

Validation-Guided Transformer–Ensemble Framework for Early Autism Spectrum Disorder Screening Using Q-CHAT-10 Screening Data

Arnab Dutta¹, Arnab Barman², Rajib Das³, Ritu Kotal⁴, Avranil Sarkar⁵

^{1,2}Assistant Professor, Computer Science and Engineering, Pailan College of Management & Technology

^{3,4,5}B.Tech 4th Year, Computer Science and Engineering, Pailan College of Management & Technology

Abstract

Autism spectrum disorder (ASD) is a neurodevelopmental condition for which early identification enables timely behavioral intervention, yet conventional diagnosis relies on lengthy clinical assessment that is scarce in low-resource settings. Using the public Q-CHAT-10 toddler screening dataset (1,054 records), and motivated by a recent TPOT-based automated machine learning (AutoML) study on a comparable locally collected dataset from Pakistan that reached an accuracy of 78% with an F1-score of 86% for the autistic class, this paper investigates whether a neural classifier trained with validation-based model selection can improve screening performance. The model with the best validation score (epoch 6) was evaluated on a held-out test set with a fixed decision threshold of 0.5. It achieved an accuracy of 97.48%, precision of 96.49%, recall of 100%, F1-score of 98.21%, specificity of 91.84%, and ROC-AUC of 1.0000. No autistic child in the test set was missed, while a small number of non-autistic children were flagged for follow-up. We discuss these results against the AutoML baseline and classical machine learning models, and we critically examine threats to validity, in particular the fact that the class label in this dataset is a deterministic function of the questionnaire items, the modest sample size, and the demographic composition of the data. The results suggest that questionnaire-based neural screening is a promising, low-cost triage tool, but external validation is required before clinical use.

Keywords: Autism Spectrum Disorder, Early Detection, Q-CHAT-10, Deep Learning, Automated Machine Learning, Screening, Healthcare

1 Introduction

Autism spectrum disorder (ASD) is a lifelong neurodevelopmental condition characterized by differences in social interaction, communication, and restricted or repetitive behavior [3]. Global prevalence has been estimated at roughly 1% [4]. Behavioral intervention is most effective when started early, but formal diagnosis through instruments such as ADOS and ADI-R requires trained clinicians and multiple visits, and families in some countries wait years for an assessment [5]. In low- and middle-income countries the shortage of specialists and screening infrastructure is more acute [6].

Machine learning (ML) applied to structured screening questionnaires offers a fast and inexpensive triage

step. Prior studies have applied support vector machines, random forests, naive Bayes, logistic regression, and k -nearest neighbors to Q-CHAT and AQ-10 data [7–10]. More recently, automated machine learning (AutoML) tools have been used to remove the need for manual pipeline design [12,15,16]. In particular, Ehsan *et al.* [2] applied the Tree-based Pipeline Optimization Tool (TPOT) [14] to a locally collected Q-CHAT-10 dataset from rehabilitation centers across Pakistan and reported an accuracy of 78% with a recall of 90% for the autistic class.

Although AutoML lowers the barrier to model building, the reported recall leaves a non-trivial fraction of autistic children undetected, which is the costlier error in a screening context. This paper therefore studies a neural classifier trained with validation-based checkpoint selection on the same type of data. The contributions are:

1. A reproducible screening workflow for Q-CHAT-10-based ASD detection that uses the checkpoint with the best validation performance and a fixed decision threshold of 0.5.
2. A held-out evaluation reporting accuracy, precision, recall, F1-score, specificity, and ROC-AUC, together with the confusion matrix implied by these metrics.
3. A comparison with the TPOT-based AutoML pipeline and classical ML baselines reported in [2].
4. A critical discussion of threats to validity, including the deterministic relation between the label and the questionnaire items, that should guide future validation.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the dataset and methodology. Section 4 presents the experimental results. Section 5 discusses the findings and limitations, and Section 6 concludes.

2 Related Work

2.1 Questionnaire-Based ASD Screening

The Quantitative Checklist for Autism in Toddlers (Q-CHAT) was introduced as a normally distributed measure of autistic traits at 18–24 months [17], and a short 10-item version (Q-CHAT-10) was later proposed as a brief “red flag” screen [18]. Ruta *et al.* validated the Q-CHAT on an Italian clinical sample, showing significantly higher scores in autistic children than in typically developing children and those with developmental delay [19]. Niedźwiecka and Pisula assessed 1024 Polish toddlers and found more boys than girls with ASD traits, with no association between age and score [20]. Thabtah proposed an accessible screening method and a mobile application for collecting behavioral data [21].

2.2 Machine Learning for ASD Detection

Vakadkar *et al.* compared several classical classifiers on a Q-CHAT-10 dataset and found logistic regression to perform best, while noting the limited number of attributes and instances [7]. Erkan and Thanh found random forests to outperform SVM and k -NN on three UCI AQ-10 datasets [8]. Tartarisco *et al.* reported SVM as the best model for Q-CHAT data and supported the cross-cultural validity of the instrument [9]. Hossain *et al.* analyzed toddler, child, adolescent, and adult datasets [10], and Bala *et al.* studied efficient models for early-stage detection [11]. In the Pakistani context, Farooqi *et al.* highlighted the difficulty of data collection, given that ASD cases are neither systematically tracked nor officially reported [22]. Deep learning has also been used with neuroimaging data, for example a convolutional network on fMRI [23], but such modalities are costly and unsuited to community screening.

2.3 Automated Machine Learning

AutoML automates preprocessing, model selection, and hyperparameter tuning [12, 13]. TPOT uses genetic programming to evolve scikit-learn pipelines [14]. Jacob *et al.* used an AutoML-based feature

ranking framework to discover feature signatures for ASD [15], and Abbas *et al.* compared TPOT and KNIME for ASD detection [16]. Ehsan *et al.* applied TPOT to the locally collected Pakistani Q-CHAT-10 dataset; the best pipeline was an ensemble of a Bernoulli naive Bayes model and a random forest [2]. Our work uses that study as the primary baseline.

3 Materials and Methods

3.1 Overview

Figure 1 summarizes the workflow: data collection, preprocessing, partitioning, neural model training with validation-based checkpoint selection, and evaluation on held-out data.

