

Clinical Decision Support System for ASD Screening Using Ensemble Machine Learning

Nur Aela Dewi¹, Aftina Zahrayani²

¹Graduate School of Informatics, Department of Information System, Gazi University, Turkey

²Department of Information System, Faculty of Computer Science, Universitas Amikom Purwokerto, Indonesia
e-mail: 24803002010@gazi.edu.tr¹, aftinazahrayani@gmail.com²

Article History	Abstract
Received : 06 January 2026 Revised : 22 February 2026 Accepted : 02 March 2026 Published : 28 March 2026 Keywords: Autism Spectrum Disorder, Early Detection, Machine Learning, Q-Chat-10, XGBoost.	Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder with a continuously rising global prevalence, where the WHO estimates that 1 in 100 children worldwide is affected by this condition. Early detection through screening instruments such as Q-CHAT-10 is crucial for appropriate intervention. This study aims to build and compare the performance of three machine learning algorithms Naive Bayes, XGBoost, and Random Forest in predicting ASD risk in toddlers using the Q-CHAT-10 dataset. Hyperparameter optimization was performed using RandomizedSearchCV, and model robustness was assessed through 10-fold stratified cross-validation. Model performance was evaluated using accuracy, precision, recall, and F1-score metrics derived from a confusion matrix. Experimental results show that the XGBoost algorithm achieved the best performance, with a test accuracy of 99.05% and a 10-fold cross-validation mean accuracy of 99.31% ($\pm 0.75\%$), demonstrating strong generalizability. Naive Bayes achieved a test accuracy of 91.00% (CV: 94.84% $\pm 2.69\%$), while Random Forest yielded a test accuracy of 95.26% (CV: 96.48% $\pm 1.30\%$). McNemar's test confirmed that the performance differences between XGBoost and each of the other models are statistically significant ($p < 0.05$). These findings indicate that XGBoost is the most suitable algorithm for clinical decision support in early ASD screening within the scope of this dataset.

*)Corresponding author : aftinazahrayani@gmail.com

1. INTRODUCTION

Autism Spectrum Disorder (ASD) is a heterogeneous group of neurodevelopmental conditions [1], characterized by persistent deficits in social communication and social interaction, along with restricted interests and repetitive behaviors [2]. The World Health Organization (WHO) estimates a global ASD prevalence of approximately 1 in 100 children [2], while recent data from the Centers for Disease Control and Prevention (CDC) reported a prevalence of 1 in 31 children (3.2%) aged eight years in the United States in 2022. This rate shows a significant increase from 1 in 36 in 2020 and 1 in 150 in 2000, reflecting improved detection and broader diagnostic criteria [3]. This rising prevalence highlights the urgency of early detection and timely intervention to improve developmental outcomes for children with ASD.

Studies indicate that early diagnosis and intervention for autism are more likely to produce substantial long-term improvements in symptoms and functional abilities [4]. Early intervention typically occurs at or before the preschool years, as early as ages 2 or 3, during which the child's brain is still developing and is therefore more capable of change than at older ages [4]. Early detection of ASD enables interventions to begin even before a formal diagnosis is established, during a critical period of neural development, which consequently leads to better outcomes and prognoses [5]. Screening instruments play an essential role in the early detection of ASD. The Quantitative Checklist for Autism in Toddlers (Q-CHAT) has been developed as a quantitative assessment tool that demonstrates strong psychometric properties across diverse settings and cultures [6]. The Q-CHAT-10, a shortened version consisting of the 10 most predictive items, was designed as a rapid marker suitable for time-limited pediatric assessments and helps reduce delays in referral [7].

In recent years, artificial intelligence and machine learning (ML) have provided new opportunities to enhance predictive accuracy and classification reliability in the screening, detection, and diagnosis of autism. A study in [8] that implemented a feature-based ML classification model for ASD identification demonstrated that computational approaches can significantly improve the efficiency of the diagnostic process. Another study reported in [9] showcased the capability of ML algorithms to predict ASD using a minimal set of medical and background features. The XGBoost model has shown robust performance with high accuracy in identifying autistic individuals[1]. Research employing various classification algorithms has reported that Random Forest achieved the highest predictive accuracy of 99.75% on an ASD diagnostic dataset [2]. A well-tuned Naive Bayes algorithm achieved a cross-validation accuracy of 92.22%, indicating strong resistance to overfitting in imbalanced datasets [10]. Comparative studies have also highlighted performance variations based on sensitivity, specificity, accuracy, and area under the curve [11].

