IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 6, December 2026, pp. 5548~5558
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i6.pp5548-5558

Improving early detection of autism through efficient machine learning models

Mou'ath A. Hourani[1], Mohammad O. Hiari[2], Abdelrahman H. Hussein[2], Omar Almomani[2], Qusai Y. Shambour[3]
[1]Department of Software Engineering, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, Jordan
[2]Department of Networks and Cybersecurity, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, Jordan
[3]Department of Software Engineering, Faculty of Information Technology, Middle East University, Amman, Jordan

Article Info

Article history:

Received Jun 13, 2025
Revised Aug 8, 2026
Accepted Sep 20, 2026

Keywords:

Autism spectrum disorder
Early screening
Feature selection
Hyperparameter optimization
Machine learning
SMOTE-ENN

ABSTRACT

Early diagnosis of autism spectrum disorder (ASD) is one of the clinical challenges that remains ever-present, and therefore requires the implementation of scalable, as well as reliable, screening approaches. This study presents a machine learning (ML) framework for early ASD detection by evaluating seven classifiers: categorical boosting (CB), decision tree (DT), k-nearest neighbors (KNN), logistic regression (LR), support vector machine (SVM), mixture discriminant analysis (MDA), and multi-layer perceptron (MLP), across two independent toddler screening datasets. To address class imbalance and feature noise, synthetic minority over-sampling technique-edited nearest neighbors (SMOTE-ENN) resampling, quantile transformer scaling, and Boruta feature selection were applied. Hyperparameter tuning was done using Bayesian optimization. The accuracy, precision, recall, F1-score, and Cohen's kappa were determined. The results demonstrate that CB, KNN, LR, and MLP have accuracy higher than 99% after feature selection. In particular, Boruta was able to improve the performance of models that are sensitive to noisy features, such as MDA and SVM, demonstrating the importance of proper feature selection for ASD screening tasks. The framework shows steady pattern in performance among datasets, indicating robust generalizability for data-driven ASD screening.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Qusai Y. Shambour
Department of Software Engineering, Faculty of Information Technology, Middle East University
Amman, Jordan
Email: q.shambour@meu.edu.jo

1. INTRODUCTION

Autism spectrum disorder (ASD) can be described as a complicated neurodevelopmental disorder that is typified by a range of behavioral impairments, social, and communicative impairments. The consequences of ASD are dramatic, as early childhood development is affected, and virtually any sphere of human life is impacted. According to a report issued by centers for disease control and prevention, about one out of thirty-six children in the United States would receive a divorce of ASD and this highlights the urgent need to have an effective diagnostic methodology [1]. The first thirty-six months of life are a critical time, after which early diagnosis allows early initiation of interventions that maximize developmental outcomes [2]. However, heterogeneity of ASD creates an extra burden on the conventional practices of diagnosis. The existing diagnostic practice is largely subjective, that is, it is based on clinical interviewing and standardized questionnaires. These methods are labor-intensive, expensive, and require specialized knowledge to read the subtle complexities of ASD, thus increasing the necessity of more objective, automated solutions [3]–[5].

The heterogeneity of ASD is the main factor that makes this disorder complex, as it implies a large number of phenotypic manifestations and is characterized by the high variability of socio-emotional manifestations. As an example, some of these children will have a striking deficiency in social communication and strong, monotonous repetitive behaviors related to certain interests, but they may not overtly have social impairments. This heterogeneity also makes the process of diagnosis more complicated and predetermines late detection and treatment. In addition, several of the currently used diagnostic plans are beyond the reach of ordinary people, especially in low-resource settings where medical know-how is limited. The need to have scalable, reliable diagnostic methods that would not lose efficacy in a wide range of populations is more pressing than ever before [3]–[5].

The recent developments in machine learning (ML) in diagnosing ASD are a promising step toward eliminating the traditional issues related to traditional diagnostic technologies. Large and complicated datasets can be used to take advantage of ML algorithms, which reveal non-intuitive tendencies related to ASD-related behaviors. In comparison to traditional approaches to diagnosis, ML is capable of incorporating a wide spectrum of data sources into creating valid forecasting models, such as behavioral indicators and demographic determinants [6]. These artificial systems assure more accurate diagnoses and less cost and time of assessment, thus allowing timely intervention [5]–[13].

Despite its promising potential, ML-based ASD diagnosis still faces challenges. One of these limitations relates to datasets that are specific to the field of ASD research. A limitation is caused by ASD-specific datasets, which tend to be biased because of class imbalances, resulting in models that do not generalize well after training. Improving the bias due to class imbalance is paramount for building fair and reliable ML models. Another obstacle is that it requires careful selection and normalization of features. Detecting salient features and making sure that they are scaled properly are essential to the quality of model performance and its ability to generalize to new, unfamiliar data [14]–[16].

To address these limitations, this study aims to present a robust methodology based on several components that enable the execution and endurance of the ML models used for the diagnosis of ASD in this work. The first approach is balancing the class distribution using the synthetic minority over-sampling technique (SMOTE)-edited nearest neighbors (ENN), which balances the performance of the model on the minority class samples and eliminates bias. This is particularly important when considering the issue of bias from the diagnosis of ASD when considering the number of cases (ASD vs non-ASD). Another contribution of importance is the introduction of a feature scaling and feature selection scheme. A quantile transformer will provide equal contribution from all features without giving the disproportionate effect of one feature over the others. Complementing this, the most predictive features from each set of data were selected using the Boruta algorithm, which increased the efficiency of the model and kept the clinical readability intact, and led to the successful incorporation into real-world clinical diagnostic processes.