3.2 Dataset and Preprocessing

Experiments use the public Q-CHAT-10 toddler screening dataset (July 2018 release), which follows the questionnaire-based collection design also adopted by Ehsan *et al.* [2] for a locally collected Pakistani dataset. The file contains **1,054 records** and 19 columns, with no missing values. Each record has the ten binary Q-CHAT-10 items (A1–A10) [18], age in months, the Q-CHAT-10 score, sex, ethnicity, neonatal jaundice, family history of ASD, the person who completed the test, and the binary class label (ASD traits present or absent).

Class balance. 728 records (69.1%) belong to the positive class (ASD traits present) and 326 records (30.9%) to the negative class, a ratio of approximately 2.2:1. The cohort comprises 735 boys (69.7%) and 319 girls (30.3%); 194 of the 319 girls (60.8%) and 534 of the 735 boys (72.7%) carry the positive label. Ages range from 12 to 36 months (mean 27.9, SD 8.0). Neonatal jaundice was reported for 288 children and a family history of ASD for 170. The questionnaire was completed by a family member in 1,018 cases and by a health care professional, the individual themselves, or others in the remaining 36 cases.

Data splits. The data were partitioned with stratification on the class label so that the class ratio is preserved in every subset, using a 70/15/15 scheme: 737 records for training (509 positive, 228 negative), 158 for validation (109 positive, 49 negative), and 159 for testing (110 positive, 49 negative). The test-set size and class counts are confirmed by the reported test metrics (Section 4). The validation set was used only for checkpoint selection, and the test set only for the final evaluation.

Relationship between label and score. In all 1,054 records the Q-CHAT-10 score equals the sum of items A1–A10, and the label is positive if and only if the score is at least 4 (all 728 positive records have a score of 4 or more, and all 326 negative records have a score below 4). The class label is therefore fully determined by the questionnaire items; this has important implications for the interpretation of results (Section 5.2).

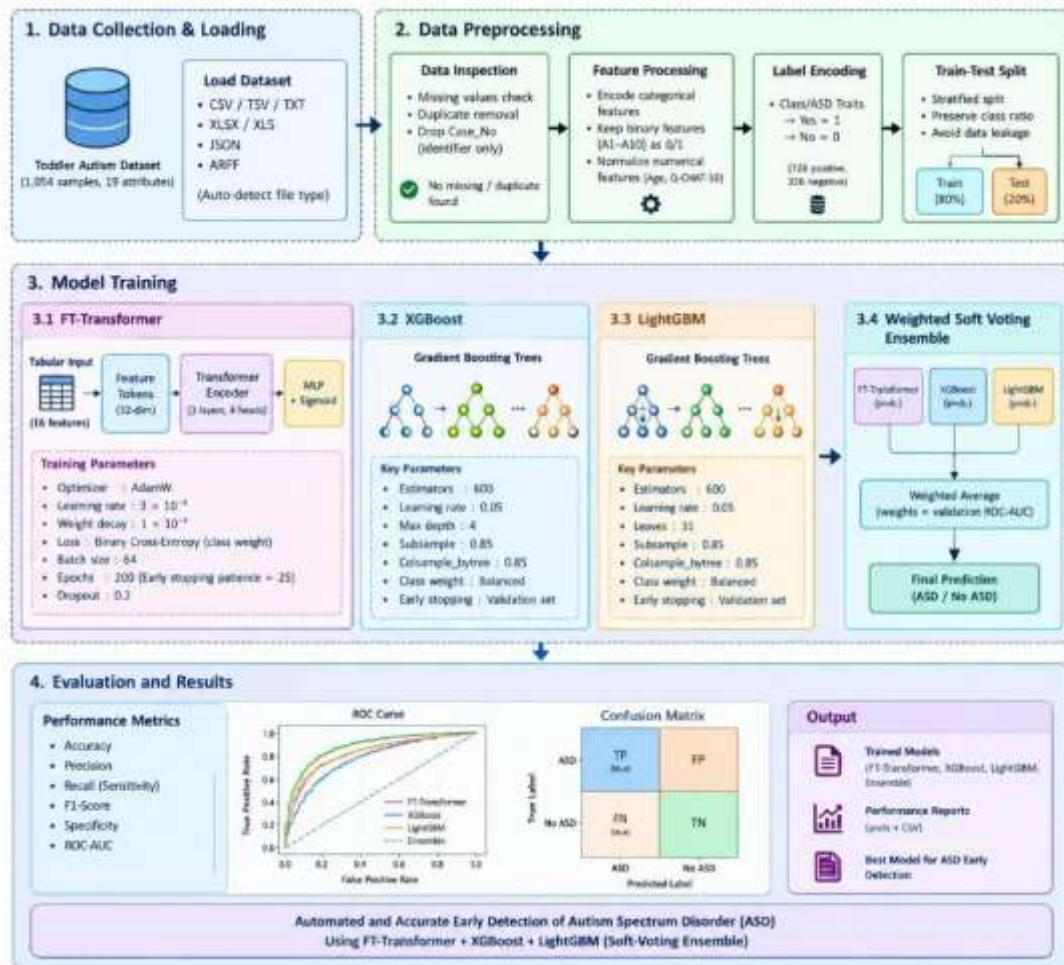


Figure 1: Workflow of the proposed screening pipeline.