Despite extensive studies, several research gaps remain. Inconsistencies persist in selecting the optimal model for ASD prediction, and many prior studies rely on non-uniform evaluation methods, resulting in models that fail to generalize effectively to new instances [7]. Additionally, most previous studies have not rigorously validated algorithm performance through cross-validation nor conducted statistical significance testing between models. Therefore, this study aims to develop and compare the performance of three machine learning algorithms (Naive Bayes, XGBoost, and Random Forest) in predicting ASD risk in toddlers, incorporating hyperparameter optimization, 10-fold cross-validation, and statistical significance testing to address these gaps. The main contributions of this study are as follows: (1) a systematic comparison of three ML algorithms on the Q-CHAT-10 dataset, incorporating hyperparameter optimization via RandomizedSearchCV; (2) rigorous validation using 10-fold stratified cross-validation and McNemar's statistical significance test; (3) identification of the most effective algorithm for early ASD detection considering accuracy, precision, recall, and F1-score trade-offs; and (4) provision of baseline performance metrics as a reference for future studies within the Indonesian population context.

2. METHOD

This study implements a machine learning approach for classifying autism in toddlers. The models used include the Naive Bayes, XGBoost, and Random Forest algorithms. The research workflow begins with data collection, followed by preprocessing, train-test splitting, SMOTE balancing, feature standardization, hyperparameter tuning, model training, evaluation, and cross-validation.

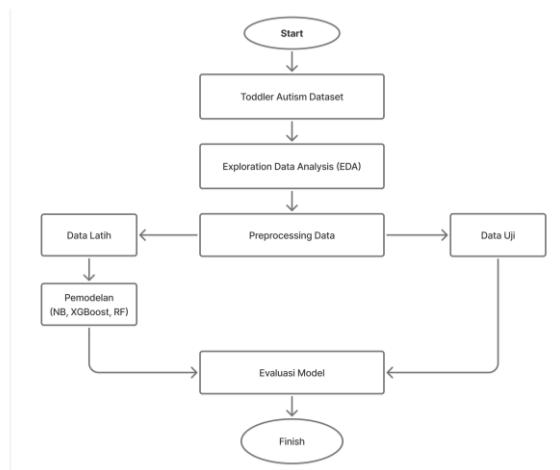


Figure 1 Research Methodology

2.1 Dataset Description

The dataset used in this study is a public dataset from Kaggle, developed by Dr. Fadi Fayez, titled “Autism Screening Data for Toddlers.” The dataset can be accessed through source [12]. This dataset contains various information related to the autism screening process in toddlers, including the 10 questions from the Q CHAT-10 score, age in months, gender, ethnicity, family history of

ASD, the respondent completing the test, and the ASD class label. The ten Q-CHAT-10 questions are presented in Table 1[13].

Table 1. Qchat-10-Score Questionnaire

Questions	
A1	Does your child turn or look at you when you call his or her name?
A2	How easy is it for you to obtain eye contact with your child?
A3	Does your child point with a finger to indicate that he or she wants something (e.g., pointing at a toy that is out of reach)?
A4	Does your child point to show you something interesting (e.g., pointing at an airplane in the sky)?
A5	Does your child enjoy pretend play (e.g., pretending to feed a doll or talking on a toy phone)?
A6	If you look at an object, does your child follow your gaze to look at the same object?
A7	If you or someone nearby appears sad or hurt, does your child show signs of wanting to comfort you?
A8	How would you describe your child's first words? (This question assesses whether the first words were social or more object-labeling.)
A9	Does your child use simple gestures to communicate (e.g., waving "bye-bye" or nodding/shaking the head)?
A10	Does your child often stare blankly at one point without a clear purpose?

A total of 1,054 participant records are included, consisting of approximately 70% males and 30% females. The dataset contains 19 columns comprising 13 numerical and 6 categorical attributes. Based on initial inspection, the data quality is considered very good, with no missing values or duplicate entries identified [12]. The class distribution shows that 69.1% of samples are labeled as No (not at risk), while 30.9% are labeled as Yes (at risk of ASD), indicating a class imbalance that requires treatment.

Table 2. Column and Sata Type Explanation

Column Name	Type Data	Description
Case No	Int64 (integers)	Case number for each row/entry
A1 - A10	Int64 (integers)	Responses to the 10 Q-CHAT-10 score questions. A value of 0 indicates "No," and a value of 1 indicates "Yes."
Age_Mons	Int64 (integers)	The child's age at the time of testing, measured in months.
Qchat-10 Score	Int64 (integers)	Total score obtained from questions A1 to A10. A score > 3 indicates a potential risk of autism.
Sex	Object (Text/categorical)	Gender (Male or Female).
Ethnicity	Object (Text/categorical)	Subject's ethnicity
Jaundice	Object (Text/categorical)	History of neonatal jaundice.
Family_mem_with_ASD	Object (Text/categorical)	Family history of ASD.