Seven different ML models were considered for the evaluation: categorical boosting (CB), decision tree (DT), k-nearest neighbors (KNN), logistic regression (LR), support vector machine (SVM), mixture discriminant analysis (MDA), and multi-layer perceptron (MLP). For every model, all hyperparameters are exhaustively optimized using Bayesian optimization strategies that explore the hyperparameter space to maximize performance while minimizing overfitting. The resulting models are very accurate and robust and can provide credible predictions within the broad spectrum of ASD related datasets.

The rest of this paper follows the structure: in section 2, the literature available on ASD detection is reviewed, which serves as the background for our research. The methodological framework including data description, data pre-processing techniques, the implementation of ML models and hyperparameter optimization techniques is provided in section 3. The results of these experiments are analyzed in detail in section 4. Finally, section 5 concludes the study and proposes further research.

2. RELATED WORKS

Recent studies have demonstrated the effectiveness of using various ML models to improve the early detection and diagnosis of ASD. Vakadkar *et al.* [17] have made valuable contributions to this domain by conducting a dedicated study on children with the help of a variety of classifiers such as SVM, random forests, and LR. Their findings showed the best accuracy to be obtained with LR, while mentioning the limitations imposed by the restricted availability of large-scale open-source ASD datasets. Similarly, Alteneiji *et al.* [18] proposed the concept of early identification, especially during the first two years of life. Their neural network model proved superior to the more traditional methods in predictive quality, sensitivity, and specificity among all age categories, highlighting the necessity of age-specific optimization and more profound incorporation of ML technology into healthcare systems, so that they could be more accessible.

Hasan *et al.* [9] proposed a unified framework that integrated four feature-scaling methods, i.e., quantile transformer, power transformer, normalizer, and max abs scaler, along with eight classifiers, such as

adaptive boosting (AdaBoost) and LDA. Their study produced high classification accuracies across four age-oriented ASD datasets, with AdaBoost achieving an accuracy of 99.25% with toddlers and 97.95% with children. However, the authors also reported the need to increase the size of the dataset with a view to improving the generalizability of the model.

Bala *et al.* [11] also offered valuable insights into the most suitable feature subsets to be used in the detection of ASD, given that they deployed advanced feature-selection methods such as Boruta and recursive feature elimination. They observed that SVM would give maximum accuracy for all the age groups on the selected subsets of features. Moreover, Shapley additive explanations (SHAP) were also used, which presented a transparent model of how each feature chosen contributes. The application of the non-parametric statistical tests of the data means that their final results should be at least partially verified with various datasets.

Preprocessing of data and feature selection have become a key landmark in developing the best model performance. Uddin *et al.* [7] made a significant contribution to this field by introducing an integrated ML framework to recognize ASD in toddlers, where the central focus was the data transformation and feature selection. Some other methods were also applied in their methodology, including recursive feature elimination and SMOTE for optimizing feature subsets and reducing class imbalance. AdaBoost proved to be the most efficient classifier after being optimized for hyperparameters to boost the prediction accuracy. Building upon this, a comparative analysis of classifiers was conducted by Talukdar *et al.* [8] that showed that the best accuracy for classifiers would be obtained when the data from adolescents and toddlers were used, with 93.69% and 93.33% respectively. These results show that tree classifiers can be used in predicting ASD with accuracy, and this shows that ensemble learning strategies can be used to improve the predictive accuracy.

In recent years, more sophisticated ensemble learning techniques were suggested, such as those presented by Hajjej *et al.* [5]. In the first step, a random forest classifier was applied to assist in the realization of the proposed objective, while in the second step, an extreme gradient boosting (XGBoost) classifier was applied, which aids in the implementation of therapeutic interventions in the areas of physical, verbal, and behavioral. This two-stage approach can be defined as a great breakthrough in the study of ASD, which has profound implications for the procedure of diagnosis as well as clinical management.

Building on these earlier studies, the current research provides some unique insights through the innovative use of two toddler-specific datasets, including the established standard ASD screening data with 1,054 instances and the Saudi Arabian dataset with 506 instances, both of which are based on the widely used Q-CHAT-10 questionnaire. In order to deal with class imbalance, the paper applies the SMOTE-ENN resampling method, and the quality of the data is improved with the help of quantile-transformer scaling. The Boruta algorithm is used to discard uninformative predictors, which ensures that only highly informative features are injected into the modelling. Lastly, predictive accuracy was carefully optimized, and overfitting was fairly eliminated by meticulously tuning the performance of seven ML classifiers through Bayesian hyperparameter optimization. Assessment of the framework on the above datasets ensured strong generalizability and strong performance, thus providing a valuable resource-saving, scalable, and practical solution for screening ASD early, especially in contexts that have limited skill and expertise to use and specifically diagnose these behaviors.