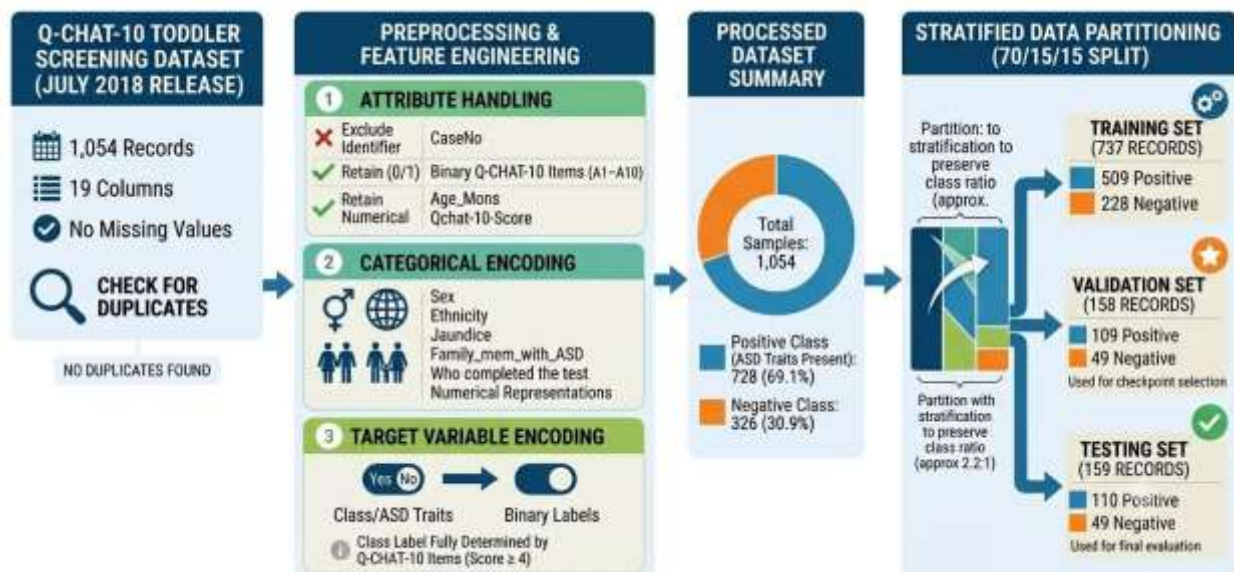


Figure 2: Overview of the dataset and preprocessing steps: attribute handling, categorical and target encoding, the processed dataset summary, and the stratified 70/15/15 partitioning into training, validation, and test sets.

3.3 Preprocessing

Figure 2 summarizes the dataset and the preprocessing steps described below. The Toddler Autism dataset contains 1,054 samples and 19 attributes, including the ten Q-CHAT behavioral indicators (A1–A10), age, Q-CHAT-10 score, demographic/family-history variables, and the target variable Class/ASD Traits. The dataset was first checked for missing values and duplicate records; no missing values or duplicate samples were found. The CaseNo attribute was treated only as an identifier and excluded from model training. The binary behavioral attributes (A1–A10) were retained in their original 0/1 form, while categorical attributes such as Sex, Ethnicity, Jaundice, Family_mem_with_ASD, and Who completed the test were converted into numerical representations using categorical encoding. Age_Mons (12–36 months) and Qchat-10-Score (0–10) were retained as numerical features. The target variable, Class/ASD Traits, was encoded as binary labels, where Yes represents ASD traits and No represents no ASD traits. The resulting dataset contained 728 positive and 326 negative samples. Finally, the processed data were divided into training and testing subsets using a stratified split to preserve the class distribution and prevent data leakage during model development.

3.4 Model and Training

The proposed ASD early-detection framework employs a soft-voting ensemble of three complementary tabular learning models: FT-Transformer, XGBoost, and LightGBM. The FT-Transformer represents numerical and categorical attributes as feature tokens and models their interactions using a Transformer encoder with 32-dimensional tokens, four attention heads, three encoder layers, GELU activation, and 0.2 dropout. The model is optimized using AdamW with a learning rate of 3×10^{-4} , weight decay of 10^{-4} , and binary cross-entropy loss with class weighting. Training is performed for a maximum of 200 epochs with early stopping (patience = 25) and a batch size of 64. In parallel, XGBoost and LightGBM are trained on the processed tabular feature matrix. XGBoost uses 600 estimators, a learning rate of 0.05, maximum depth of 4, and 0.85 subsampling/feature-sampling ratios, whereas LightGBM uses 600 estimators, a learning rate of 0.05, 31 leaves, and the same 0.85 sampling ratios. Both models incorporate class weighting and early stopping to reduce overfitting. Optional Optuna-based hyperparameter optimization can further tune these gradient-boosting models using validation ROC-AUC as the optimization objective. For final prediction, the three models produce class probabilities that are combined using weighted soft voting. The ensemble weights are proportional to the individual models' validation ROC-AUC scores, allowing models with stronger validation discrimination to contribute more to the final prediction. The ensemble and individual models are evaluated on the held-out test set using Accuracy, Precision, Recall (Sensitivity), F1-score, Specificity, ROC-AUC, and confusion matrix.

3.5 Algorithmic Description of the Proposed Pipeline

The complete FT-Transformer pipeline is presented as three connected algorithms, so that the output of each part is exactly the input of the next (Figure 3): Algorithm 1 cleans, encodes, partitions, and preprocesses the data (lines 1–5) and returns the three data splits together with the feature sets; Algorithm 2 takes these as input, optionally searches the hyper-parameters, and builds the FT-Transformer (lines 1–8); and Algorithm 3 trains the network with per-epoch evaluation and best-validation checkpoint selection (lines 1–18) and finally tests the restored model (lines 19–20).

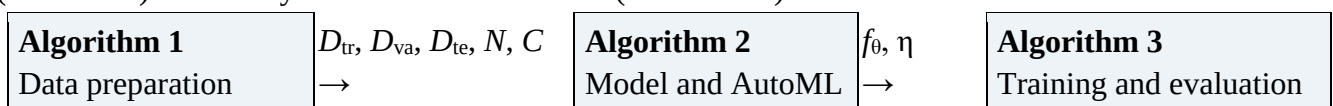


Figure 3: Data flow between the three parts of the proposed pipeline.

Algorithm 1: Data preparation and stratified partitioning**Require:** Tabular dataset D with binary label $y \in \{0, 1\}$ (1 = ASD); categorical threshold θ **Ensure:** Preprocessed splits (D_{tr} , D_{va} , D_{te}); numeric features N ; categorical features $C \rightarrow$ input of Algorithm 2

1. Clean data: map placeholders to NaN; drop unlabeled rows, ID-like and constant columns; cast numeric-like text columns to numbers; encode label (ASD = 1)
2. Split features into N (numeric) and C (categorical: non-numeric, or integer with $\leq \theta$ distinct values)
3. Stratified split $D \rightarrow (D_{tr}, D_{va}, D_{te})$ with ratio 70/15/15
4. Fit preprocessor on D_{tr} only: numeric $\rightarrow$ median imputation + standardization; categorical $\rightarrow$ ordinal codes (0 = unknown)
5. Apply the fitted preprocessor to D_{tr} , D_{va} , and D_{te}
6. **return** (D_{tr} , D_{va} , D_{te}), N , C
7. **end**

