



Early Detection of Autism Spectrum Disorder in Children Using Different Machine Learning Algorithms

Sabreen Waheed Kadhum^{1*}✉, Mohammed Ali Tawfeeq²id✉

¹ Computer Science Department, College of Science, University of Baghdad, Baghdad, Iraq

² Artificial Intelligence Engineering Department, College of Engineering, Al-Naji University, Baghdad, Iraq

Article Info	ABSTRACT
Received: 03 June 2026 Revised: 25 June 2026 Accepted: 26 June 2026 Available online: 30 June 2026 Keywords: <i>Autism Spectrum Disorder</i> <i>Diagnostic Engineering</i> <i>Machine learning,</i> <i>Sustainable Development Goals</i>	Autism spectrum disorder (ASD) is a neurological condition marked by impaired communication abilities, social detachment, and repetitive behaviors in individuals. Global health organizations are facing difficulties in establishing an effective ASD diagnostic system that facilitates precise analysis and early autism prediction. It is a scientific issue that necessitates resolution. This research presents an approach to early prediction of children with ASD using machine learning (ML) methods. Three stages comprise the suggested technique. First, a 1250-case ASD dataset was identified and preprocessed. Five extremely effective traits with a high Pearson correlation coefficient (PCC) are chosen from a 10-feature ASD dataset through its paces using five ML techniques: Naïve Bayes (NB), K-Nearest Neighbor (kNN), Decision Tree (DT), Support Vector Machine (SVM), and AdaBoostM1 (ABM1). The proposed framework is assessed in the third phase using five metrics: accuracy, precision, prediction time, recall, and F1-score. The findings revealed that NB and K-NN approaches exhibit superior accuracy rates of 99.2% and 97.2%, respectively, with minimal prediction times of approximately 0.3 and 0.45 seconds, respectively. Conversely, the DT and AdBM1 methods demonstrate a minor decline in accuracy, achieving 94.8% and 87.6%, respectively, along with increased prediction times. Nonetheless, the SVM approach performs the worst, achieving an accuracy of 80.4% with a prediction time of 0.84 seconds. This research aligns with several Sustainable Development Goals (SDGs), including SDG 3, which focuses on good health, and SDG 10, which addresses reducing inequalities, including those related to disability. Moreover, SCG9 aims to strengthen technological capabilities, particularly in scientific research.

1. Introduction

Approximately twenty million individuals globally are impacted by autism spectrum disorder (ASD), a neurodevelopmental condition. This condition is marked by significant challenges in communication and social interaction, as well as highly repetitive interests and behaviors [1]. Functional impairment can be significant in young individuals with ASD. Consequently, early and accurate diagnosis is essential for improving

everyday life, reducing symptom severity, and reducing maladaptive behaviors.

Notwithstanding extensive research, the brain mechanisms of ASD remain ambiguous. Consequently, ASD is diagnosed based on behavior rather than etiology or mechanism. Following the diagnosis of ASD, genetic testing is commonly employed to identify inherited variables; these tests just indicate risk. Nonetheless, research has progressed considerably. [2]

Autism is a complicated neurological condition marked by repetitive and inflexible behaviors, compromised communication skills, inadequate sensory processing, and cognitive inflexibility. It exerts a lifelong influence and usually presents itself before to the age of three[3]. According to the Centers for Disease Control and Prevention (CDC), one in every 54 children has been diagnosed with autism. Approximately 1% of the worldwide population is considered to have ASD. Accurate and precise prognostication is essential in the early stages of autism spectrum disorder. When children receive care from professionals, they might acquire essential communication and social skills. State or federal disability payments may also be accessible to assist with therapy expenses. [4]

Conventional behavioral examination between 1-3 years forecasts ASD. Postponing intervention during crucial social communication and learning phases in affected youngsters may diminish its effectiveness. Automated systems that capture a child's video and audio during public interactions and utilize ML to detect high-risk ASD children will facilitate early risk assessment and intervention. The newly preprocessed and balanced ASD dataset can facilitate autism research. Proper preprocessing produces superior outcomes compared to the original ASD dataset. [5]

Recent research has highlighted the repercussions of technology underutilization in early ASD detection. These studies underline the importance of interdisciplinary approaches that combine engineering innovation with public health tactics. For example, this provides crucial data for engineering models designed at developing scalable, automated ASD prediction systems. Here are some of most relevant studies to autism detection:

Alwidian et al. [6] Investigated the Association Classification (AC) for predicting the occurrence of autism in patients. Seven notable methods are selected to evaluate the effectiveness of the AC method in identifying correlations within the data. Imbalanced datasets with 21 different characteristics.

Vakadkar et al.[7] Developed prediction models based on the results by utilizing ML algorithms such as Naïve Bayes (NB), Random Forest Classifier (RFC) , Support Vector Machines (SVM), Logistic Regression (LR), and K Nearest Neighbors (KNN). The primary limitation of this research is the restricted availability of open-source and extensive ASD datasets. Constructing an

accurate model necessitates a substantial dataset. The dataset employed in this study was insufficiently populated with cases.