3. PROPOSED METHOD

The proposed methodology employs a unified ML framework aimed at supporting the early and timely identification of ASD in toddlers. Its effectiveness was validated on two independent datasets, which demonstrated consistent and generalizable performance. These results suggest the framework offers a scalable and practical solution for early ASD detection in diverse settings.

3.1. Data preprocessing

At the preprocessing phase, extreme care was taken to narrow down and alter data to provide desirable input features in order to develop machine-learning models. Some preliminary data-cleaning was done, which entailed the methodical elimination of attributes that were considered potentially non-informative in either of the datasets. Specifically, the variables that were not included in the analysis because they have minimal effect on predictive objectives are: ‘Case no’ and ‘Who is completing the test’. A careful inspection showed the two datasets had full datasets with no missing values; in addition, extensive duplicate-checking revealed that there were no duplicate records, which eliminated the need for an imputation or duplicate-handling measure. Further data pre-processing steps such as data balancing, data transformations, and data feature selection are explored in the following sections. This detailed preprocessing framework forms a background infrastructure of cleaning and organizing datasets before building robust and generalizable ML models.

3.1.1. Data balancing

This study utilizes two distinct datasets to improve early ASD detection in toddlers. The first dataset, the ASD screening data for toddlers, is a set of 1,054 cases, 735 of which are classified as ASD, and 319 as non-ASD. The second dataset, the ASD screening data for toddlers in Saudi Arabia, consists of 506 cases, of which 341 were ASD cases and 165 non-ASD cases. In both datasets, the degree of class imbalance is pronounced and poses a threat of biasing ML models in favor of the majority class during training [19]. To mitigate this limitation, SMOTE-ENN resampling methodology [20], a multi-purpose strategy that homogenizes both SMOTE and ENN, was used.

Unlike conventional oversampling techniques, SMOTE-ENN first creates synthetic samples of the minority group to decrease the disparity between classes, and then removes the indistinct or noisy examples-usually belonging to the majority group-with the help of the ENN method. ENN works by looking at the local neighborhood of all samples and eliminating those samples that do not get the majority of their nearest neighbors supporting their labels, which eliminates noise. This two-step solution not only balances the dataset but also removes noise, which would have otherwise affected the capability of the classifier in distinguishing between discriminative patterns. As a result, SMOTE-ENN improves the result of model classification on skewed classification tasks by combining synthetic sample generation with neighborhood-based sample removal.

3.1.2. Feature transformation

Data preprocessing is largely dependent on feature transformation, which is used to alter the form or representation of data attributes in a dataset to improve the quality and relevance of the data and make the data more appropriate for modeling tasks. The general aim of the transformation of features is to create more informative, discriminative, or analyzable attributes without altering the structure of the dataset. Two feature-transformation methods were used consistently in both datasets in the current work, namely, categorical encoding and feature normalization.

Encoding is the process of converting categorical features to numerical features that are acceptable to ML models. The categorical features of the two datasets, namely sex, jaundice, ethnicity, region, Family_mem_with_ASD, and Class/ASD_Traits, were encoded using label encoding, whereby the categorical features are assigned integer values, but the same categories are given the same value. After the encoding process, these variables are encoded to be machine-readable, but they still have their inherent categorical meaning.

Normalization of features is important for ensuring that all variables are treated equally in the model training. The quantile transformer was utilized in this study, which is a non-linear transformation method that maps the distribution of all features to a specified target distribution (e.g., uniform, normal). This method leverages the quantile information to create new feature values, such that the set of new feature values follows a target-specified distribution and sharply mitigates the influence of outliers and skewed data. The re-distribution of features achieves the goal of normalizing the distribution of features, while still maintaining the rank order of the original set of values. The result is a set of comparable (same scale), normalized features that are more suitable for models that are sensitive to the input distributions.

3.1.3. Feature selection

Feature selection methods are utilized to address high dimensionality in datasets that are only supported by a limited number of relevant features, with the remainder having low contributions to the predictability of the model. The effect of independent features is negated by problems such as overfitting, noise, and high computation time. Within this study, the Boruta method was employed to select the relevant features that influence model performance and interpretability.

The Boruta method is a feature selection scheme for ML to be employed on low-dimensional datasets. The method compares and ranks the importance of features in a dataset to the importance of the same features in shuffled data. Anytime features show consistently higher importance in unshuffled data, the features are likely kept, while the feature scores below a certain threshold tend to be discarded. The Boruta method will be especially helpful if there are mixtures of relevant features and irrelevant features, as Boruta provides a systematic and data-driven approach to selecting features. Boruta improves model interpretability, reduces dimensionality, is resistant to overfitting, and is easily applicable to various data types [7].

3.2. Machine learning models

The proposed methodological framework systematically integrates and assesses seven ML algorithms through the full scikit-learn library in Python. The library includes a large number of implementations of ML, such as ensemble methods, probabilistic classifiers, and neural network

architectures. Each of these classifiers will be tested systematically to establish the most effective algorithmic solutions for classifying ASD with predictive accuracy and model generalizability. The chosen ML models are as follows:

3.2.1. Categorical boosting

CB is a high-performance gradient-boosting algorithm that builds an ensemble of DT for both classification and regression tasks. Critically, CB automatically supports categorical variables, including their coding in a way that retains their statistical data. The algorithm has features of ordered boosting and advanced regularization, which are expected to decrease overfitting and target leakage. CB is an efficient, scalable, highly predictive model that requires little parameter tuning, is interpretable, and can handle heterogeneous data.

