not ideal for distance-based modelling or subgroup interpretation. Future pipelines should use appropriate categorical encoding and report subgroup counts before drawing fairness conclusions.

H. Explainable AI, local attribution, and clinical interpretation

Interpretability is provided at complementary levels. Logistic Regression coefficients provide a global linear weighting of the model, and SHAP provides an explanation of how each feature influences a chosen prediction, compared to a reference prediction [16]. As in the known construction of the endpoint, the behavioral items dominated the local explanation in the valid held-out example. The agreement with the coefficient profile and the SHAP case is good for implementation, as it means that the explanation layer reproduces the fitted classifier. The result should not, however, be interpreted as a clinical ranking of the symptoms, as a large model attribution does not indicate that the feature is more important to the autism pathophysiology.

The what-if simulator also provides local sensitivity analysis as it changes one valid input to the model and computes the model probability again. This kind of tool can be useful to the user to understand the behavior of the model but should not be treated as a treatment-effect or causal simulator.

LIME provides a complementary local perspective by approximating the fitted classifier within the neighborhood of an individual observation [19]. Presenting the explanation in native questionnaire units improves readability because the conditions correspond directly to valid screening responses rather than standardized numerical thresholds. Concordance between LIME and SHAP may be useful as an implementation-consistency check; however, agreement between two post-hoc explanation methods does not constitute independent clinical validation. Both approaches describe how the fitted model behaves and do not establish biological importance, diagnostic validity, or causal relationships between individual features and ASD.

I. Deployment strengths and safeguards

The Streamlit deployment is a practical contribution because it translates the training pipeline into a reproducible interface that reuses the saved feature order, encoders, scaler, and model artefacts. It exposes not only single-record screening predictions but also interactive EDA, SHAP/LIME explanations, all-model confusion matrices, downloadable screening reports, batch inference, and what-if sensitivity analysis. This broadens the system from a prediction demo into a transparent research workflow in which data structure, model performance, individual explanations, and scenario sensitivity can be inspected within one application.

Deployment also increases the need for safeguards. The interface should clearly distinguish screening from diagnosis, restrict inputs to valid questionnaire values, avoid causal wording in explanations, and direct users to qualified professionals when concerns arise. Clinical deployment would additionally require privacy, security, governance, monitoring, and regulatory review.

J. Reference integrity and scientific reporting

Scientific reporting also depends on accurate references and publication status. The bibliography was checked against publisher or DOI information, and evidence was interpreted in context rather than ranked solely by headline accuracy. This is particularly important in rapidly evolving AI literature, where retractions, corrections, or expressions of concern can change the evidential value of a source.

Prevalence statements were likewise reported with population and time qualifiers. The WHO global estimate and the CDC ADDM estimate refer to different populations and ascertainment systems [9], [26]; presenting them as interchangeable would overstate comparability.

K. Strengths of the study

Key strengths include a common nine-model hold-out benchmark, explicit exclusion of the pre-computed score, an audited reconciliation of dataset counts and endpoint construction, paired statistical testing, SMOTE ablation, separation of exploratory tuning from deployed models, serialized preprocessing artefacts, and integrated

SHAP/LIME explanation. The extended application adds valid-domain questionnaire controls, interactive EDA, all-model confusion matrices, what-if sensitivity analysis, downloadable screening reports, and batch inference. The availability of executable code allowed inconsistencies in narrative documentation to be identified rather than propagated. The study also retains negative methodological findings, including model-dependent SMOTE effects and leakage risk in auxiliary validation utilities, which strengthens transparency even where it complicates the performance narrative.