Siddiqui et al. [8] Introduced an on-body biological sensor and classification system for the detection of autistic gestures. The proposed system identifies and categorizes gestures. A server for data collection and classification acquires information from Bluetooth-enabled wearable biosensors. Ten autistic adolescents do each of the 24 gestures ten times. Temporal and spectral characteristics derived from sensor data are categorized utilizing K-NN, Decision

Trees, Neural Networks NN, and RF techniques. The primary limitation of this study is that children with ASD were in motion while standing, which is restrictive. This device has not been evaluated with biosensors from adolescents with ASD in various body positions. A solitary wrist bio-sensor gathered data, rendering the detection of intricate actions unfeasible.

Rajab [9] Constructed models utilizing pre-determined datasets through attribute selection techniques and three classification algorithms: AdaBoost, kNN, and ID3. Additionally, it is necessary to illustrate which attribute selection methodologies discern essential features influencing the first diagnostic procedure. The categorization of screening tools is profoundly influenced by these methodologies. The efficacy of features chosen from a real-world dataset of toddlers using five different techniques: Chi-square (CHI), Correlation Feature Set (CFS), Information Gain (IG), Gini index (GI), and Fast Correlated Based Filter (FCBF) were assessed by ML models.

Pan et al [10] utilized data from the Chinese version of the M-CHAT-R combined with sociodemographic and environmental factors to improve classification performance. By applying multiple feature selection methods and ML algorithms, the study identified a reduced subset of key items capable of effectively distinguishing ASD from typically developing children.the best-performing model achieved in a seven-item classifier with 92.5 % sensitivity, 90.1 % specificity, and 10.0% positive predictive value.

Bala et al. [11] Implemented multiple classification methods on datasets comprising toddlers, children, teens, and seniors. Additionally, employed several feature selection techniques to

enhance these baselines and optimize feature sets. This model examined significant features but was not trained on diverse multivariate datasets. Anita and

Shinde and Patil [12] Developed a multi-classification approach including various classifiers to enhance the precision of ASD predictions. The model's performance was assessed using several ML approaches applied to three imbalanced datasets comprising children, teens, and adults.

Khudhur and Khudhur [13] Utilized various predictive algorithms on four accessible non-clinical ASD screening datasets. These datasets encompass children, adolescents, toddlers, and adults. Models for ASD identification was assessed based on various performance metrics. Examples include DT, K-NN, DT, NBs, and RF.

Farooq et al. [14] employed Federated Learning (FL) to diagnose ASD in children and adults by training two ML classifiers, SVM and LR. The intricacy of the suggested architecture and the training of federated learning models across numerous devices with constrained processing and storage capabilities complicate the application of the proposed model to more sophisticated tasks necessitating deep neural networks or other advanced machine learning models. The variability of suggested data between devices and locations complicates the development of universally effective models.

Das et al. [15] utilized a diverse selection of foundational models, such as LR, K-NN, NNs, DT, and SVM, to optimize feature selection. The methodology's incorporation of intricate feature selection and model stacking techniques is a critical element. The primary objective is to identify the four primary characteristics that are most diagnostic of ASD in order to facilitate future training in the early detection of autism in preschoolers.

Researches on the diagnosis and prediction of ASD presents several limitations. Limited homogeneous sample numbers restrict the generalizability to diverse groups. Numerous research incorporates subjective assessments, such as clinician observations and parental reports, which may skew results. The limited application of sophisticated ML techniques presents additional difficulty, as certain models exhibit deficiencies in interpretability and robustness. The dependence on discrete attributes, as opposed to biological, external

factors and behavioral factors, has also constrained comprehension. These challenges must be addressed to enhance the identification of ASD and prediction. Accurate diagnosis and treatment of ASD require clinical judgment and standardized assessments. Therapies that influence early development necessitate the identification of ASD.

This study employs medical information from 1,250 cases encompassing ten distinct attributes, including "Gender, speech delay, learning disability, intellectual disability, Social Issues, Depression, anxiety disorder, jaundice, Genetic disorder and family history". The proposed algorithms categorize children's instances into two classifications: non_ASD and ASD. The prognosis of these cases relies on the chosen data utilizing five different ML algorithms: Naïve Bayes (NB), k-Nearest Neighbors (k-NN), Decision Trees (DT), Support Vector Machines (SVM), and AdaBoostM1 (ABM1). These predictive methods are formulated based on a set of ten features that signify the longitudinal reversal of ASD, enabling the model to learn from labeled instances and classify new, unobserved cases.

A variety of features can improve model accuracy. Features that reflect important data will help the model find patterns. Extraneous or noisy elements may reduce accuracy. Due to "overfitting," the model learns non-beneficial patterns from training data, affecting its performance on novel data. Thus, dimensionality reduction is essential for finding the best features and boosting model performance.