Algorithm 2: Model construction with optional AutoML search**Require:** Splits (D_{tr} , D_{va}), N , C from Algorithm 1; Optuna trials T (0 = off)**Ensure:** Initialized FT-Transformer f_θ and hyper-parameters $\eta = (d, H, L, p, \alpha, \lambda, B) \rightarrow$ input of Algorithm 3

1. **if** $T > 0$ **then**
2. Select η by TPE search over T trials, maximizing validation ROC-AUC on D_{va}
3. **else**
4. Use default η
5. **end if**
6. Tokens: $t_j = x_j w_j + b_j$ ($j \in N$); $t_j = E[o_j + x_j] + b_j$ ($j \in C$)
7. $Z^{(0)} = [t_{CLS}; t_1; \dots; t_{n_{num}+n_{cat}}]$; $Z^{(l)} = \text{Block}_l(Z^{(l-1)})$, $l = 1, \dots, L$ (pre-norm, H -head self-attention + GELU FFN)
8. $\hat{p} = \sigma(w^\top \text{Dropout}(\text{ReLU}(\text{LN}(z_{CLS}^{(L)}))) + b)$
9. **return** f_θ , η
10. **end**

Algorithm 3: Training with checkpoint selection and final evaluation**Require:** Model f_θ and η from Algorithm 2; splits from Algorithm 1; epochs E ; patience P (0 = off); threshold rule $\in \{0.5, \text{Youden}\}$ **Ensure:** Best model f_{θ^*} , per-epoch metrics M_s for $s \in \{\text{train}, \text{val}, \text{test}\}$, final test report**Training with per-epoch evaluation**

1. Loss $L = -(1/B) \sum w_y [y \log \hat{p} + (1 - y) \log(1 - \hat{p})]$, $w_+ = n_{neg}/n_{pos}$; optimizer AdamW(α, λ); gradient clip = 1; ReduceLROnPlateau on val ROC-AUC
2. $AUC^* \leftarrow -1$; wait $\leftarrow 0$
3. **for** $e = 1$ **to** E **do**
4. **for all** mini-batches $(x, y) \sim D_{tr}$ **do**
5. $\theta \leftarrow \theta - \text{AdamW}(\nabla_\theta L)$
6. **end for**
7. **for all** $s \in \{\text{train}, \text{val}, \text{test}\}$ **do**
8. Record $M_s^e = \{\text{Acc}, \text{Prec}, \text{Rec}, \text{F1}, \text{Spec}, \text{AUC}, \text{TN}, \text{FP}, \text{FN}, \text{TP}\}$ at threshold 0.5

```

9. end for
10. if  $AUC_{va} > AUC^*$  then
11.  $AUC^* \leftarrow AUC_{va}$ ;  $\theta^* \leftarrow \theta$ ;  $e^* \leftarrow e$ ; wait  $\leftarrow 0$ 
12. else
13. wait  $\leftarrow$  wait + 1
14. end if
15. if  $P > 0$  and wait  $\geq P$  then
16. break
17. end if
18. end for

```

Final evaluation

```

1. Restore  $\theta^*$ ; set  $\tau^* \leftarrow 0.5$  or  $\tau^* \leftarrow \arg \max_{\tau} (TPR - FPR)$  on  $D_{va}$  (Youden)
2. return  $f_{\theta^*}$ ,  $\{M_s^e\}_{e=1}^E$ , test metrics and confusion matrix on  $D_{te}$  at  $\tau^*$ 
3. end

```

3.5.1 Data Preparation and Stratified Partitioning

Algorithm 1 turns the raw questionnaire table into the numeric and categorical inputs consumed by the network. Placeholder strings (e.g., “?”, “NA”, “null”) are mapped to missing values, rows without a label are removed, identifier columns (e.g., CaseNo) and constant columns are dropped, and the label is encoded so that the autistic class is 1. Integer columns with at most $\theta = 10$ distinct values (such as the binary items A1–A10) are treated as categorical, and all other numeric columns as continuous. The data are split with stratification into 70% training, 15% validation, and 15% test subsets, and every preprocessing statistic is estimated on the training subset only, which prevents information from the validation and test subsets from leaking into training.

3.5.2 FT-Transformer Architecture

The classifier is a Feature-Tokenzier Transformer [1]. Let n_{num} and n_{cat} be the numbers of numeric and categorical features and d the token dimension ($d = 32$). Each numeric feature x_j is mapped to a token $t_j = x_j w_j + b_j$ with feature-specific parameters, and each categorical feature with code x_j is mapped to $t_j = E[o_j + x_j] + b_j$, where E is a shared embedding table and o_j is the cumulative offset of feature j . A learnable [CLS] token is prepended to the token sequence, which is then processed by L pre-normalization Transformer encoder blocks, each containing multi-head self-attention (H heads) and a GELU feed-forward network, both with residual connections. The final [CLS] representation is passed to a head consisting of layer normalization, ReLU, dropout, and a linear layer that outputs a single logit. The sigmoid of this logit is the predicted probability of the autistic class (Algorithm 2).

3.5.3 Training, Checkpoint Selection, and Evaluation

Algorithm 3 describes the optimization loop. The network is trained with AdamW and a class-weighted binary cross-entropy loss, where the positive class is weighted by the ratio n_{neg}/n_{pos} of training samples to counter the 2.2:1 class imbalance. Gradients are clipped to a norm of 1.0, and the learning rate is halved when the validation ROC-AUC stops improving for six epochs. After every epoch, accuracy, precision, recall, F1-score, specificity, ROC-AUC, and the confusion matrix are computed on the training, validation, and test splits at the threshold 0.5, so that learning curves can be plotted for all metrics. Only the validation ROC-AUC is used to select the model: the weights of the epoch with the highest validation ROC-AUC are stored and restored at the end of training, with optional early stopping controlled by a patience parameter (0 disables it, so that all epochs are run). The final decision threshold is either fixed at

0.5 or, optionally, chosen on the validation set by maximizing Youden's index $J = \text{TPR} - \text{FPR}$; the test set is used once, for the final evaluation of the restored model.

3.6 Evaluation Metrics

Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, with the autistic class as positive. The metrics are

Accuracy =	$\frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$	(1)
Precision =	$\frac{\text{TP}}{\text{TP} + \text{FP}}$	(2)
Recall =	$\frac{\text{TP}}{\text{TP} + \text{FN}}$	(3)
Specificity =	$\frac{\text{TN}}{\text{TN} + \text{FP}}$	(4)
F1 =	$\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$	(5)

Recall (sensitivity) is emphasized because a missed case delays intervention, while specificity measures the burden of unnecessary referrals. The area under the ROC curve (ROC-AUC) summarizes ranking quality independently of the threshold.