LIMITATIONS

The study has some notable limitations, however. Most importantly, the target is a deterministic Q-CHAT-derived screening label and not an independent clinical diagnosis, and the analysis is performed on a single public convenience dataset, without prospective or external validation. However, the pre-processing stage before the split is not necessarily fold-local, and the auxiliary 10-fold analysis in the repository is optimistic because SMOTE and scaling are performed before the folds are made. After integer encoding of nominal variables, standard SMOTE is also performed. The sample size of the held out sample is limited, there is no external testing of probability calibration, no subgroup fairness analysis and no usability, privacy, security, or clinical-impact evaluation of the Web application. The current batch utility requires numeric or already encoded inputs, and the LIME explanations remain sensitive to the local perturbation distribution, discretization procedure, kernel width, sampling density, and choice of reference neighborhood. Accordingly, explanation stability and clinician interpretability should be evaluated systematically before the XAI interface is used in clinical-facing settings. Furthermore, the class distribution is not a representative one for screening in the general population. Lastly, SHAP and LIME are post hoc explanations that rely on reference data and perturbation selections and are not causal explanations or clinically proven decision aids.

FUTURE RESEARCH

A more robust follow-up should be based on a clinician-confirmed external dataset and a prespecified pipeline, where imputation and categorical encoding and resampling/class weighting and scaling are performed in each training fold. Model selection should be done in a manner that the selected model should be tested within the model selection group and an external cohort that has not been touched should be used for the final evaluation. The deterministic Q-CHAT score should be kept as a clear baseline, and any machine-learning model should prove to be of incrementally greater value.

Evaluation should not be limited to discrimination but also include calibration plots, Brier score, PPV and NPV at realistic prevalence and decision-curve analysis. Subgroup analyses by sex, age, ethnicity, language, and respondent type should be provided for larger cohorts. The incremental value of adding speech, gaze, motion, EEG, or imaging features should be balanced against the increased acquisition cost and privacy concerns [2], [7], [8], [12], [17], [27]. XAI should not only be evaluated with respect to visual plausibility, but also for stability, concordance, and clinician usability.

Another interesting pathway is feature reduction. Reducing the number of items in the subset while maintaining sensitivity to independent diagnoses would alleviate caregiver burden, and would be consistent with previous compact-screening studies [1] and [22]. Feature selection must be done only during the training folds and validated in a prospective manner.

Human factors testing and governance should be part of any move towards clinical deployment. The user should be aware of the meaning of the probability, when to repeat screening, what to do if a screening is positive and how to deal with errors. Formal monitoring would be required for model versions, input schemas, drift and for post-deployment performance.

Table 7. Contextual comparison with selected verified literature. Metrics are not ranked across studies because datasets, labels, and modalities differ.

Study	Data/endpoint	Approach	Reported lesson	Relevance to present study
Achenie et al. [1]	M-CHAT-R archival toddler data	ML item reduction	Strong screening classification; reduced-item objective	Shows value of ML when reducing screening burden
Ahmed et al. [2]	Eye tracking	ML/DL	High classification performance	Independent behavioural modality but requires acquisition hardware
Chen et al. [11]	Medical claims; large cohort	LR/RF	AUROC below questionnaire-derived near-perfect scores	Endpoint is later ASD diagnosis; more clinically independent predictors
Hossain et al. [14]	Public ASD questionnaire datasets	Multiple ML models	High performance varies by cohort	Supports multi-model benchmarking but labels remain dataset-dependent
Luongo et al. [17]	Game-based motion trajectories	ANN / motion analysis	Useful group classification	Adds motor information independent of questionnaire sum
Shambour et al. [20]	Non-clinical toddler Q-CHAT-style data	Six classical models	LR reported 100%	Converges with current LR result; circularity must be considered
Sollis et al. [22]	QCHAT-derived training + clinical external set	Feature selection + ML	Clinical transfer AUROC materially below assessment prediction	Direct evidence that score reconstruction does not equal clinical diagnosis
Sturner et al. [23]	18-month community screening + diagnostic evaluation	Q-CHAT/M-CHAT validation	Sensitivity/specificity trade-offs	Provides clinical benchmark for screening interpretation
Yin et al. [27]	Longitudinal structural MRI	Siamese DL + path signatures	Promising early classification	Independent modality with substantially greater acquisition burden

CONCLUSION

In this study, an early ASD screening framework was developed involving nine machine-learning/ANN classifiers, leakage-aware Q-CHAT-10 preprocessing, statistical evaluation, explainable AI, and Streamlit deployment. Once duplicates and the identifier were removed there were 975 unique records. Logistic Regression achieved zero error on the 195-record held-out test set (ROC-AUC 1.0000), and MLP-ANN, QDA and SVM-RBF also performed well, with McNemar's exact test preferring Logistic Regression over the ANN ($p=0.0156$). The SMOTE effects depended on the models and were not always positive.