Who completed the test	Object (Text/categorical)	Relationship of the respondent completing the test to the child (Father, Mother, or Guardian).
Class/ASD Traits	Object (Text/categorical)	Target variable or classification label indicating whether the subject shows ASD related characteristics.

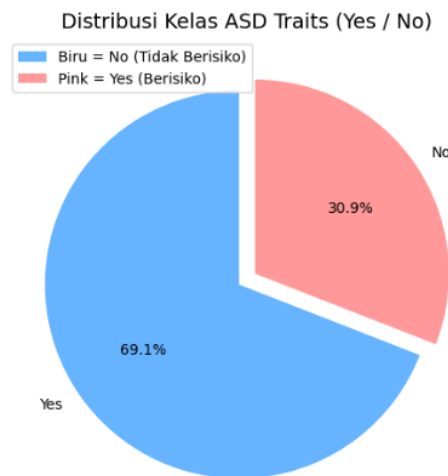


Figure 2. ASD Traits Class Distribution

Figure 2 presents the class distribution of the Class/ASD Traits variable, which consists of two categories: Yes (at risk of ASD) and No (not at risk of ASD). The pie chart visualization shows that the No category dominates the dataset with a proportion of 69.1%, while the Yes category accounts for only 30.9% of the total data. This distribution indicates that the dataset exhibits class imbalance, where the number of non-ASD-risk samples is significantly higher than ASD-risk samples. This condition is important to consider during the modeling process, as classification algorithms may become biased toward the majority class. Therefore, a balancing technique such as SMOTE (Synthetic Minority Over sampling Technique) is required to ensure that the model does not learn skewed patterns and is able to recognize both classes more accurately. This dataset serves as the foundation for training machine learning models to predict autism spectrum indications in toddlers based on Q-CHAT-10 scores and other supporting features.

2.2 Exploratory Data Analysis (EDA)

The EDA phase was conducted to understand the dataset's characteristics before modeling. Descriptive statistical calculations were performed on numerical variables, covering minimum and maximum values, mean, median, and standard deviation. Frequency distribution analysis was carried out for categorical variables, with particular focus on the target variable Class/ASD Traits to assess class imbalance. Correlation analysis was also performed to examine relationships between variables, particularly between individual Q-CHAT-10 behavioral indicators and the target. The heatmap revealed strong correlations between behavioral indicators A3, A5, and A9 and the target label. The Qchat-10-Score variable showed a very high correlation with the target, as it is a direct summation of the behavioral indicators. To prevent data leakage, the total score variable was removed from the training features.

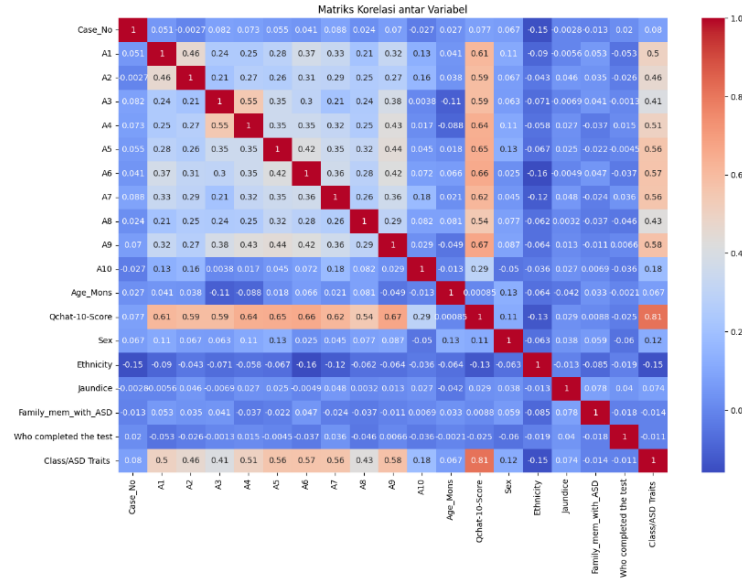


Figure 3. Correlation between variables

2.3 Preprocessing

The initial preprocessing stage was carried out to prepare the data before splitting it into the training and testing sets. The first step involved transforming categorical variables, namely Sex, Ethnicity, Jaundice, Family_mem_with_ASD, and Who completed the test into numerical representations using the LabelEncoder from sklearn.preprocessing. This transformation was necessary to ensure that the dataset is compatible with machine learning algorithms, which require numerical input.

2.3.1 Categorical variable Transformation (Label Encoding)

Categorical variables (Sex, Ethnicity, Jaundice, Family_mem_with_ASD, and Who completed the test) were transformed into numerical representations using the LabelEncoder from sklearn.preprocessing [14], [15]. This transformation was necessary to ensure dataset compatibility with machine learning algorithms, which require numerical input [16].