In the context of feature analysis for ML, there are several methods to ascertain the correlation between attributes in a dataset to understand their interdependencies and connections. Improve the performance of a model by grouping features based on how they are related to each other. The Pearson Correlation Coefficient (PCC) evaluates the linear association between two continuous variables and is efficient. Statisticians, academics, and data analysts employ the PCC to comprehend correlations among variables.

Facilitates the elimination of superfluous characteristics (highly correlated features) to mitigate multicollinearity and identify pertinent features with a robust correlation to the goal variable (dependent variable) while maintaining a diminished correlation among themselves.

The purpose of the present research is not only to facilitate the identification of personalized delays in brain function or structure that are linked to ASD, but it also demonstrates the efficacy of relevant data in elucidating the relationships among features regarding their impact on model performance for predictive models of early autism indicators in physiological data samples. This has been the focus of ASD prediction efforts through the utilization of the PCC method with ten diverse attributes to enhance precision measurements.

This work is organized into five sections: Section 2 provides a detailed explanation of the suggested technique and introduces the adopted prediction models that will be employed. Section 3 discusses the criteria for assessment. The findings, when compared with compression to state_of_the art investigations, are presented in Section 4. Final thoughts based on the work's primary findings and potential future directions are explained in Section 5.

2. Methodology

The primary aim of this research is to forecast the likelihood of ASD at an early stage and to better specify the characteristics that have a more substantial impact on the prediction of increased accuracy. This could reduce the costs of various

patient trials, as not all traits significantly influence the outcome prediction.

Figure 1 presents the proposed framework for the early prediction of ASD among clinical controls, which is based on the dataset and can be useful for predicting whether a child will develop ASD in the early stages. The initial source of data is the Kaggle site [17].

The PCC technique was employed to ascertain the relationship among the implemented ten features. After extracting five unique features with the highest positive correlation values, the data is subsequently allocated to five distinct ML algorithms (NB, k-NN, DT, SVM, and ABM1) for analysis.

2.1 Data Preprocessing

In the data partitioning process, the dataset including 1,250 instances is divided into two segments for subsequent analysis. 80% of the data is allocated for learning, whilst the remaining 20% is allocated for test. Samples from the dataset used with the ten key features are shown in Table 1. The first column in this table denotes gender (0 for females and 1 for males).

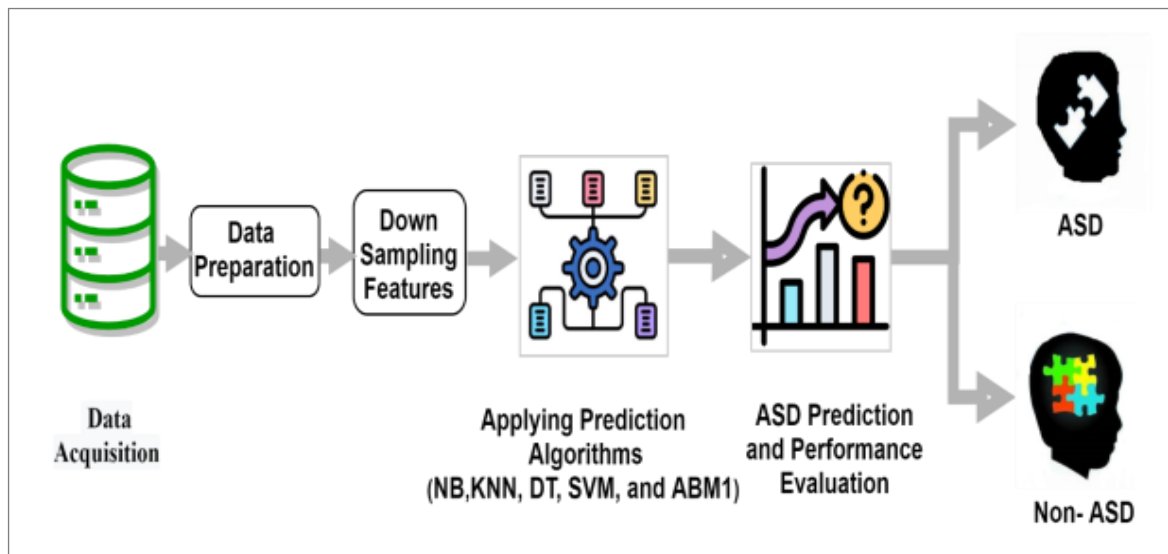


Figure 1. Formulation of a framework

Table 1: Examples of the executed dataset.

Case	Gender	Speech Delay	Learning disorder	Intellectual disability	Social Issues	Depression	Anxiety disorder	Jaundice	Genetic disorder	Family history	ASD_traits
1	0	1	1	1	1	1	1	1	1	0	1
2	1	1	0	1	0	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	0	1
4	1	1	1	1	1	1	1	0	0	0	1
5	0	1	1	1	1	1	1	0	0	0	0
6	1	1	1	1	1	1	1	0	0	0	1
7	1	1	1	1	1	1	1	1	1	0	1
8	1	1	1	1	1	1	1	1	1	0	1
9	1	1	1	1	1	1	1	0	0	0	0
10	1	1	1	1	1	1	1	0	0	0	1
11	1	1	1	1	1	1	1	1	1	1	1
12	1	1	1	1	1	1	1	1	1	0	1
13	0	1	1	1	1	1	1	1	1	0	0
14	0	1	1	1	1	1	1	1	1	0	1
15	1	1	1	1	1	1	1	1	0	0	0

2.2. Selection of Features

Feature reduction is vital to prediction system evolution, improving performance and efficiency. Removing unnecessary or duplicate features reduces dataset dimensionality, speeding up training and inference. Feature reduction also improves model interpretability, making the feature-predicted relationship easier to grasp.