3.2.2. Decision tree

DT is a non-parametric model, simple to use, and can be applied to classification and regression. A DT recursively divides the data space into homogeneous regions based on the feature and threshold value chosen to best separate the classes at each node. The DT structure is similar to that of a flow chart, with the internal nodes presenting a decision rule and the terminal nodes presenting the label of classes. It enables researchers to process massive amounts of data, especially when it comes to visualizing and understanding the data.

3.2.3. K-nearest neighbors

KNN uses an instance-based learning paradigm, in which classification choices are based on a majority labeling of the k nearest instances. This non-parametric algorithm is very sensitive to both the scaling of features and the manner in which data is distributed; it can, however, be seen to perform pretty well in cases where the data in question has a strong clustering structure. Thus, KNN continues to be a simple, useful tool in a variety of fields, renowned for its simplicity and interpretability.

3.2.4. Logistic regression

LR is a linear model that estimates the probability of a binary outcome based on a linear combination of input features. Through the use of the logistic (sigmoid) function, the LR model can map the predicted values to the range [0, 1], therefore allowing a probabilistic description of whether or not a class membership would occur. The resulting coefficients are used to estimate the effect of each predictor on the outcome, which can then be explained easily. The LR is computationally efficient and still best adapted when the relationship among covariates and the target is close to being linear.

3.2.5. Support vector machine

SVM performs complex classification tasks by identifying an optimal hyperplane that maximizes the margin between classes in a high-dimensional feature space. Through the use of kernel functions, it effectively captures non-linear decision boundaries without explicitly transforming the data. This makes SVM particularly well suited to high-dimensional data, and, when combined with class-weighted penalties, to handling class imbalance.

3.2.6. Mixture discriminant analysis

MDA can be considered an extension of classical linear discriminant analysis in that each class is modeled as a mixture over several multivariate Gaussian distributions instead of a single one. This provides MDA with the flexibility to represent more complex, multimodal, class-conditional densities, and thus provide a more accurate representation of underlying data structures, which might not be linearly separable. The expectation-maximization (EM) algorithm estimates the parameters of these component distributions, following which it becomes possible to construct estimates of the probability of class membership of novel observations. MDA is beneficial when the context is characterized by high heterogeneity or subpopulations within the class distributions because it can be used to draw finer decision boundaries.

3.2.7. Multi-layer perceptron

MLP is a feedforward neural network configuration with multiple layers of neurons. Learning the parameters is done with backpropagation over the MLP model, and MLPs have proven they can capture the complicated nonlinear relationships found in the data structure of interest very well. Their architectural flexibility and adaptive ability with respect to learning make MLPs particularly relevant in a supervised systems context to extract and process complex images or patterns.

3.3. Hyperparameter optimization

Systematic hyperparameter optimization is key to the robustness and generalizability of ML models. This process requires tuning of a few conceptual configuration parameters that control the learning and predictive behavior of a model in a selective manner. The aim is to achieve the best compromise between bias and variance, and to prevent overfitting and promote the generalizability of the model to entirely new data. This careful calibration has shown superiority in increasing the classification accuracy in diagnosing ASD.

This study employs the Bayesian optimization technique [21] to address hyperparameter tuning, which is considered a novel and advanced method for traversing intricate hyperparameter spaces. This approach builds a surrogate model based on Gaussian processes that makes a probabilistic connection between hyperparameters and the metric to be optimized. The efficiency of this method lies in the fact that it includes an iterative cycle, which balances the exploration of new areas of parameters and exploitation of good areas with additional knowledge of promising areas with the help of surrogate prediction. This framework has found application empirically, with remarkable performance in terms of optimization speed and good reliability in finding effective hyperparameter configurations within a small number of epochs.

For every ML classifier in our study, Bayesian optimization was used to determine the optimum values of critical parameters calling for learning rate dynamics, constraints on the depth of trees, regularization coefficients, and kernel configurations. Table 1 presents a detailed breakdown of these optimized hyperparameter values. The tuning of these hyperparameters helped optimize an ML model for toddler ASD detection.

Table 1. Hyperparameter configuration for ML models

ML model	Parameters
CB	learning_rate: 0.29629, depth: 4, iterations: 990, l2_leaf_reg: 9
DT	criterion: 'gini', max_depth: 10, min_samples_leaf: 1, min_samples_split: 2, splitter: 'best'
KNN	metric: 'manhattan', n_neighbors: 4, weights: 'distance'
LR	C: 498.93243, l1_ratio: 0.1, penalty: 'l2', solver: 'saga'
SVM	C: 1000.0, gamma: 'auto', kernel: 'rbf'
MDA	n_components: 3, max_iter: 200
MLP	activation: 'relu', alpha: 0.00194, learning_rate: 'adaptive', n_layers: 3, n_neurons: 150, solver: 'adam'

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1. Dataset selection

The investigation of the behavioral and demographic markers of early-stage ASD in toddlers involved two datasets, which provide different perspectives on early ASD manifestation. This dual dataset provides a stronger justification for developing better ML models in the early detection of ASD.