2.3.2 Train-test-Split

After encoding, the dataset was separated into features (X) and the target variable (y). The Qchat-10-Score column was also excluded from X to prevent data leakage. The data was then split using an 80:20 ratio through the train_test_split function from sklearn.model_selection, resulting in 843 training samples and 211 testing samples. Stratified sampling (stratify=y) was applied to ensure consistent class distribution across both subsets.

2.3.3 Handling Imbalance with SMOTE

In the descriptive analysis stage, the variation in values across behavioral indicators A1 to A10 shows a dynamic distribution. The class distribution in the training set revealed an imbalance, with 582 samples in the majority class (Yes/ASD) and 261 samples in the minority class (No/not ASD). This imbalance may cause classifiers to become biased toward the majority class. To mitigate this, the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training data — not the test set — to prevent data leakage [17]. SMOTE generates synthetic minority-class samples through interpolation between neighboring samples using the k-nearest neighbors (KNN) algorithm, as formulated below:

$$X_{syn} = X_i + (X_{knn} - X_i) \times \delta \quad (1)$$

Information:

X_{syn} : Synthesis Data

X_i : Data to be replicated

X_{knn} : Data that has the closest distance to data
 δ : Random value between 0 and 1

The application of SMOTE successfully balanced the class distribution, resulting in 582 samples for each category. This adjustment ensures that both classes receive equal representation during training, providing the model with an equal opportunity to learn from minority-class instances and reducing the risk of bias toward the majority class.

2.3.4 Feature Standarization

StandardScaler was applied after SMOTE, fitting exclusively on the SMOTE-balanced training data and transforming both training and test sets using the learned parameters [18]. This step ensures that all numerical features are on a uniform scale (mean = 0, standard deviation = 1) and prevents data leakage from the test set influencing the scaling parameters. The standardization formula is:

$$\hat{X} = \frac{x-\mu}{\sigma} \quad (2)$$

Subsequently, the same standardization parameters are applied to the test data to prevent data leakage. This standardization step is essential for stabilizing probabilistic calculations in Naive Bayes, optimizing the learning process in XGBoost, and maintaining consistent feature distributions before being processed by Random Forest.

2.4 Hyperparameter Optimization

To ensure each model was evaluated at its best potential configuration, hyperparameter optimization was performed using RandomizedSearchCV with 5-fold cross-validation internal to the search. This approach was applied to XGBoost and Random Forest, as these models have meaningful hyperparameters that significantly affect performance. Naive Bayes (GaussianNB) does not have critical hyperparameters and was used with its default configuration.

For XGBoost, the search space included `n_estimators` (100–500), `max_depth` (3–9), `learning_rate` (0.01–0.3), `subsample` (0.6–1.0), and `colsample_bytree` (0.6–1.0), with 50 random combinations evaluated. The optimal configuration identified was: `n_estimators` = 500, `max_depth` = 5, `learning_rate` = 0.1, `subsample` = 0.6, `colsample_bytree` = 0.7 (CV accuracy: 98.37%).

For Random Forest, the search space included `n_estimators` (100–500), `max_depth` (None, 5, 10, 15, 20), `min_samples_split` (2, 5, 10), `min_samples_leaf` (1, 2, 4), and `max_features` ('sqrt', 'log2'), with 50 combinations evaluated. The optimal configuration was: `n_estimators` = 400, `max_depth` = 5, `min_samples_split` = 2, `min_samples_leaf` = 2, `max_features` = 'log2' (CV accuracy: 96.82%).

2.5 Machine Learning Modelling

In the modeling stage, three machine learning algorithms are implemented for the classification task, namely Naive Bayes, XGBoost, and Random Forest.

2.5.1 Naïve Bayes

Gaussian Naïve Bayes (GaussianNB) was employed as one of the classification methods. This algorithm is a supervised, probability-based learning model that utilizes Bayes' Theorem under the assumption that all features are mutually independent [19], [20]. The posterior probability for a given class is computed as:

$$P(C|X) = \frac{P(X|C)P(c)}{P(x)} \quad (3)$$

Information:

$P(C|X)$ (Posterior) : The probability to be calculated (class c based on evidence x)
 $P(c)$ (Prior Kelas) : How often does class c appear in general?
 $P(X|C)$ (Likelihood) : How often does evidence x appear in class c?
 $P(x)$ (Prior Atribut) : How often does evidence x occur in general?

This approach is commonly used for binary classification and can perform well on high-dimensional datasets with low computational cost, as demonstrated in study [21], which highlighted the effectiveness of Naïve Bayes for classification tasks. In this research, Gaussian Naïve Bayes is also employed to handle numerical data after the encoding and standardization processes.