Evaluation of attribute correlations while maintaining the dataset's integrity is often achieved using techniques such as PCC. Ultimately, ML algorithms focus on the most germane information by efficiently reducing features. The PCC approach is employed in this study to determine the likelihood of the ten features “gender, speech delay, learning disorder, intellectual disability, social issues, melancholy, anxiety disorder, jaundice, genetic disorder, and family history”.

A linear relationship between two continuous variables can be measured by the PCC, which is frequently represented as (r). In the interval [-1,+1], Here is the process[18 ,19]

$$r(X, Y) = \frac{\sum_{n=1}^N (X_n - \bar{X}) (Y_n - \bar{Y})}{\sqrt{\sum_{n=1}^N (X_n - \bar{X})^2 \sum_{n=1}^N (Y_n - \bar{Y})^2}} \quad (1)$$

Where:

X and Y are two features that are being compared, and N is the number of samples.

$\bar{X}$ and $\bar{Y}$ denotes the mean of particular attributes correspondingly.

Calculate the value of the Pearson correlation coefficient (r) for each feature pair in the dataset.

r = 1: Optimal positive correlation

r = -1: Optimal negative correlation

r = 0: No correlation

Create a correlation matrix by utilizing the correlation coefficients of feature pairs. This matrix facilitates the identification of robust relationships. Table 2 illustrates the relationships among the 10 features implemented.

Table 2. Ten-feature dataset's correlation matrix

	Gender	Speech Delay	Learning disorder	intellectual disability	Social Issues	Depression	Anxiety disorder	Jaundice	Genetic Disorder	family_history
Gender	1	0.22682	0.23091	0.22682	0.22862	0.22826	0.22764	-0.04632	0.2046	0.05934
Speech Delay	0.22682	1	0.960363	0.99394	1	1	0.99102	-0.04894	0.89153	0.236755
Learning disorder	0.23091	0.96036	1	0.95413	0.96028	0.963274	0.95174	-0.04597	0.85619	0.230762
intellectual disability	0.22682	0.99394	0.95414	1	0.99394	0.993923	0.9850	-0.04894	0.89153	0.241181
Social Issues	0.22861	1	0.96028	0.99393	1	1	0.9910	-0.04972	0.89129	0.23651
Depression	0.22826	1	0.963274	0.99392	1	1	0.9909	-0.04877	0.89114	0.239457
Anxiety disorder	0.2276	0.9910	0.9517	0.9850	0.991	0.9909	1	-0.0496	0.8828	0.2279
Jaundice	-0.0463	-0.0489	-0.0459	-0.0489	-0.0497	-0.0487	-0.049	1	0.01236	0.13342
Genetic Disorders	0.2046	0.89153	0.856197	0.89153	0.89129	0.891139	0.8828	0.01236	1	0.216754
family_history	0.05934	0.23675	0.23076	0.24118	0.23652	0.23945	0.2279	0.13342	0.21675	1

A significant positive correlation value exists between the two variables (speech delay, learning disorder, intellectual disability, social issues, depression, and anxiety disorder) around (0.981). In contrast, the negative correlation for “Jaundice” is the highest, with a value of -0.05, suggesting an inverse relationship with all nine attributes. It is essential to comprehend this connection to provide genetic counseling, risk assessment, and early intervention to families with a history of autism or related conditions.

To conclude, the correlation approach that is most suitable is determined by the distribution assumptions and the datasets that are employed. Each technique has particular benefits and is suitable for a variety of analyzes.

2.3 Prediction

This research employed five supervised ML techniques. The primary algorithms are predominantly trained with the labeled dataset. The qualifying model classifies an unlabeled testing dataset into suitable categories. The subsequent subsections present the implemented algorithms for autism prediction[13, 10]

2.3.1 Naïve Bayes (NB)

NB is a straightforward and efficient classification technique that is predicated on Bayes' theorem and feature independence. The most probable label is assigned by calculating class probabilities based on characteristics. It is particularly effective with small datasets and has rapid computing capabilities; however, its efficacy may be compromised when features are correlated.

NB continues to be favored due to its simplicity and efficiency.

- The Intermediate node: manages a variety of properties.
- Leaf nodes signify the outcome of each test.

2.3.2 Support Vector Machine (SVM)

A supervised ML method, SVM is used for prediction, and regression. is efficient in high dimensional spaces and resistant to ‘overfitting’, especially when features outnumber samples. SVMs are efficient since the decision function uses just part of the training data.