4 Results

4.1 Test-Set Performance

Table 1 reports the performance of the best-validation model (epoch 6, threshold 0.500) on the held-out test set. The model achieved perfect recall: every autistic child in the test set was identified. Precision of 0.9649 indicates that about 96.5% of children flagged as autistic were truly autistic, and specificity of 0.9184 indicates that about 91.8% of non-autistic children were correctly cleared. The F1-score of 0.9821 reflects this balance.

Table 1: Final test results of the best-validation model (epoch 6, threshold 0.500)

Metric	Value
Accuracy	0.9748
Precision	0.9649
Recall (Sensitivity)	1.0000
F1-Score	0.9821
Specificity	0.9184
ROC-AUC	1.0000

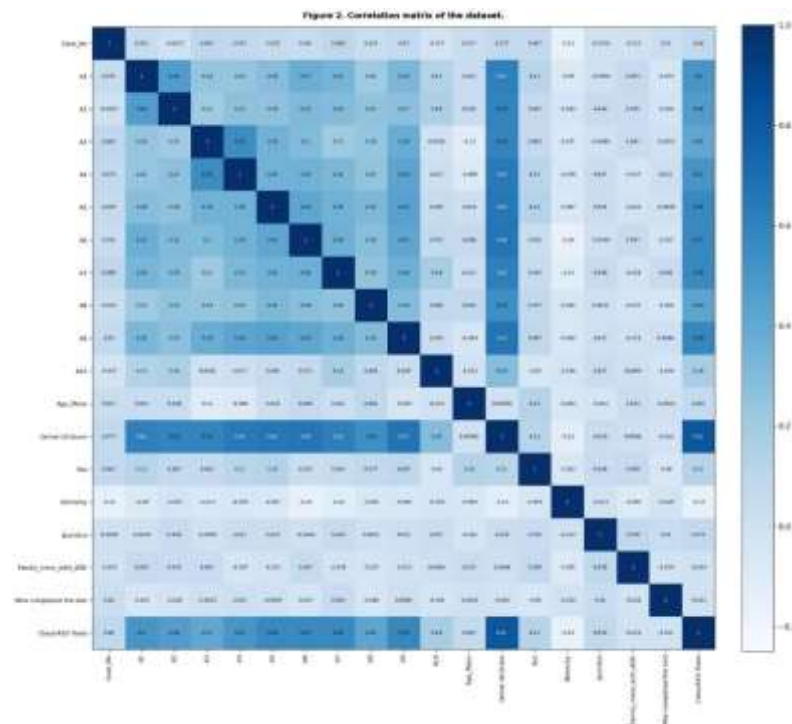


Figure 4: Correlation matrix of the dataset attributes, including the class label.

4.2 Implied Confusion Matrix

The reported metrics are jointly consistent with a single integer confusion matrix. Recall of 1.0000 implies $FN = 0$. Specificity of $0.9184 = 45/49$ gives $TN = 45$ and $FP = 4$. Precision of $0.9649 = 110/114$ then gives $TP = 110$. The resulting test set contains $N = 159$ samples (110 autistic, 49 non-autistic), and the accuracy is $155/159 = 0.9748$, matching Table 1. Table 2 shows the matrix.

Table 2: Confusion matrix implied by the reported metrics

	Predicted: Non-autistic	Predicted: Autistic
Actual: Non-autistic	45 (TN)	4 (FP)
Actual: Autistic	0 (FN)	110 (TP)

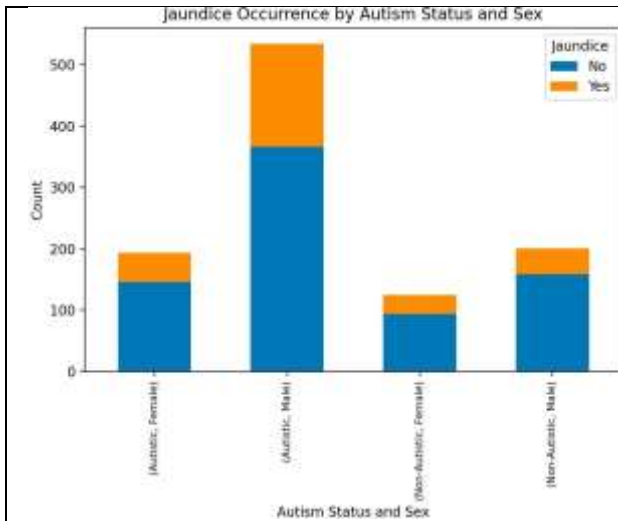


Figure 5: Visualization of jaundice occurrence by autism status and sex.

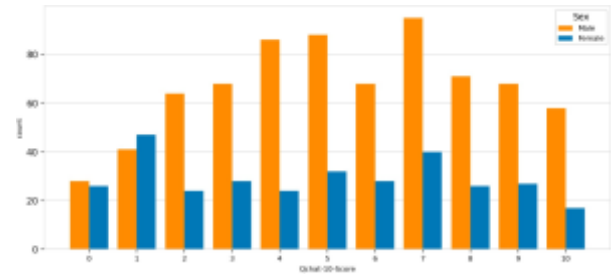


Figure 6: Visualization of the Q-CHAT-10 score according to gender.

4.3 Threshold Behavior

A ROC-AUC of 1.0000 means that, on this test set, every autistic sample received a higher predicted probability than every non-autistic sample. At $\tau = 0.5$, however, four non-autistic samples were classified as autistic (Table 2). This indicates that a threshold slightly above 0.5 could separate the classes perfectly on this particular test set. We deliberately did not tune the threshold on the test data, since doing so would bias the estimate; if a different operating point is desired, it should be chosen on the validation set.

4.4 Comparison with Prior Work

Table 3 compares the proposed classifier with the TPOT-based AutoML pipeline and classical ML models reported by Ehsan *et al.* [2]. For the AutoML pipeline, the autistic-class precision, recall, and F1-score are taken from that study's per-class report; its abstract reports an overall accuracy of 78%, whereas its result tables list 0.83, so we report both values. The comparison is indicative only, because the baseline results were obtained on a different, locally collected dataset with a different partition.