2.5.2 XGBoost

XGBoost (eXtreme Gradient Boosting) is an enhanced form of gradient boosting based on decision trees, designed to be faster, more efficient, and capable of handling complex feature structures [22]. The algorithm constructs the model sequentially by minimizing the loss function using a gradient descent approach while incorporating regularization mechanisms (L1 and L2) to reduce the risk of overfitting. The mathematical representation of XGBoost can be expressed as follows:

$$\hat{y}_i^{(t)} = \sum_{k=1}^t f_k(x_i) \quad (4)$$

Information :

- $\hat{y}_i^{(t)}$: Prediction for the i-th sample after the t-th iteration
- t : Iteration index (number of trees built)
- f_k : The function of the k-th tree, namely the decision tree that predicts the residuals at the k-th iteration
- X_i : Features or inputs up to i

The implementation of XGBoost in this study follows the practices adopted in previous ASD prediction research, where tree-based boosting has proven effective in capturing nonlinear patterns and feature interactions. The model is trained using the standardized training data, while parameters such as the number of estimators, tree depth, and learning rate are configured based on commonly used baseline settings in similar studies. After the training process is completed, the model is applied to the testing data to generate predictions as part of the evaluation of its classification performance for ASD detection.

2.5.3 Random Forest

This method is an ensemble model based on bagging, which constructs multiple decision trees randomly and combines their predictions to produce more stable and accurate classification results [23]. Each tree is trained on different bootstrap samples, and feature selection at each split is performed randomly, reducing overfitting. The classification formula is:

$$\hat{y} = \text{mode}(\{h_1(x), h_2(x), \dots, h_T(x)\}) \quad (5)$$

Information :

- T : Number of trees in the Algorithm
- $h_T(x)$: Prediction of the t-th Tree for input x
- $\text{Mode}(\dots)$: The class most predicted by the tree

This approach is consistent with previous studies on ASD prediction, where Random Forest has demonstrated strong performance on both clinical and survey-based data due to its ability to capture nonlinear relationships among variables. In this study, the Random Forest model was trained using standardized training data and subsequently used to generate predictions on the test set as part of the model evaluation process.

2.6 Model Evaluation

Model evaluation was conducted using several key metrics: precision, recall, F1 score, and accuracy, which were calculated based on the values of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Additionally, a confusion matrix was employed to provide a detailed overview of the prediction error patterns. The selection of these metrics follows the evaluation approaches described in prior studies [24], [25].

2.6.1 Precision

According to prior studies, precision is used to measure the proportion of correctly predicted positive cases (TP) relative to all predicted positive cases (TP + FP) [25]. This metric evaluates the model's ability to avoid false positives, making it particularly important in scenarios where incorrect positive predictions carry significant consequences. The formula for precision is expressed as follows:

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

2.6.2 Recall

Similarly, recall measures the model's ability to identify all actual positive cases (TP) out of the total true positives (TP + FN) [25]. This metric is crucial in medical contexts or sensitive detection tasks, including autism spectrum screening, where missing a positive case may lead to serious implications. The recall metric is formulated as follows:

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

2.6.3 F1-Score

The F1-score, as defined in prior work, represents the harmonic mean between precision and recall [25]. This metric provides a balanced assessment when both false positives and false negatives are equally important. The formula is defined as follows:

$$F1_{Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

2.6.4 Accuracy

Accuracy measures the proportion of all correct predictions (TP + TN) relative to the total number of samples. Although widely used, previous studies highlight that accuracy may be misleading when dealing with imbalanced datasets, as a model may appear "accurate" simply by predicting the majority class [24]. The accuracy metric is calculated using the following formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

2.6.5 Confusion Matrix

The confusion matrix presents the distribution of model predictions in a tabular format containing TP, FP, FN, and TN. As noted in recent research[26], the confusion matrix provides detailed insights into error patterns and is an essential tool for evaluating classification performance, particularly in medical or clinical risk prediction tasks.

Through the combined use of these metrics, the effectiveness of each model can be evaluated more comprehensively, enabling the identification of the most suitable model for early ASD risk screening.

2.7 Cross-Validation

To assess model robustness and generalizability beyond a single train-test split, 10-fold stratified cross-validation was applied to all three models using the SMOTE-balanced, scaled training data. In each fold, the model was trained on 9 folds and validated on the remaining fold, repeating 10 times. The mean accuracy and standard deviation across folds are reported to provide a reliable estimate of each model's generalization performance and stability.

2.8 Statistical Significance Testing

McNemar's Test was applied to determine whether performance differences between model pairs are statistically significant. The test compares the disagreements between two classifiers on the same test set, constructing a 2×2 contingency table of correct/incorrect predictions. The null hypothesis (H₀) states that there is no significant difference between two models. A p-value < 0.05 is used as the threshold for rejecting H₀, indicating statistically significant differences in predictive performance.