i) Dataset I: the first ASD toddler dataset is taken from Thabtah *et al.* [22]. This dataset contains 1,054 cases with 17 categorical variables. The 10 behavioral features mainly stem from the Q-CHAT-10 assessment tool, along with the relevant demographic features. There are 735 cases positive for ASD and 319 negatives for it in this dataset; this set of features has been constantly occupying itself in demonstrating ASD-specific behavioral patterns. These features have proved to be significant ones acting as surface discriminators for ASD against non-ASD cases, and hence have been used as instruments in evolving the classification algorithms.

ii) Dataset II: the second dataset is from the most recent work of Alkahtani *et al.* [23], who conducted ASD screenings among toddlers aged 12-36 months across Saudi Arabia. This dataset, gathered through online questionnaires, contained 506 cases with 17 attributes. Notable aspects include an Arabic adaptation of the Q-CHAT-10, alongside essential demographic variables such as age and gender, geographical distribution of the cases, and family history of autism. It contained 341 positives for autism and 165 negatives. This can shed some light on the manifestations of autism in diverse cultural backgrounds.

Together, these datasets enhance one another, providing insights into behavior and demographics drawn from various community backgrounds. This variety of data sources lays the groundwork for developing machine-learned models that are appropriate for a wide array of early-ASD manifestations across different demographic and cultural ecological settings. These sources are complementary, which is why our analysis for early detection is both robust and generalizable.

4.2. Performance evaluation metrics

A variety of complementary evaluation metrics were used to assess the performance of ML classifiers for screening toddlers with ASD. As these metrics are based upon a single evaluation framework, they provide a broad view of different behaviors of classifiers during the differentiation of ASD from

non-ASD in toddlers [24], [25]. The foundational assessment is accuracy, which is the sum of correct classifications in both positive and negative cases. However, although this metric provides an overall indication of the classifier's performance, its lack of detail regarding classifier behavior necessitates further assessments. Therefore, precision was included to evaluate the degree to which positive predictions for ASD were reliable; an important issue in clinical detection of ASD is avoiding false positives. Recalling the opposite end of the spectrum, recall (sensitivity) evaluates how well our models identified actual cases of ASD. This metric has proven invaluable in understanding the capability of our models to detect individuals that actually have ASD, while at the same time addressing the importance of minimizing false negatives; a key aspect of early intervention. The F1-score is a balanced evaluation metric between precision and recall as it calculates the geometric mean between the two metrics.

Cohen's kappa statistic provided additional insight into the degree of agreement between predicted classifications and actual classifications. While accuracy can only provide a measure of performance, Kappa incorporates agreement arising merely by chance when evaluating performance. Furthermore, Kappa scores range from -1 to 1, where 1 represents perfect agreement, 0 indicates random agreement (i.e., "coin flip "), and values less than 0 indicate systematic disagreement between predicted classifications and actual classifications. Incorporating Kappa provided us with a better comprehension of the performance of our classification algorithms; thus potentially eliminating any false assumptions associated with previous interpretations. By rigorously applying these metrics, a solid framework was established for assessing the performance of ML classifiers, which served as the basis for creating a dependable diagnostic tool aimed at early detection of ASD. This thorough level of evaluation ensures that our models not only meet the stringent standards necessary for clinical application but also offer transparency in how their performance is assessed.

4.3. Comparison of machine learning models

4.3.1. Performance evaluation using dataset I

The performance evaluation of the seven ML classifiers was assessed on dataset I, as shown in Table 2. Each algorithm has been tested in two different settings: using the complete feature set and using features that have been selected by the Boruta algorithm. Five evaluation metrics were used for a comprehensive assessment of both predictive power and class agreement.

Table 2. Comparative analysis of models' performance on dataset I

Evaluation metric	Features selection	CB	DT	KNN	LR	SVM	MDA	MLP
Accuracy	All features	0.9849	0.9509	0.9849	0.9925	0.9245	0.8642	0.9925
	Boruta	0.9925	0.9547	0.9887	0.9962	0.9321	0.9698	0.9962
Precision	All features	0.9853	0.9526	0.9853	0.9926	0.9358	0.8968	0.9926
	Boruta	0.9926	0.9569	0.9889	0.9963	0.9413	0.9710	0.9963
Recall	All features	0.9849	0.9509	0.9849	0.9925	0.9245	0.8642	0.9925
	Boruta	0.9925	0.9547	0.9887	0.9962	0.9321	0.9698	0.9962
F1-score	All features	0.9849	0.9506	0.9849	0.9924	0.9249	0.8642	0.9924
	Boruta	0.9924	0.9544	0.9887	0.9962	0.9324	0.9699	0.9962
Kappa	All features	0.9691	0.8990	0.9691	0.9846	0.8493	0.7332	0.9846
	Boruta	0.9846	0.9066	0.9768	0.9923	0.8641	0.9388	0.9923

Overall, the use of the Boruta feature selection was found to uniformly improve the performance of all the classifiers. LR and MLP stood out as the best performers, achieving the highest accuracy (99.62%) and kappa (99.23%) under Boruta, with precision, recall, and F1-scores all exceeding 99.6%. These findings are consistent with those of Vakadkar *et al.* [17], who reported strong performance of LR on the toddler ASD dataset following proper preprocessing. CB and KNN also benefited from feature selection, with CB's accuracy rising from 98.49% with all features to 99.25% with Boruta, and KNN improving from 98.49% to 98.87%. Both classifiers showed similar improvements in precision, recall, F1-score, and kappa, supporting the positive influence of Boruta on predictive performance.