2.3.3 AdaboostM1 (ABM₁)

ABM₁ collaboratively learning supervised classifier. The adaptive improvement method integrates multiple weak classifiers into a strong classifier for enhanced classifying. The preliminary stage assigns equal weight to observations. Estimation error values determine classifier coefficients, while weak coefficients influence observation weights. Classifier coefficients represent errors. Consequently, the ABM1 algorithm can assign greater weight to misclassified data and lesser weight to well-established data. A linear combination integrates multiple weak classifiers to build a strong, precise classifier. The ASD dataset is subjected to training and evaluation through these methodologies.

3. Evaluation Criteria

Five predictive algorithms were utilized on the same balanced dataset including 1,250 cases to ascertain the most effective way by

evaluating statistical parameters. The employed algorithms were NB, KNN, DT, SVM, and ABM1. Table 3 illustrates the confusion matrix of the proposed research to evaluate the efficacy of each implemented approach.

Table 3. Confusion matrix of ASD prediction

	ASD	non-ASD
predicted as ASD	TP	FP
not predicted as ASD	FN	TN

For qualitative variables, precision, recall, and F-measure estimate interrater agreement from identified and expected accuracy to evaluate algorithm efficiency.

Precision is an excellent indication for validating the proposed ML model with predicted and actual outcomes. The predicted-to-actual positive ratio is estimated. This requires TP and FP. [20,21]

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall, another helpful assessment statistic to employ when attempting to estimate the number of positives that may be reasonably expected is the percentage of positives that were correctly classified. Recall is quantified using the values of TP and FN.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

The accuracy and recall of a classifier are balanced by the F-measure. The F-measure score is a

numerical value between (0 and 1) that represents the statistically significant metrics of recall and accuracy.

$$F - \text{measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (4)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \times 100\% \quad (5)$$

Where:

True positive (TP): The child has ASD and it is predicted correctly.

False positive (FP): The child does not have ASD, but predicted incorrectly(the child has ASD). This is Type 1 error.

True negative (TN): The child does not have ASD and predicted correctly.

False negative (FN): The child has ASD, but predicted incorrectly (the child does not have ASD). This is Type 2 error.

4. Results

In Tables (4 and 5), the assessment details of the five algorithms that were executed via a total of 1,250 cases for 10 characteristics and after using PCC technique to redact features to half (a total of five) separate features which is, gender, speech delay, jaundice, genetic condition, and familial history are presented.

Table 4. Ten-feature algorithm assessment metrics comparison

	NB		K-NN		DT		ABM1		SVM	
Accuracy	98.8		96.4		93.6		86.8		74.4	
F1-score	98.7		96.3		93.7		91.3		77.4	
Recall	99.1		97.5		91.6		86.4		69.1	
Precision	98.4		95.2		96		96.8		88	
Confusion matrix	123	2	119	6	120	5	121	4	110	15
	1	124	3	122	11	114	19	96	49	76

Table5. Five-feature algorithm assessment metrics comparison

	NB		K-NN		DT		ABM1		SVM	
Accuracy	99.2		97.2		94.8		87.6		80.4	
F1-score	99.1		97.1		94.8		88.6		82.9	
Recall	98.4		96.8		93		81.7		73.4	
Precision	99.9		97.6		96.8		96.8		95.2	
Confusion matrix	125	0	122	3	121	4	121	4	119	6
	2	123	4	121	9	116	27	98	43	82

The examination of the two tables reveals a notable observation: by downsampling the characteristics to

half, depending on the approach applied, has shown a large improvement across a variety of metrics

considered in the study. This suggests that feature selection has a significant role in improving the performance of prediction models by removing characteristics that are either redundant or irrelevant.

By focusing on the most relevant qualities, the model achieves greater accuracy while reducing computational complexity. This enhancement highlights the necessity of implementing robust down-sampling feature approaches to optimize machine learning algorithms for improved speed and reliability. Figure 2 depicts the predictive accuracy of the five algorithms used for all 10 characteristics, compared to the down-sampled features dataset.

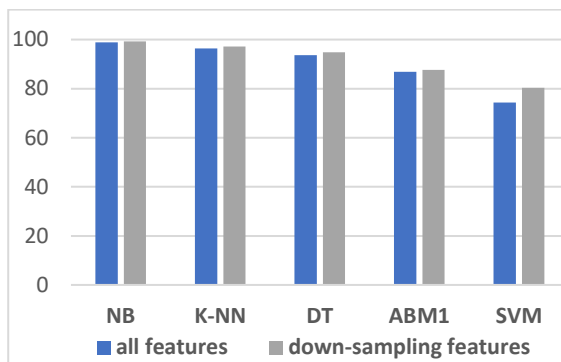


Figure 2. Accuracy of prediction

In the context of both scenarios, NB outperforms all other classifiers with an accuracy of 99.9% and 98.8%. The lowest accuracy is achieved by the SVM algorithm, at 74.4% and 80.4%, respectively.