3. RESULTS AND DISCUSSION

The final models used were: GaussianNB (default configuration) for Naive Bayes; XGBClassifier with `n_estimators = 500`, `learning_rate = 0.1`, `max_depth = 5`, `subsample = 0.6`, `colsample_bytree = 0.7`, and `random_state = 42` for XGBoost; and RandomForestClassifier with `n_estimators = 400`, `max_depth = 5`, `min_samples_split = 2`, `min_samples_leaf = 2`, `max_features =`

'log2', and random_state = 42 for Random Forest. All models were trained on the SMOTE-balanced, standardized training data and evaluated on the held-out test set.

Table 3. Comparison of Model Performance Metrics on test Set

	NB No ASD	Naïve Bayes	XGB No ASD	XGBoost	RF No ASD	Random Forest
Precision	1.00	0.88	0.97	1.00	0.92	0.97
Recall	0.71	1.00	1.00	0.99	0.92	0.97
F1-Score	0.83	0.94	0.98	0.99	0.92	0.97
Accuracy	91.00%		99.05%		95.26%	

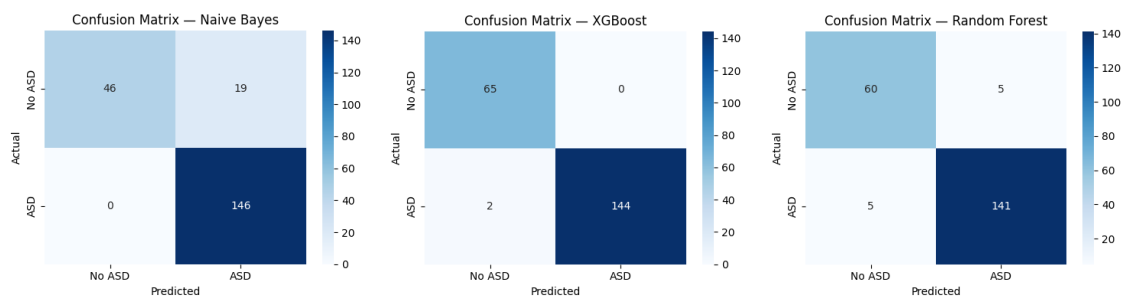


Figure 4 Confusion Matrix Naïve Bayes, XGBoost, Random Forest

Table 4. 10-Fold Cross-Validation Result

	Mean CV Accuracy	Std Deviation
Naïve Bayes	94.84%	±2.69%
XGBoost	99.31%	±0.75%
Random Forest	96.48%	±1.30%

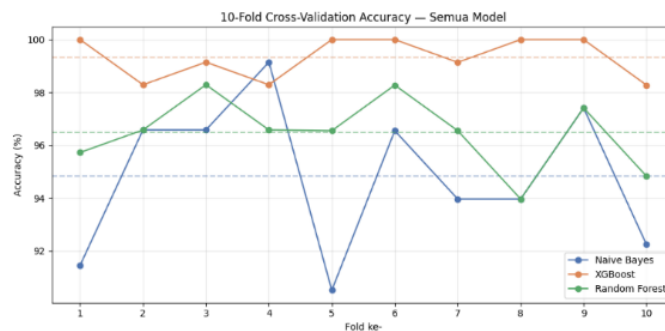


Figure 5 10-Fold Cross-Validation

Table 5 McNemar's Statical Significance Test Result

Model Comparison	P-Value	Conclusion
XGBoost vs Naive Bayes	0.0002	Significant ($p < 0.05$)
XGBoost vs Random Forest	0.0078	Significant ($p < 0.05$)
Naive Bayes vs Random Forest	0.0636	Not Significant ($p > 0.05$)

Referring to Table 3, the Naive Bayes model achieved a test accuracy of 91.00%. Its most notable characteristic is its perfect precision for the non-ASD class (1.00) and perfect recall for the ASD class (1.00), indicating that the model never misses an at-risk toddler. However, its recall for the non-ASD class is considerably lower (0.71), meaning it tends to over-predict ASD risk. The 10-fold cross-

validation mean accuracy of 94.84% ($\pm 2.69\%$) indicates moderate generalizability with relatively high variance across folds, suggesting some instability.

The XGBoost model achieved the highest test accuracy at 99.05%, with near-perfect scores across all metrics (Precision ASD: 1.00, Recall No ASD: 1.00, F1-Score: 0.98–0.99). The 10-fold cross-validation confirmed this performance with a mean accuracy of 99.31% ($\pm 0.75\%$), demonstrating both strong performance and high stability. The low standard deviation indicates that XGBoost generalizes consistently across different data subsets, making it the most reliable model in this study. The Random Forest model achieved a test accuracy of 95.26%, with balanced and consistent precision and recall values across both classes (0.92–0.97). The 10-fold cross-validation mean accuracy of 96.48% ($\pm 1.30\%$) confirms its stability, though it falls below XGBoost in both performance and consistency.