The key interpretive finding is that Qchat-10-Score is the same as A1-A10 and Class/ASD Traits is the same as the threshold rule score ≥ 4 . The benchmark thus represents a reconstruction of the questionnaire-derived screening classification rather than independent clinical diagnosis. Logistic Regression is particularly effective because its linear additive model is very similar to this rule.

The deployed system reuses serialized preprocessing and models and introduces SHAP, LIME, valid domain what-if analysis, interactive EDA, all model comparison views, downloadable screening reports and prototype

batch inference. These functions further facilitate inspection of the research pipeline and make it practically reproducible, but external clinician confirmation, fold-local preprocessing, categorical-aware imbalance handling, calibration, fairness analysis, and evaluation of explanation stability still need to be done prior to clinical use. The system should thus be considered as a prototype of a research screening system, not as a diagnostic medical device.

REFERENCES

1. Achenie, L. E. K., Scarpa, A., Factor, R. S., Wang, T., Robins, D. L., & McCrickard, D. S. (2019). A machine learning strategy for autism screening in toddlers. *Journal of Developmental & Behavioral Pediatrics*, 40(5), 369-376. <https://doi.org/10.1097/DBP.0000000000000668>
2. Ahmed, I. A., Senan, E. M., Rassem, T. H., Ali, M. A. H., Shatnawi, H. S. A., Alwazer, S. M., & Alshahrani, M. (2022). Eye tracking-based diagnosis and early detection of autism spectrum disorder using machine learning and deep learning techniques. *Electronics*, 11(4), 530. <https://doi.org/10.3390/electronics11040530>
3. Allison, C., Baron-Cohen, S., Wheelwright, S., Charman, T., Richler, J., Pasco, G., & Brayne, C. (2008). The Q-CHAT (Quantitative CHECKlist for Autism in Toddlers): A normally distributed quantitative measure of autistic traits at 18-24 months of age: Preliminary report. *Journal of Autism and Developmental Disorders*, 38(8), 1414-1425. <https://doi.org/10.1007/s10803-007-0509-7>
4. Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief 'red flags' for autism screening: The Short Autism Spectrum Quotient and the Short Quantitative Checklist for Autism in Toddlers in 1,000 cases and 3,000 controls [corrected]. *Journal of the American Academy of Child & Adolescent Psychiatry*, 51(2), 202-212.e7. <https://doi.org/10.1016/j.jaac.2011.11.003>
5. Allison, C., Matthews, F. E., Ruta, L., Pasco, G., Soufer, R., Brayne, C., Charman, T., & Baron-Cohen, S. (2021). Quantitative Checklist for Autism in Toddlers (Q-CHAT). A population screening study with follow-up: The case for multiple time-point screening for autism. *BMJ Paediatrics Open*, 5(1), e000700. <https://doi.org/10.1136/bmjpo-2020-000700>
6. Alzakari, S. A., Allinjaw, A., Aldrees, A., Zamzami, N., Umer, M., Innab, N., & Ashraf, I. (2025). Early detection of autism spectrum disorder using explainable AI and optimized teaching strategies. *Journal of Neuroscience Methods*, 413, 110315. <https://doi.org/10.1016/j.jneumeth.2024.110315>
7. Assaf, R., Shehabeddine, Z., & Ramesh, V. (2025). Screening autism spectrum disorder in children using machine learning on speech transcripts. *Scientific Reports*, 15, 34134. <https://doi.org/10.1038/s41598-025-01500-6>
8. Bosl, W. J., Tager-Flusberg, H., & Nelson, C. A. (2018). EEG analytics for early detection of autism spectrum disorder: A data-driven approach. *Scientific Reports*, 8, 6828. <https://doi.org/10.1038/s41598-018-24318-x>
9. Centers for Disease Control and Prevention. (2025). Data and statistics on autism spectrum disorder: ADDM Network 2022. <https://www.cdc.gov/autism/data-research/index.html>
10. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357. <https://doi.org/10.1613/jair.953>
11. Chen, Y.-H., Chen, Q., Kong, L., & Liu, G. (2022). Early detection of autism spectrum disorder in young children with machine learning using medical claims data. *BMJ Health & Care Informatics*, 29(1), e100544. <https://doi.org/10.1136/bmjhci-2022-100544>
12. Ganai, U. J., Ratne, A., Bhushan, B., & Venkatesh, K. S. (2025). Early detection of autism spectrum disorder: Gait deviations and machine learning. *Scientific Reports*, 15, 873. <https://doi.org/10.1038/s41598-025-85348-w>
13. Ghazal, T. M., Munir, S., Abbas, S., Athar, A., Alrababah, H., & Khan, M. A. (2023). Early detection of autism in children using transfer learning. *Intelligent Automation & Soft Computing*, 36(1), 11-22. <https://doi.org/10.32604/iasc.2023.030125>
14. Hossain, M. D., Kabir, M. A., Anwar, A., & Islam, M. Z. (2021). Detecting autism spectrum disorder using machine learning techniques: An experimental analysis on toddler, child, adolescent and adult datasets. *Health Information Science and Systems*, 9(1), 17. <https://doi.org/10.1007/s13755-021-00145-9>