The recall, accuracy, F1-score, and specificity of 10 characteristics, together with the down-sampling features, are employed to assess the importance of the results, as seen in Figures (3 and 4), respectively.

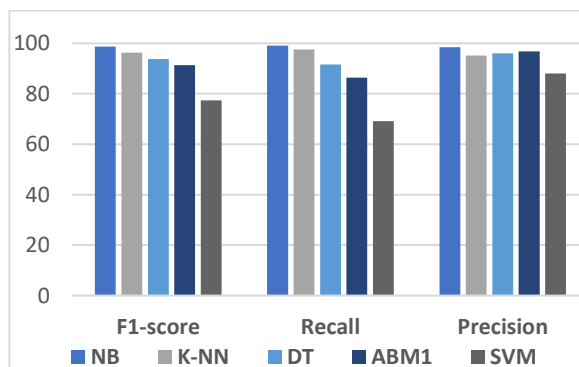


Figure 3: Performance metrics employing ten features

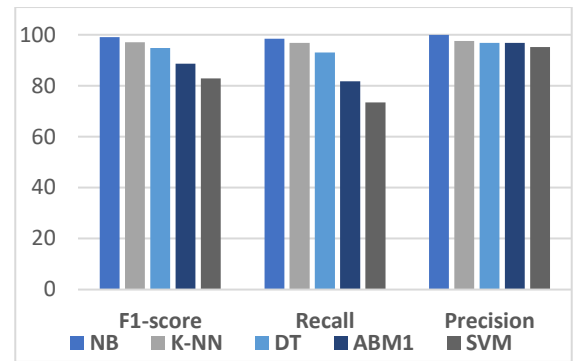


Figure 4. Performance metrics employing (down-sampling) features

In early autism screening, relying solely on accuracy can be misleading, even when the dataset is balanced. Accuracy does not reflect how well the model identifies true positive cases, which is critical in medical diagnosis. Sensitivity (recall) is a more important metric in this context because it measures the model's ability to correctly detect children with ASD. A model with high accuracy but low sensitivity may miss a significant number of true ASD cases (false negatives), leading to delayed diagnosis and missed opportunities for early intervention. Therefore, emphasizing sensitivity alongside other metrics such as specificity provides a more clinically meaningful evaluation of screening performance.

For classification algorithms, the unit of time can serve as a measure of the computational effort necessary to train and execute these algorithms. The time required for various ML algorithms depends on a variety of factors, including the size of the dataset, the complexity of the model, and the availability of computational resources.

The average time required to predict the outcome of each case of the data used for testing purposes and for each algorithm has been calculated the objective to determine the efficacy of the implemented algorithms. The time required for a model to predict or infer the output for new data is referred to as inference or prediction time. The prediction process was implemented using Python, and the reported execution time reflects pure inference only. Specifically, the measured time excludes any overhead related to model loading or initialization, focusing solely on the model's prediction performance. This ensures that the reported results accurately represent the computational efficiency of the inference stage. Figure 4 illustrates the total time required to

process the ASD dataset for each of the implemented algorithms.

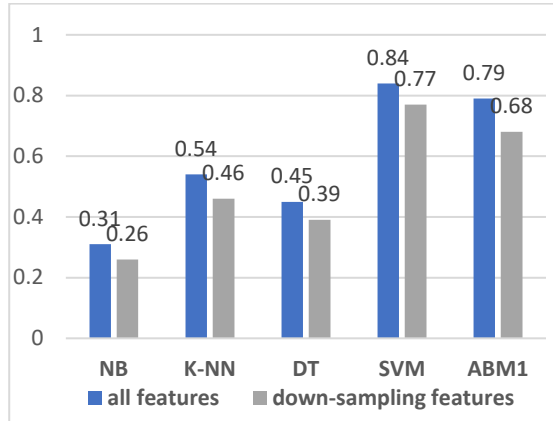


Figure 4. Time to estimate patient condition for Algorithms

It is evident that the minimum prediction time of the NB algorithm is 0.26 seconds, while the k-NN algorithm significantly increases this time to 0.46 seconds. Nevertheless, the DT algorithm has resulted in a minor decrease in time, with a value of 0.39

seconds. This is approximately half the time recorded by the SVM and ABM1 algorithms.

The contrasting findings with other works demonstrate that the proposed technique with down-sampling characteristics can accomplish improved accuracy along with prediction time. The test is also less inefficient due to the reduced number of characteristics. Therefore, the ASD prediction is less time-consuming, and the accuracy and overall performance are improved by ML algorithms that employ the feature reduction technique. Table 6 provides a comparison of the proposed techniques with state-of-the-art works.

Table 6: A comparison of the presented models with previous research.