The McNemar's Test results presented in Table 5 confirm that the performance advantage of XGBoost over both Naive Bayes ($p = 0.0002$) and Random Forest ($p = 0.0078$) is statistically significant at the 0.05 level. The difference between Naive Bayes and Random Forest, however, is not statistically significant ($p = 0.0636$), suggesting their performance difference may be attributed to chance within this dataset.

3.1 Discussion

This study aimed to develop and compare three machine learning algorithms (Naive Bayes, XGBoost, and Random Forest) for ASD risk prediction in toddlers using the Q-CHAT-10 screening instrument. The methodological framework incorporated rigorous validation strategies including SMOTE-based class balancing, hyperparameter optimization via RandomizedSearchCV, 10-fold stratified cross-validation, and McNemar's statistical significance testing, addressing key limitations identified in prior research [7]. The experimental results demonstrate that XGBoost is the superior model for this classification task, achieving both the highest test accuracy (99.05%) and the highest cross-validation mean accuracy (99.31% $\pm 0.75\%$). Its consistently low standard deviation across folds confirms robust generalizability. The perfect precision on the ASD class ensures that positive predictions are highly reliable, which is critical in clinical decision-support applications where false alarms may cause unnecessary anxiety for families. Statistical testing further confirmed that XGBoost's superiority over both other models is not a result of chance.

Naive Bayes, while achieving lower accuracy (91.00%), demonstrated perfect recall for the ASD class, meaning it never misses an at-risk toddler. This property makes it potentially valuable in large-scale population screening where the primary goal is to minimize missed cases, even at the cost of higher false positives. However, its high variance in cross-validation ($\pm 2.69\%$) suggests it may be less stable across different data distributions. Random Forest provided a balanced performance (accuracy 95.26%, CV 96.48% $\pm 1.30\%$), offering a reasonable compromise between precision and recall. Its lower variance compared to Naive Bayes indicates more consistent generalization, though it does not match XGBoost's level of performance or reliability.

The limitations of this study include the relatively small and single-source dataset (1,054 records from Kaggle), which limits generalizability to broader clinical populations. The dataset also relies solely on Q-CHAT-10 behavioral responses without incorporating multimodal clinical features. Furthermore, the demographic composition of the dataset may not reflect the Indonesian toddler population specifically. Future research should consider external validation using independent datasets from diverse clinical settings, integration of multimodal data sources, and the development of explainable AI mechanisms to enhance clinical trust and interpretability.

CONCLUSION

This study successfully implemented and compared Naive Bayes, XGBoost, and Random Forest algorithms for early detection of ASD in toddlers using the Q-CHAT-10 screening instrument. Methodological rigor was ensured through SMOTE-based class balancing, hyperparameter optimization (RandomizedSearchCV), 10-fold stratified cross-validation, and McNemar's statistical significance testing. Within the scope of the Kaggle Q-CHAT-10 dataset used in this study, XGBoost demonstrated the best overall performance, achieving a test accuracy of 99.05% and a cross-validation mean accuracy of 99.31% ($\pm 0.75\%$), with statistically significant superiority over both other models ($p < 0.05$). These

results suggest that XGBoost is the most suitable algorithm for clinical decision-support systems prioritizing diagnostic accuracy and reliability in this context. Naive Bayes showed perfect recall for the at-risk class, making it potentially appropriate for large-scale screening scenarios where minimizing missed cases is the primary objective. Random Forest provided stable and balanced performance as a viable alternative. It is important to note that these findings are bounded by the characteristics of the dataset used, and broader generalization requires further validation on larger and more diverse clinical datasets. Future work should explore multimodal feature integration, external dataset validation, and the development of explainable AI approaches to support clinical adoption.