The DT classifier showed a marginally higher accuracy increase (accuracy increasing from 95.09% to 95.47%), an increase that can be credited to the nature of the feature-selection procedures involved in the tree. In contrast, MDA witnessed a great improvement, with accuracy improving from 86.42% to 96.98% and kappa rising substantially to 93.88%. This significant increase indicates that Boruta can effectively isolate informative predictors, especially when the algorithm being used is susceptible to irrelevant/redundant data, which should also be noted in the case of Boruta-based selection on ASD datasets reported by Bala *et al.* [11]. SVM showed consistent but relatively small improvements in all measures, reflecting its resilience and adaptability, but also indicates possible weaknesses when dealing with more involved feature interactions, similar to prior ASD classification research with models based on SVMs [7].

In sum, these findings support the importance of robust feature selection, such as that performed by Boruta, in maximizing the performance of a classifier, particularly when it occurs on models that are highly sensitive to feature relevance. The reported improvements in accuracy, F1-score, and Kappa of most classifiers highlight the significance of feature selection as a key process towards the establishment of reliable early ASD classification systems (see Table 2 for detailed results).

4.3.2. Performance evaluation using dataset II

The performance evaluation of ML classifiers on dataset II is summarized in Table 3. Again, each model was assessed using the full feature set and with features selected by the Boruta algorithm. CB achieved perfect predictive performance, with accuracy, precision, recall, F1-score, and kappa all reaching 100% in both the all-features and Boruta-selected feature scenarios, which is consistent with other previous findings indicating that ensemble and gradient-boosting classifiers display strong generalization in ASD screening tasks [5]. KNN also showed the best results following the use of Boruta feature selection, achieving 100% across all evaluation metrics, a sign of highly performant instance-based learning in the screening of ASD data as reported in comparative studies on toddler datasets [9].

Table 3. Comparative analysis of models' performance on dataset II

Evaluation metric	Features selection	CB	DT	KNN	LR	SVM	MDA	MLP
Accuracy	All features	1.0000	0.9921	0.9921	0.9841	0.9444	0.9286	0.9921
	Boruta	1.0000	0.9921	1.0000	0.9921	0.9683	0.9444	0.9921
Precision	All features	1.0000	0.9922	0.9922	0.9846	0.9508	0.9388	0.9922
	Boruta	1.0000	0.9922	1.0000	0.9922	0.9704	0.9479	0.9922
Recall	All features	1.0000	0.9921	0.9921	0.9841	0.9444	0.9286	0.9921
	Boruta	1.0000	0.9921	1.0000	0.9921	0.9683	0.9444	0.9921
F1-score	All features	1.0000	0.9921	0.9921	0.9841	0.9447	0.9289	0.9921
	Boruta	1.0000	0.9921	1.0000	0.9921	0.9684	0.9447	0.9921
Kappa	All features	1.0000	0.9838	0.9838	0.9674	0.8884	0.8571	0.9838
	Boruta	1.0000	0.9838	1.0000	0.9838	0.9358	0.8879	0.9838

The DT classifier consistently showed high performance, with an accuracy of 99.21% and similar values for precision, recall, F1-score, and kappa under both configurations, which suggests that its internal structure is efficient in addressing redundant features. Both LR and MLP reported high and consistent results, with accuracy and kappa values of 99.21% and 98.38%, respectively, after Boruta was applied, which again demonstrates the strength and applicability of these models since their results were virtually identical irrespective of feature-selection procedures. Boruta was even more beneficial to SVM and MDA, as accuracy went up, respectively, from 94.44% to 96.83%, and its kappa coefficient rose from 88.84% to 93.58%. Likewise, MDA improved from 92.86% to 94.44% in accuracy, with a corresponding increase in kappa from 85.71% to 88.79%. These enhancements testify to the favorable impact of the chosen feature selection, especially on models that are sensitive to insignificant or redundant features, and similar findings are supported by previous research, which highlights the importance of feature-subset optimization and resampling in achieving improved ASD classification [7], [11].

To sum up, the Boruta feature selection method either maintained or even enhanced the performance of classifiers across all models. The most significant improvements were observed in KNN, SVM, and MDA, while CB, DT, LR, and MLP showed remarkable stability and dependability regardless of the feature selection strategy used. These results underscore the critical role of effective feature selection in optimizing the predictive accuracy and interpretability of machine-learned models for early ASD detection. Enhancements in accuracy, F1-score, and Kappa reinforce the necessity of integrating the latest advancements in feature-selection techniques for creating dependable classification systems, refer to Table 3 for detailed results.