15. Jacob, S. G., Sulaiman, M. M. B. A., & Bennet, B. (2023). Feature signature discovery for autism detection: An automated machine learning based feature ranking framework. *Computational Intelligence and Neuroscience*, 2023, 6330002. <https://doi.org/10.1155/2023/6330002>
16. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774. <https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
17. Luongo, M., Simeoli, R., Marocco, D., Milano, N., & Ponticorvo, M. (2024). Enhancing early autism diagnosis through machine learning: Exploring raw motion data for classification. *PLOS ONE*, 19(4), e0302238. <https://doi.org/10.1371/journal.pone.0302238>
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
19. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why Should I Trust You?': Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144. <https://doi.org/10.1145/2939672.2939778>
20. Shambour, Q., Qandeel, N., Alrabanah, Y., Abumariam, A., & Shambour, M. K. (2024). Artificial intelligence techniques for early autism detection in toddlers: A comparative analysis. *Journal of Applied Data Sciences*, 5(4), 1754-1764. <https://doi.org/10.47738/jads.v5i4.353>
21. Simeoli, R., Rega, A., Cerasuolo, M., Nappo, R., & Marocco, D. (2026). Using machine learning for motion analysis to early detect autism spectrum disorder: A systematic review. *Review Journal of Autism and Developmental Disorders*, 13, 278-297. <https://doi.org/10.1007/s40489-024-00435-4> (published online 2024)
22. Sollis, L. J., Wall, D. P., & Washington, P. Y. (2025). Compact subsets of autism screening items predict clinical diagnoses with a machine learning analysis of the QCHAT-10. *Scientific Reports*, 15, 39091. <https://doi.org/10.1038/s41598-025-26131-9>
23. Sturmer, R., Howard, B., Bergmann, P., Attar, S., Stewart-Artz, L., Bet, K., Allison, C., & Baron-Cohen, S. (2022). Autism screening at 18 months of age: A comparison of the Q-CHAT-10 and M-CHAT screeners. *Molecular Autism*, 13, 2. <https://doi.org/10.1186/s13229-021-00480-4>
24. Tan, S. (2024). A DNN-based diagnosis on autism spectrum disorder in children. *Applied and Computational Engineering*, 67, 13-20. <https://doi.org/10.54254/2755-2721/67/20240606>
25. Thabtah, F. (2018). Autism screening data for toddlers [Data set]. Kaggle. <https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers>
26. World Health Organization. (2025, September 17). Autism. <https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders>
27. Yin, Z., Ding, X., Zhang, X., Wu, Z., Wang, L., Xu, X., & Li, G. (2024). Early autism diagnosis based on path signature and Siamese unsupervised feature compressor. *Cerebral Cortex*, 34(13), 72-83. <https://doi.org/10.1093/cercor/bhae069>