	Vakadkar et al. [7]	Farooq et al. [14]	Proposed Method
Dataset	records:292 children	records:1054 children attributes:18	records:1250 children attributes:5
Method	NB, SVM, LR, CNN, NN	NB, SVM, LR, KNN, RF	NB, SVM, ABM1, KNN, DT
Accuracy	NB:97.53%, SVM:96.30%, CNN:95%, NN:94.5% LR:96.88%,	NB:94.7% SVM:93.8% LR:93.1% KNN:90.5% RF: 81.5%	NB:99.2% KNN:97.2% SVM:80.4% ABM1:87.6% DT:94.8%
Limitation	does not forecast the severity of ASD. In some instances, the symptoms utilized for diagnosing autism may not consistently reflect a true occurrence of the illness.	The dataset exhibited a significant imbalance, predominantly favoring ASD class, due to its primary compilation from autism-focused sources.	difficult to develop a universally applicable paradigm due to the fact that ASD presents uniquely in each child. Therefore, it is imperative to rigorously validate models with diverse populations.
Strength	High precision is attained by the utilization of five different ML classification approaches.	Five machine learning strategies were constructed and tested using the forward feature selection technique.	Balanced datasets consisting of 1,250 kids have been processed to predict Autism.

The correlation between the quantity of features and prediction accuracy is contingent upon the characteristics of the dataset and the caliber of the features.

5. Conclusions

Methods involving machine learning have been employed to predict the onset of the disease, as they are capable of accurately estimating the

disease's rate. The efficacy of ML methods for prediction was evaluated in the current study using a balanced ASD dataset with 1,250 records that were labeled. Furthermore, the importance of selecting the appropriate attributes to obtain an accurate prediction was evaluated for all of the algorithms that were employed. This work sought the most efficient ML approaches among several well-

accepted, simple-to-construct, fast, and well-performing algorithms.

The research results reveal that the Naïve Bayes(NB) method has higher performance relative to four other algorithms, achieving an accuracy of 99.2% in minimal prediction time, whereas the Support Vector Machine(SVM) attained the lowest accuracy of 80.4% with maximum prediction time when handling large datasets. SVM machine implementations need considerable training duration. likewise, the results clearly indicate that employing the PCC feature reduction strategy significantly improved the classifiers' overall performance by decreasing the number of features.

ML algorithms utilizing these selected characteristics provide enhancements in performance metrics as well as prediction time. The primary purpose of this work was to anticipate an essential event in children early and correctly using learning techniques. In the future, construct a hybrid model that integrates ML methods with optimization approaches to create an adaptive diagnostic model for autism.

This research is closely linked to several Sustainable Development Goals (SDGs), including SDG3, which focuses on good health and well-being for all ages. It also aligns with SDG10, which addresses reducing inequalities, including those related to disability. Furthermore, the research's contribution to enhancing technological capabilities, particularly in scientific research, helps achieve part of SDG9.

Acknowledgements

The authors express their sincere appreciation and gratitude to the University of Baghdad (www.uobaghdad.edu.iq) in Baghdad, Iraq, and to Al-Naji University (https://en.alnaji-uni.edu.iq) in Baghdad, Iraq, for their support.

Abbreviations

ASD	Autism spectral disorder
ABM1	AdaBostM1
AC	Association classification
ADHD	Attention-deficit/hyperactivity disorder
CDC	Center for disease control
CFS	Correlation feature set
CHI	Chi-square
DT	Decision tree
DT	Decision tree
FCBF	Fast correlation-based filter

FL	Federated learning
GI	Gini index
IG	Information gain
KNN	K-nearest neighbors
LR	Logistic regression
ML	Machine learning
NB	Naive Bayes
NN	Neural network
PCC	Pearson correlation coefficient
RF	Random forest
SVM	Support vector machine

Conflict of interest

The authors declare no conflicts of interest concerning this research."

Author Contribution

SWK. developed the theoretical framework verified the analytical methods. MAT proposed the research problem. Both authors participated in the discussion of the results and contributed to writing the manuscript.

AI Declaration Statement

The authors confirm that the manuscript has been written without the assistance of generative AI or AI-based writing tools.

References

- [1] S. R. Reeja and S. Mounika, "Autism Spectrum Disorder Classification Using Machine Learning and Deep Learning-A Survey," *EAI Endorsed Trans. Pervasive Health Technol.*, vol. 9, no. 1, May 2023. <https://doi.org/10.4108/eetpht.9.4240>
- [2] S. Hashem, S. Nisar, A. A. Bhat, et al., "Genetics of structural and functional brain changes in autism spectrum disorder," *Transl. Psychiatry*, vol. 10, art. no. 229, 2020. <https://doi.org/10.1038/s41398-020-00921-3>
- [3] V. Sirigiri and S. Katta, "Detection of Autism Spectrum Disorder Using Machine Learning: A Review," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 12, no. 3, pp. 824–830, Mar. 2024. <https://doi.org/10.22214/ijraset.2024.58939>
- [4] M. M. Rahman, O. L. Usman, R. C. Muniyandi, S. Sahran, S. Mohamed, and R. A. Razak, "A Review of Machine Learning Methods of Feature Selection and Classification for Autism Spectrum Disorder,"