5. CONCLUSION

This research introduces a systematic and robust ML framework aimed at the early detection of ASD in toddlers by directly tackling significant challenges such as class imbalance, feature relevance, and model optimization. By conducting a comparative analysis of seven ML classifiers across two separate ASD screening datasets, the findings reveal that selecting features is crucial for enhancing predictive performance and ensuring model reliability. Notably, the Boruta algorithm consistently improved or upheld classification accuracy, offering substantial benefits to models that are sensitive to noisy or irrelevant features, like MDA and SVM. Meanwhile, CB and KNN reached nearly perfect predictive accuracy following feature selection, while LR and MLP exhibited strong and stable performance. These consistent improvements across various

datasets and learning paradigms indicate that the proposed framework possesses excellent generalization and adaptability capabilities, making it suitable for real-world screening environments where resources and clinical expertise may be constrained. Overall, the findings highlight the need to combine effective feature selection with thorough model validation to create reliable, accurate, and scalable ASD screening instruments, thus aiding the screening and early intervention.

FUNDING INFORMATION

The authors state no funding is involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Mou'ath A. Hourani	✓			✓		✓				✓	✓	✓		
Mohammad O. Hiari	✓	✓			✓	✓		✓	✓					
Abdelrahman H. Hussein		✓	✓			✓	✓		✓		✓			
Omar Almomani	✓		✓	✓			✓	✓		✓				
Qusai Y. Shambour	✓	✓			✓		✓		✓	✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

DATA AVAILABILITY

The data supporting the findings of this study are publicly available on Kaggle at

– https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers, reference [22]
– https://www.kaggle.com/datasets/asdpredictioninsaudi/asd-screening-data-for-toddlers-in-saudi-arabia, reference [23].

REFERENCES

[1] M. J. Maenner *et al.*, "Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 Sites, United States, 2020," *MMWR. Surveillance Summaries*, vol. 72, no. 2, pp. 1–14, Mar. 2023, doi: 10.15585/mmwr.ss7202a1.
[2] C. P. Johnson and S. M. Myers, "Identification and evaluation of children with autism spectrum disorders," *Pediatrics*, vol. 120, no. 5, pp. 1183–1215, Nov. 2007, doi: 10.1542/peds.2007-2361.
[3] F. Thabtah and D. Peebles, "Early autism screening: a comprehensive review," *International Journal of Environmental Research and Public Health*, vol. 16, no. 18, Sep. 2019, doi: 10.3390/ijerph16183502.
[4] A. S. Mohanty, P. Parida, and K. C. Patra, "Usage of ML techniques for ASD detection," in *Machine Learning and Deep Learning in Medical Data Analytics and Healthcare Applications*, Boca Raton, Florida: CRC Press, 2022, pp. 91–112, doi: 10.1201/9781003226147-5.
[5] F. Hajjej, S. Ayouni, M. A. Alohali, and M. Maddeh, "Novel framework for autism spectrum disorder identification and tailored education with effective data mining and ensemble learning techniques," *IEEE Access*, vol. 12, pp. 35448–35461, 2024, doi: 10.1109/ACCESS.2024.3349988.
[6] F. Rangani *et al.*, "Revolutionizing heart disease diagnosis: a novel machine learning approach," in *Machine Intelligence for Research and Innovations (MAiTRI 2024)*, Singapore: Springer, 2026, pp. 467–475, doi: 10.1007/978-981-96-7614-9_35.
[7] M. J. Uddin *et al.*, "An integrated statistical and clinically applicable machine learning framework for the detection of autism spectrum disorder," *Computers*, vol. 12, no. 5, Apr. 2023, doi: 10.3390/computers12050092.
[8] J. Talukdar, D. K. Gogoi, and T. P. Singh, "A comparative assessment of most widely used machine learning classifiers for analysing and classifying autism spectrum disorder in toddlers and adolescents," *Healthcare Analytics*, vol. 3, Nov. 2023, doi: 10.1016/j.health.2023.100178.