Table 3: Comparison with results reported in [2] (autistic class)

Model	Acc.	Prec.	Rec.	F1
SVM	0.710	0.718	1.000	0.836
k-NN	0.740	0.785	0.888	0.833
Naive Bayes	0.800	0.847	0.887	0.866
Logistic Reg.	0.757	0.805	0.875	0.838
Decision Tree	0.700	0.804	0.769	0.786
Random Forest	0.780	0.821	0.888	0.854
AutoML	0.830	0.868	0.868	0.868
Proposed method	0.975	0.965	1.000	0.982

The proposed classifier improves accuracy by more than 14 percentage points over the best figure reported for the AutoML pipeline, raises recall from 0.90 to 1.00, and raises the autistic-class F1-score from 0.86

to 0.98. The SVM baseline also attains recall of 1.000, but at an accuracy of only 0.71, which indicates that it labels most children as autistic and would generate many unnecessary referrals. In contrast, the proposed model combines perfect recall with a specificity of 0.9184.

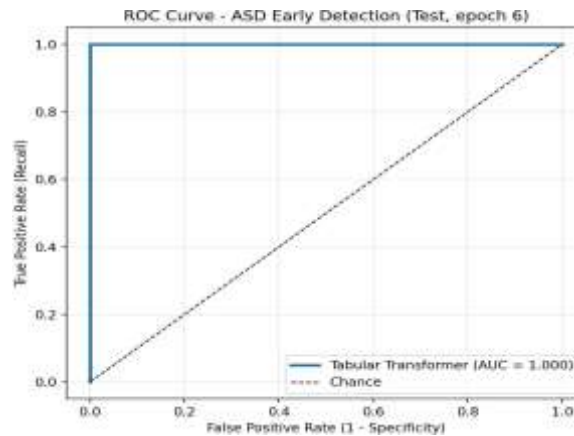


Figure 8: ROC curve of the best-validation model (epoch 6) on the held-out test set (AUC = 1.000).

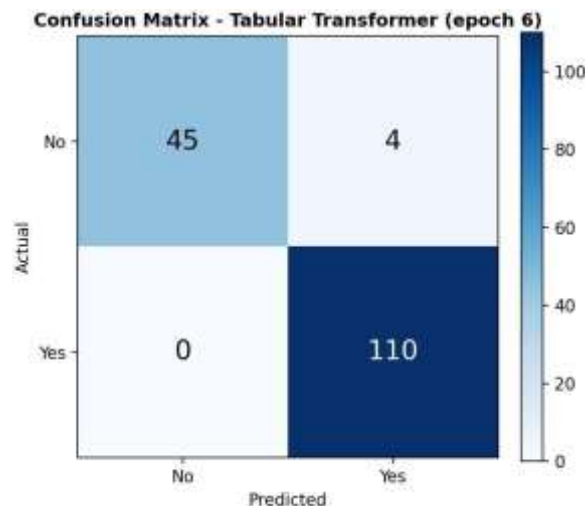


Figure 7: Confusion matrix of the best-validation model (epoch 6) on the held-out test set.

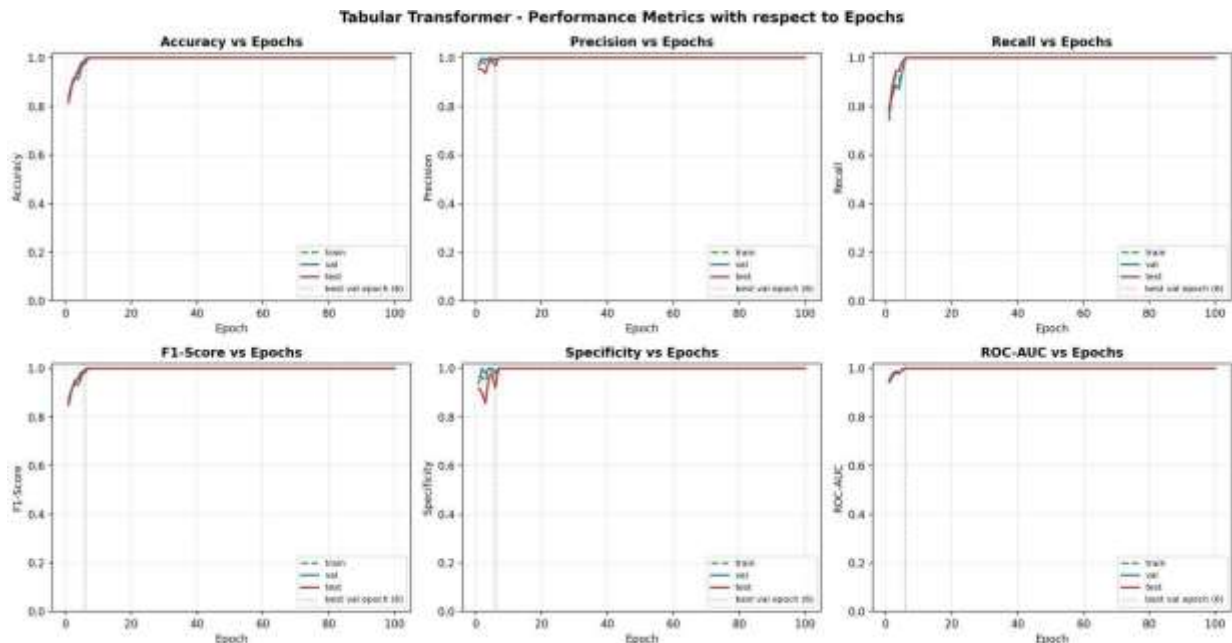


Figure 9: Performance metrics of the Tabular Transformer with respect to training epochs on the training, validation, and test sets.

5 Discussion

5.1 Clinical Interpretation

For a first-line screening tool, sensitivity is the priority, because false negatives postpone intervention during the period when it is most effective [3]. The proposed model missed no autistic child in the test set, while incurring four false alarms among 49 non-autistic children. Such false alarms lead to a referral for clinical assessment rather than a diagnosis, which is an acceptable cost for triage. The model is intended to supplement, not replace, assessment by clinical psychologists using instruments such as ADOS and ADI-R.

5.2 Threats to Validity

The very high scores should be interpreted with caution for the following reasons.