BIBLIOGRAPHY

- [1] A. Issac *et al.*, “The global prevalence of autism spectrum disorder in children: a systematic review and meta-analysis,” Feb. 01, 2025, *Korea Centers for Disease Control and Prevention*. doi: 10.24171/j.phrp.2024.0286.
- [2] World Health Organization (WHO), “Autism,” who.int. Accessed: Nov. 28, 2025. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders>
- [3] U.S. Centers For Disease Control and Prevention, “Data and Statistics on Autism Spectrum Disorder,” cdc.gov. Accessed: Dec. 01, 2025. [Online]. Available: <https://www.cdc.gov/autism/data-research/index.html>
- [4] S. Daniolou, N. Pandis, and H. Znoj, “The Efficacy of Early Interventions for Children with Autism Spectrum Disorders: A Systematic Review and Meta-Analysis,” *J. Clin. Med.*, vol. 11, no. 17, p. 5100, Aug. 2022, doi: 10.3390/jcm11175100.
- [5] Y. Alonso-Esteban and F. Alcantud-Marín, “Screening, Diagnosis and Early Intervention in Autism Spectrum Disorders,” *Children*, vol. 9, no. 2, p. 153, Jan. 2022, doi: 10.3390/children9020153.
- [6] G. Tartarisco *et al.*, “Use of machine learning to investigate the quantitative checklist for autism in toddlers (Q-CHAT) towards early autism screening,” *Diagnostics*, vol. 11, no. 3, Mar. 2021, doi: 10.3390/diagnostics11030574.
- [7] C. Allison *et al.*, “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 Paediatr. Open*, vol. 5, no. 1, p. e000700, May 2021, doi: 10.1136/bmjpo-2020-000700.
- [8] A. Novianto and M. D. Anasanti, “Autism Spectrum Disorder (ASD) Identification Using Feature-Based Machine Learning Classification Model,” *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 17, no. 3, p. 259, Jul. 2023, doi: 10.22146/ijccs.83585.
- [9] S. S. Rajagopalan, Y. Zhang, A. Yahia, and K. Tammimies, “Machine Learning Prediction of Autism Spectrum Disorder from a Minimal Set of Medical and Background Information,” *JAMA Netw. Open*, vol. 7, no. 8, Aug. 2024, doi: 10.1001/jamanetworkopen.2024.29229.
- [10] M. P. Pulungan, A. Purnomo, and A. Kurniasih, “Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Kepribadian MBTI Menggunakan Naive Bayes Classifier,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 5, pp. 1033–1042, Oct. 2024, doi: 10.25126/jtiik.2024117989.
- [11] 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,” *J. Biomed. Inform.*, vol. 118, Jun. 2021, doi: 10.1016/j.jbi.2021.103797.
- [12] Fadi, “Autism screening data for toddlers,” kaggle.com. Accessed: Dec. 01, 2025. [Online]. Available: <https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers/data>
- [13] “SCREENING--AUTISM--QCHAT-10 Question Autism Survey for Toddlers”.
- [14] S. O. Akinola and O. J. Oyabugbe, “Accuracies and Training Times of Data Mining Classification Algorithms: An Empirical Comparative Study,” *Journal of Software Engineering and Applications*, vol. 08, no. 09, pp. 470–477, 2015, doi: 10.4236/jsea.2015.89045.
- [15] M. Arif, maruf Setiawan, A. Dwi Hartono, M. Arif Ma, and ruf Setiawan, “Menggunakan Metode Machine Learning Untuk Memprediksi Nilai Mahasiswa Dengan Model Prediksi Multiclass,” *Jurnal Informatika: Jurnal pengembangan IT*, vol. 10, no. 1, p. 2025, doi: 10.30591/jpit.v9ix.xxx.

-
- [16] F. Pedregosa FABIANPEDREGOSA *et al.*, “Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot,” 2011. [Online]. Available: <http://scikit-learn.sourceforge.net>.
- [17] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” 2002.
- [18] scikit-learn, “StandarScaler,” <https://scikit-learn.org/>. Accessed: Dec. 01, 2025. [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>
- [19] scikit-learn, “1.9. Naive Bayes,” scikit-learn.org. Accessed: Dec. 01, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/naive_bayes.html
- [20] Eda Kavlakoglu, “What are Naïve Bayes classifiers?,” [ibm.com](https://www.ibm.com/think/topics/naive-bayes). Accessed: Dec. 01, 2025. [Online]. Available: <https://www.ibm.com/think/topics/naive-bayes>
- [21] H. Junianto, P. Arsi, B. A. Kusuma, D. Intan, and S. Saputra, “SINTECH Journal | 27 Evaluasi Aplikasi Raileo Melalui Analisis Sentimen Ulasan Playstore Dengan Metode Naive Bayes”, [Online]. Available: <https://doi.org/10.31598>
- [22] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
- [23] Eda Kavlakoglu, “What is random forest?,” [ibm.com](https://www.ibm.com/think/topics/random-forest). Accessed: Dec. 01, 2025. [Online]. Available: <https://www.ibm.com/think/topics/random-forest>
- [24] D. M. W. Powers and Ailab, “EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION.”
- [25] N. Haque, T. Islam, and M. Erfan, “An exploration of machine learning approaches for early Autism Spectrum Disorder detection,” Jun. 01, 2025, *Elsevier Inc.* doi: 10.1016/j.health.2024.100379.
- [26] Kyle Coddington, “Child Autism Statistics,” [prosperahealthcare.com](https://prosperahealthcare.com/autism-statistics/). Accessed: Nov. 28, 2025. [Online]. Available: <https://prosperahealthcare.com/autism-statistics/>