[9] S. M. M. Hasan, M. P. Uddin, M. Al Mamun, M. I. Sharif, A. Ulhaq, and G. Krishnamoorthy, "A machine learning framework for early-stage detection of autism spectrum disorders," *IEEE Access*, vol. 11, pp. 15038–15057, 2023, doi: 10.1109/ACCESS.2022.3232490.
[10] R. Al-Dmour, H. Al-Dmour, E. B. Amin, and A. Al-Dmour, "Impact of AI and big data analytics on healthcare outcomes: An empirical study in Jordanian healthcare institutions," *Digital Health*, vol. 11, Jan. 2025, doi: 10.1177/20552076241311051.
[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, May 2022, doi: 10.3390/a15050166.
[12] A. Shalaldeh *et al.*, "Designing predictive models: a comparative evaluation of machine learning algorithms for predicting body carcass fat in ewes at weaning," *Agriculture*, vol. 16, no. 4, Feb. 2026, doi: 10.3390/agriculture16040488.
[13] A. A. A.-Shareha, M. M. Abualhaj, M. A. Alsharaiah, A. Al-Saaidah, and A. Achuthan, "Diabetes prediction through classification using pima dataset: survey and evaluation," *Journal of Soft Computing and Data Mining*, vol. 6, no. 1, Jun. 2025, doi: 10.30880/jscdm.2025.06.01.001.
[14] A. S. Albahri *et al.*, "Towards physician's experience: development of machine learning model for the diagnosis of autism spectrum disorders based on complex T-spherical fuzzy-weighted zero-inconsistency method," *Computational Intelligence*, vol. 39, no. 2, pp. 225–257, Apr. 2023, doi: 10.1111/coin.12562.
[15] T. Shrivastava, V. Singh, and A. Agrawal, "Autism spectrum disorder detection with kNN imputer and machine learning classifiers via questionnaire mode of screening," *Health Information Science and Systems*, vol. 12, no. 1, Mar. 2024, doi: 10.1007/s13755-024-00277-8.
[16] M. Bala and M. H. Ali, "Early stage prediction model of autism spectrum disorder traits of toddlers," in *International Conference on Computer, Communication, and Signal Processing*, 2023, pp. 3–17, doi: 10.1007/978-3-031-39811-7_1.
[17] K. Vakadkar, D. Purkayastha, and D. Krishnan, "Detection of autism spectrum disorder in children using machine learning techniques," *SN Computer Science*, vol. 2, no. 5, Sep. 2021, doi: 10.1007/s42979-021-00776-5.
[18] M. R. Alteneiji, L. Mohammed, and M. Usman, "Autism spectrum disorder diagnosis using optimal machine learning methods," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 9, Sep. 2020, doi: 10.14569/IJACSA.2020.0110929.
[19] M. Lham and A. Majda, "A comparative study of balancing genomic data approaches in autism classification," in *International Conference of Machine Intelligence and Computer Science Applications*, Cham, Switzerland: Springer, May 2026, pp. 26–36, doi: 10.1007/978-3-032-19963-8_3.
[20] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, Jun. 2004, doi: 10.1145/1007730.1007735.
[21] J. Wu, X. Y. Chen, H. Zhang, L. D. Xiong, H. Lei, and S. H. Deng, "Hyperparameter optimization for machine learning models based on Bayesian optimization," *Journal of Electronic Science and Technology*, vol. 17, no. 1, pp. 26–40, Mar. 2019, doi: 10.11989/JEST.1674-862X.80904120.
[22] F. Thabtah, F. Kamalov, and K. Rajab, "A new computational intelligence approach to detect autistic features for autism screening," *International Journal of Medical Informatics*, vol. 117, pp. 112–124, Sep. 2018, doi: 10.1016/j.ijmedinf.2018.06.009.
[23] H. Alkahtani, T. H. H. Aldhyani, M. Y. Alzahrani, and A. A. Alqarni, "Efficient deep learning and machine learning models for early-stage identification of autism spectrum disorder in toddlers: evidence from Saudi Arabia," *Journal of Disability Research*, vol. 2, no. 4, 2023, doi: 10.57197/JDR-2023-0048.
[24] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, Mar. 2024, doi: 10.1038/s41598-024-56706-x.
[25] S. A. Hicks *et al.*, "On evaluation metrics for medical applications of artificial intelligence," *Scientific Reports*, vol. 12, Apr. 2022, doi: 10.1038/s41598-022-09954-8.

BIOGRAPHIES OF AUTHORS



Mou'ath A. Hourani is a professor of Software Engineering in the Department of Software Engineering at Al-Ahliyya Amman University, Jordan. He received his B.Sc. degree in Computer Engineering from Al-Ahliyya Amman University, Jordan, in 1999, his M.Sc. degree in Software Engineering from the University of Western Sydney, Australia, in 2004, and his Ph.D. degree in Software Engineering from Newcastle University, Australia, in 2007. His research focuses on software engineering, recommender systems, and bioinformatics. He can be contacted at email: mhourani@ammanu.edu.jo.



Mohammad O. Hiari is a lecturer at Al-Ahliyya Amman University. He received his first degree in Software Engineering from Philadelphia University, Jordan, in August 2004, and his master's degree in Computer Science from Al Balqa Applied University, Jordan, in February 2016. His research area of interest includes VoIP, cybersecurity, machine learning, and optimization. He can be contacted at email: m.hyari@ammanu.edu.jo.



Abdelrahman H. Hussein is a professor at the Department of Networks and Cybersecurity at Al-Ahliyya Amman University, Jordan. He received his first degree in Computer Science from Jordan University of Science and Technology, Jordan, in July 2000, a master's degree in Computer Science from Jordan University, Jordan, in July 2003, and a Ph.D. from Anglia Ruskin University, United Kingdom, in 2010. His main research interests lie in the areas of VoIP, mobile ad-hoc networking, and artificial intelligence. He can be contacted at email: a.husein@ammanu.edu.jo.



Omar Almomani completed his bachelor's and master's degrees in Telecommunication Technology at the Institute of Information Technology at the University of Sindh in 2002 and 2003, respectively. He received his Ph.D. from the University Utara Malaysia in computer network and security in 2010. He is currently a professor in the Department of Networks and Cybersecurity, Faculty of Information Technology at Al-Ahliyya Amman University. His research interests involve network performance, network quality of service (QoS), IoT, network modeling and simulation, network and cyber security, and machine learning. He can be contacted at email: o.almomani@ammanu.edu.jo.



Qusai Y. Shambour earned his B.Sc. degree in Computer Science from Yarmouk University, Jordan in 2001, M.Sc. degree in Computer Networks from the University of Western Sydney, Australia in 2003, and Ph.D. degree in Software Engineering from the University of Technology Sydney, Australia in 2012. Currently, he is the Dean of the Faculty of Information Technology at Middle East University, Jordan. His research interests include AI applications, information filtering, recommender systems, machine learning, and data science. He can be contacted at email: q.shambour@meu.edu.jo.