- Brain Sci., vol. 10, no. 12, art. no. 949, Dec. 2020. <https://doi.org/10.3390/brainsci10120949>
- [5] M. Duda, R. Ma, N. Haber, and D. P. Wall, "Use of machine learning for behavioral distinction of autism and ADHD," *Translational Psychiatry*, vol. 6, no. 2, pp. e732-e732, 2016/02/01 2016, doi: <https://doi.org/10.1038/tp.2015.221>.
- [6] J. Alwidian, A. Elhassan, and R. Ghnemat, "Predicting Autism Spectrum Disorder Using Machine Learning Technique," *Int. J. Recent Technol. Eng. (IJRTE)*, vol. 8, no. 5, pp. 4139–4143, Jan. 2020. <https://doi.org/10.35940/ijrte.E6016.018520>
- [7] K. Vakadkar, D. Purkayastha, and D. Krishnan, "Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques," *SN Comput. Sci.*, vol. 2, no. 5, art. no. 399, Sep. 2021. <https://doi.org/10.1007/s42979-021-00776-5>
- [8] [U. A. Siddiqui et al., "Wearable-Sensors-Based Platform for Gesture Recognition of Autism Spectrum Disorder Children Using Machine Learning Algorithms," *Sensors*, vol. 21, no. 10, art. no. 3319, May 2021. <https://doi.org/10.3390/s21103319>
- [9] K. D. Rajab, A. Padmavathy, and F. Thabtah, "Machine Learning Application for Predicting Autistic Traits in Toddlers," *Arab. J. Sci. Eng.*, vol. 46, no. 4, pp. 3793–3805, Apr. 2021. <https://doi.org/10.1007/s13369-020-05165-3>
- [10] N. Pan et al., "Developing a Simplified Measure to Predict the Risk of Autism Spectrum Disorders: Abbreviating the M-CHAT-R Using a Machine Learning Approach in China," *Psychiatry Res.*, vol. 344, art. no. 116353, Feb. 2025. <https://doi.org/10.1016/j.psychres.2025.116353>
- [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, May 2022. <https://doi.org/10.3390/a15050166>
- [12] . V. Shinde and D. D. Patil, "A Multi-Classifer-Based Recommender System for Early Autism Spectrum Disorder Detection Using Machine Learning," *Healthcare Anal.*, vol. 4, art. no. 100211, Dec. 2023. <https://doi.org/10.1016/j.health.2023.100211>
- [13] D. D. Khudhur and S. D. Khudhur, "The Classification of Autism Spectrum Disorder by Machine Learning Methods on Multiple Datasets for Four Age Groups," *Meas. Sensors*, vol. 27, art. no. 100774, Jun. 2023. <https://doi.org/10.1016/j.measen.2023.100774>
- [14] M. S. Farooq, R. Tehseen, M. Sabir, and Z. Atal, "Detection of Autism Spectrum Disorder (ASD) in Children and Adults Using Machine Learning," *Sci. Rep.*, vol. 13, no. 1, art. no. 9315, Dec. 2023. <https://doi.org/10.1038/s41598-023-35910-1>
- [15] A. Das, P. Pattanaik, S. Mukherjee, S. Turjya, and A. Bandopadhyay, "Early Autism Spectrum Disorder Screening in Toddlers: A Comprehensive Stacked Machine Learning Approach," *Int. J. Comput. Digit. Syst.*, vol. 17, no. 1, pp. 1–12, Jan. 2025. <https://doi.org/10.12785/ijcds/1571010950>
- [16] R. Kumar et al., "Classification of COVID-19 from Chest X-Ray Images Using Deep Features and Correlation Coefficient," *Multimed. Tools Appl.*, vol. 81, no. 19, pp. 27631–27655, Aug. 2022. <https://doi.org/10.1007/s11042-022-12500-3>
- [17] Kaggle, "Medical Data." [Online]. Available: <http://www.kaggle.com/datasets?search=medical>. [Accessed: Feb. 10, 2025].
- [18] P. Chen, F. Li, and C. Wu, "Research on Intrusion Detection Method Based on Pearson Correlation Coefficient Feature Selection Algorithm," *J. Phys.: Conf. Ser.*, vol. 1757, no. 1, art. no. 012054, Feb. 2021. <https://doi.org/10.1088/1742-6596/1757/1/012054>
- [19] Z. Šverko, M. Vrankić, S. Vlahinić, and P. Rogelj, "Complex Pearson Correlation Coefficient for EEG Connectivity Analysis," *Sensors*, vol. 22, no. 4, art. no. 1477, Feb. 2022. <https://doi.org/10.3390/s22041477>
- [20] W. Al-Salman, Y. Li, A. Y. Oudah, and S. Almagd, "Sleep Stage Classification in EEG Signals Using the Clustering Approach-Based Probability Distribution Features Coupled with Classification Algorithms," *Neurosci. Res.*, vol. 188, pp. 51–67, Mar. 2023. <https://doi.org/10.1016/j.neures.2022.09.009>

- [21] F. Hajjej, S. Ayouni, M. A. Alohal, 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. <https://doi.org/10.1109/ACCESS.2024.3349988>