Label determined by the inputs. As shown in Section 3, in this dataset the class label is positive exactly when the sum of the ten binary items is at least 4. The label is therefore a deterministic function of A1–A10 (and of the score derived from them). A classifier that receives these items, or the score, can reach near-perfect performance by learning this threshold rule, which is consistent with the ROC-AUC of 1.0000. The results should thus be read as showing that the network can recover the Q-CHAT-10 scoring rule, not that it has learned to diagnose ASD or that it outperforms methods evaluated on clinically confirmed labels. Comparisons with studies whose labels come from independent clinical assessment are not like-for-like. A meaningful evaluation requires a target that is independent of the inputs, such as a clinical diagnosis, or an experiment in which the items that define the label are withheld.

Sample size and variance. With 159 test samples, of which 49 are negative, a difference of one or two samples changes specificity by 2–4 percentage points, and the confidence intervals are wide. Repeated stratified splits or cross-validation with confidence intervals should be reported. *Cohort composition.* The cohort is male-dominated (69.7%), 69.1% of records are positive, and 96.6% of questionnaires were

completed by a family member. Precision and accuracy depend on prevalence, so performance in community screening, where prevalence is far lower, would differ. Ehsan *et al.* [2] likewise note that demographic and cultural specificity may limit generalizability. *Respondent variability.* Questionnaires were completed by parents, caregivers, medical staff, or clinicians, which introduces response variability.

5.3 Comparison with AutoML

AutoML pipelines such as TPOT provide interpretable, reproducible scikit-learn pipelines and require little ML expertise [12,14]. Our results indicate that a validation-selected neural classifier can yield higher test scores on the same type of data, although at the price of manual design decisions and lower interpretability. A fair comparison would evaluate both approaches on identical splits with repeated runs.

6 Conclusion and Future Work

This paper studied early ASD detection from Q-CHAT-10 screening data with a neural classifier selected by validation performance. The epoch-6 checkpoint achieved an accuracy of 0.9748, precision of 0.9649, recall of 1.0000, F1-score of 0.9821, specificity of 0.9184, and ROC-AUC of 1.0000 on held-out data, which is higher than the figures reported for an AutoML baseline on a comparable dataset. Because the label in this dataset is fully determined by the questionnaire items, the small test set, and the composition of the cohort, these results demonstrate that the model recovers the screening rule rather than clinical diagnostic ability, and should be seen as preliminary. Future work includes (i) evaluation against independent clinical diagnoses and ablation of the items that define the label, (ii) repeated cross-validation with confidence intervals, (iii) evaluation on public datasets such as the UCI ASD screening data and on data from other regions, (iv) calibration and validation-based threshold selection, (v) model explainability, and (vi) comparison with other AutoML libraries.

7 Acknowledgements

The authors gratefully acknowledge the researchers who created and openly shared the Q-CHAT-10 toddler screening dataset used in this study [21], as well as the parents, caregivers, and clinicians whose questionnaire responses made such data available. The authors also thank Ehsan *et al.* [2], whose work on automated machine learning for early autism screening motivated this investigation and provided the baseline for comparison. The authors are grateful to their respective departments and institutions for providing the academic environment and computing resources used to carry out the experiments, and to the developers of the open-source scientific Python libraries on which the analysis relied. Finally, the authors thank the anonymous reviewers for their time and constructive feedback.

8 Data Availability

The dataset is available on: <https://www.kaggle.com/datasets/arnabdutta123/toddler-dataset>

9 Code Availability

The code is available on: [/kaggle/input/models/arnabdutta123/ft-transformer/transformers/default/1](https://kaggle.com/input/models/arnabdutta123/ft-transformer/transformers/default/1)

References

1. Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 18932–18943, 2021. Available:

- https://proceedings.neurips.cc/paper_files/paper/2021/file/9d86d83f925f2149e9edb0ac3b49229c-Paper.pdf
2. K. Ehsan, K. Sultan, A. Fatima, M. Sheraz, and T. C. Chuah, "Early detection of autism spectrum disorder through automated machine learning," *Diagnostics*, vol. 15, no. 15, Art. no. 1859, 2025. <https://doi.org/10.3390/diagnostics15151859>
 3. L. Zwaigenbaum, J. A. Brian, and A. Ip, "Early detection for autism spectrum disorder in young children," *Paediatrics & Child Health*, vol. 24, no. 7, pp. 424–443, 2019. <https://doi.org/10.1093/pch/pxz119>
 4. J. Zeidan, E. Fombonne, J. Scoriah, A. Ibrahim, M. S. Durkin, S. Saxena, A. Yusuf, A. Shih, and M. Elsabbagh, "Global prevalence of autism: A systematic review update," *Autism Research*, vol. 15, no. 5, pp. 778–790, 2022. <https://doi.org/10.1002/aur.2696>
 5. R. Romero-García, R. Martínez-Tomás, P. Pozo, F. de la Paz, and E. Sarriá, "Q-CHAT-NAO: A robotic approach to autism screening in toddlers," *Journal of Biomedical Informatics*, vol. 118, Art. no. 103797, 2021. <https://doi.org/10.1016/j.jbi.2021.103797>
 6. M. Marlow, C. Servili, and M. Tomlinson, "A review of screening tools for the identification of autism spectrum disorders and developmental delay in infants and young children: Recommendations for use in low- and middle-income countries," *Autism Research*, vol. 12, no. 2, pp. 176–199, 2019. <https://doi.org/10.1002/aur.2033>
 7. K. Vakadkar, D. Purkayastha, and D. Krishnan, "Detection of autism spectrum disorder in children using machine learning techniques," *SN Computer Science*, vol. 2, Art. no. 386, 2021. <https://doi.org/10.1007/s42979-021-00776-5>
 8. U. Erkan and D. N. H. Thanh, "Autism spectrum disorder detection with machine learning methods," *Current Psychiatry Research and Reviews*, vol. 15, no. 4, pp. 297–308, 2019. <https://doi.org/10.2174/2666082215666191111121115>
 9. G. Tartarisco, G. Cicceri, D. Di Pietro, E. Leonardi, S. Aiello, F. Marino, F. Chiarotti, A. Gagliano, G. M. Arduino, F. Apicella, F. Muratori, D. Bruneo, C. Allison, S. Baron-Cohen, D. Vagni, G. Pioggia, and L. Ruta, "Use of machine learning to investigate the quantitative checklist for autism in toddlers (Q-CHAT) towards early autism screening," *Diagnostics*, vol. 11, no. 3, Art. no. 574, 2021. <https://doi.org/10.3390/diagnostics11030574>
 10. D. Hossain, M. A. Kabir, A. Anwar, and Z. Islam, "Detecting autism spectrum disorder using machine learning techniques: An experimental analysis on toddler, child, adolescent and adult datasets," *Health Information Science and Systems*, vol. 9, no. 1, Art. no. 17, 2021. <https://doi.org/10.1007/s13755-021-00145-9>
 11. M. Bala, M. H. Ali, M. S. Satu, K. F. Hasan, and M. A. Moni, "Efficient machine learning models for early stage detection of autism spectrum disorder," *Algorithms*, vol. 15, no. 5, Art. no. 166, 2022. <https://doi.org/10.3390/a15050166>
 12. J. Waring, C. Lindvall, and R. Umeton, "Automated machine learning: Review of the state-of-the-art and opportunities for healthcare," *Artificial Intelligence in Medicine*, vol. 104, Art. no. 101822, 2020. <https://doi.org/10.1016/j.artmed.2020.101822>
 13. A. Truong, A. Walters, J. Goodsitt, K. Hines, C. B. Bruss, and R. Farivar, "Towards automated machine learning: Evaluation and comparison of AutoML approaches and tools," in *Proc. IEEE 31st Int. Conf. Tools with Artificial Intelligence (ICTAI)*, Portland, OR, USA, pp. 1471–1479, 2019. <https://doi.org/10.1109/ICTAI.2019.00209>

14. R. S. Olson and J. H. Moore, "TPOT: A tree-based pipeline optimization tool for automating machine learning," in *Proc. Workshop on Automatic Machine Learning*, PMLR, pp. 66–74, 2016. Available: https://proceedings.mlr.press/v64/olson_tpot_2016.html
15. S. G. Jacob, M. M. B. A. Sulaiman, B. Bennet, and P. Gastaldo, "Feature signature discovery for autism detection: An automated machine learning based feature ranking framework," *Computational Intelligence and Neuroscience*, vol. 2023, Art. no. 6330002, 2023. <https://doi.org/10.1155/2023/6330002>
16. R. T. Abbas, K. Sultan, M. Sheraz, and T. C. Chuah, "A comparative analysis of automated machine learning tools: A use case for autism spectrum disorder detection," *Information*, vol. 15, no. 10, Art. no. 625, 2024. <https://doi.org/10.3390/info15100625>
17. C. Allison, S. Baron-Cohen, S. Wheelwright, T. Charman, J. Richler, G. Pasco, and C. Brayne, "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*, vol. 38, no. 8, pp. 1414–1425, 2008. <https://doi.org/10.1007/s10803-007-0509-7>
18. C. Allison, B. Auyeung, and S. Baron-Cohen, "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," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 51, no. 2, pp. 202–212.e7, 2012. <https://doi.org/10.1016/j.jaac.2011.11.003>
19. L. Ruta, F. Chiarotti, G. M. Arduino, F. Apicella, E. Leonardi, R. Maggio, C. Carrozza, N. Chericoni, V. Costanzo, N. Turco, G. Tartarisco, A. Gagliano, C. Allison, S. Baron-Cohen, G. Pioggia, and F. Muratori, "Validation of the Quantitative CHecklist for Autism in Toddlers (Q-CHAT) in an Italian clinical sample of young children with autism and other developmental disorders," *Frontiers in Psychiatry*, vol. 10, Art. no. 488, 2019. <https://doi.org/10.3389/fpsy.2019.00488>
20. A. Niedźwiecka and E. Pisula, "Symptoms of autism spectrum disorders measured by the Quantitative Checklist for Autism in Toddlers in a large sample of Polish toddlers," *International Journal of Environmental Research and Public Health*, vol. 19, no. 5, Art. no. 3072, 2022. <https://doi.org/10.3390/ijerph19053072>
21. F. Thabtah, "An accessible and efficient autism screening method for behavioural data and predictive analyses," *Health Informatics Journal*, vol. 25, no. 4, pp. 1739–1755, 2019. <https://doi.org/10.1177/1460458218796636>
22. N. Farooqi, F. Bukhari, and W. Iqbal, "Predictive analysis of autism spectrum disorder (ASD) using machine learning," in *Proc. Int. Conf. Frontiers of Information Technology (FIT)*, Islamabad, Pakistan, pp. 305–310, 2021. <https://doi.org/10.1109/FIT53504.2021.00063>
23. Z. Sherkatghanad, M. Akhondzadeh, S. Salari, M. Zomorodi-Moghadam, M. Abdar, U. R. Acharya, R. Khosrowabadi, and V. Salari, "Automated detection of autism spectrum disorder using a convolutional neural network," *Frontiers in Neuroscience*, vol. 13, Art. no. 1325, 2020. <https://doi.org/10.3389/fnins.2019.01325>
24. A. Das, P. K. Pattnaik, A. Bandyopadhyay, N. Sharma, A. Singh, and S. Potharaju, "Privacy preserving federated framework for autism spectrum disorder screening with ensemble stacking and deep learning approaches," *Discover Artificial Intelligence*, vol. 6, no. 1, Art. no. 426, 2026. <https://doi.org/10.1007/s44163-026-01103-8>

25. A. Das, P. K. Pattnaik, and K. K. Jha, “Deep multimodal edge-deployable neural architecture for IoT-based autism detection using EEG and facial biomarkers,” *Recent Advances in Computer Science and Communications*, 2026. <https://doi.org/10.2174/0126662558437934260226205950>
26. A. Das, P. K. Pattnaik, and A. Bandyopadhyay, “Early intervention in autism spectrum disorder: Hybridizing convolutional neural networks with ensemble methods for enhanced detection using facial imagery,” in *Proc. 3rd Int. Conf. Intell. Syst., Adv. Comput. Commun. (ISACC)*, pp. 781–786, 2025. <https://doi.org/10.1109/ISACC65211.2025.10969308>
27. A. Das and P. K. Pattnaik, “Study of machine learning-based autism spectrum disorder detection,” in A. Bhattacharya, S. Dutta, M. A. Razzak, and D. Samanta, Eds., *Data Mining and Information Security*, pp. 405–419. Springer Nature, Singapore, 2026. https://doi.org/10.1007/978-981-96-6053-7_27