Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.062643

ARTICLE

Awareness with Machine: Hybrid Approach to Detecting ASD with a Clustering

Gozde Karatas Baydogmus* **and Onder Demir**

Department of Computer Engineering, Marmara University, Istanbul, 34854, Turkey

*Corresponding Author: Gozde Karatas Baydogmus. Email: gkaratas@marmara.edu.tr

Received: 23 December 2024; Accepted: 20 May 2025; Published: 03 July 2025

ABSTRACT: Detection of Autism Spectrum Disorder (ASD) is a crucial area of research, representing a foundational aspect of psychological studies. The advancement of technology and the widespread adoption of machine learning methodologies have brought significant attention to this field in recent years. Interdisciplinary efforts have further propelled research into detection methods. Consequently, this study aims to contribute to both the fields of psychology and computer science. Specifically, the goal is to apply machine learning techniques to limited data for the detection of Autism Spectrum Disorder. This study is structured into two distinct phases: data preprocessing and classification. In the data preprocessing phase, four datasets—Toddler, Children, Adolescent, and Adult—were converted into numerical form, adjusted as necessary, and subsequently clustered. Clustering was performed using six different methods: K-means, agglomerative, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), mean shift, spectral, and Birch. In the second phase, the clustered ASD data were classified. The model's accuracy was assessed using 5-fold cross-validation to ensure robust evaluation. In total, ten distinct machine learning algorithms were employed. The findings indicate that all clustering methods demonstrated success with various classifiers. Notably, the K-means algorithm emerged as particularly effective, achieving consistent and significant results across all datasets. This study is expected to serve as a guide for improving ASD detection performance, even with minimal data availability.

KEYWORDS: ASD; ASD detection; machine learning; clustering methods

1 Introduction

In today's world, ASD has gained significant attention. Diagnostic methods typically focus on behavioral symptoms in social, sensory, and motor skills. Recent studies indicate that ASD affects approximately 1 in 36 children in the United States, with prevalence rates rising globally [1]. This increase highlights the urgent need for accessible, efficient, and scalable diagnostic methods. Recent advances in technology, machine learning, and data analysis are improving quantitative and ecological validation methods. However, clinical screening tests remain expensive and time-consuming [2].

Machine learning has shown remarkable success across fields, providing techniques for learning, detection, data analysis, and pre-processing. Rapid developments in computer science have enabled prediction models that integrate multiple disciplines [3,4]. This study aims to develop a machine learning-based estimator for ASD detection across different age groups using limited data.

Early diagnosis is crucial, yet obtaining well-structured datasets remains a challenge. A dataset covering four age groups—Toddler, Children, Adolescent, and Adult—was selected. Six clustering techniques (k-means, agglomerative, DBSCAN, mean shift, spectral, and birch) and ten machine learning classifiers (logistic regression, support vector machines, k-nearest neighbor, multi-layered perceptron, extra tree classifier, Gaussian process classifier, passive aggressive classifier, ridge, stochastic gradient descent, and linear support vector machines) were employed. The prediction performance of these models was evaluated through clustering on ASD datasets.

Despite numerous studies, no research aligns precisely with the objectives of this study. Existing approaches primarily focus on popular machine learning and deep learning models with an emphasis on feature extraction rather than processing. Proper feature normalization significantly impacts model performance. This research aims to bridge this gap by designing a clustering-assisted classification model, contributing to cost-effective and efficient ASD screening solutions. Considering these aspects, the study aimed to address the following questions:

- How do alternative machine learning algorithms, apart from the commonly used ones, affect the model's performance?
- How can the model's performance be enhanced without reducing the number of features?

The designed model consists of two phases: data preprocessing and classification. Numerical transformations of the four ASD datasets selected in the data preprocessing phase were performed, and clustering algorithms were used. Machine learning algorithms selected with a clustered dataset were trained in the classification phase, and the prediction results were observed. In the next part of the study, information about the related work is given in Section 2; in Section 3, the materials used in the development of the study and the proposed model are explained; in Section 4, experimental results are given; in Section 5, the experimental results are discussed, and in Section 6, the results are concluded.

2 Related Work

ASD detection with Machine Learning (ML) has just begun to attract attention; not many journal studies on this subject have been found in the literature. Therefore, conference publications have been added to the relevant studies' titles.

Abdelwahab and others explored the use of ML to improve ASD diagnosis [5]. Using publicly available datasets from Kaggle and UCI ML, the researchers tested several ML algorithms. Data preprocessing involved feature selection, encoding, and normalization. Among the algorithms, Random Forest achieved the highest accuracy at 99.75%, while Logistic Regression also performed well at 96.69%. The findings highlight ML's potential to complement traditional ASD diagnosis, enabling earlier intervention and reducing costs.

In 2024, researchers compared two AutoML tools—TPOT and KNIME—for ASD detection using data from rehabilitation centers in Pakistan [6]. Both tools automated feature selection and model tuning using the Q-CHAT-10 questionnaire. TPOT achieved 85.23% accuracy, while KNIME reached 83.89%, with the Q-CHAT-10 score identified as the most important predictor. The study highlights AutoML's potential to streamline ASD diagnosis, making ML more accessible to healthcare professionals while improving early detection and treatment.

Xu et al. developed a method to detect ASDs in EEG (Electroencephalogram) datasets without using data augmentation methods [7]. They collected data from 97 ASD and 92 typically developing individuals from publicly available datasets. The data was collected during rest and while performing a task. They designed and implemented a combined network based on convolutional neural network (CNN) and long short-term memory (LSTM) for ASD detection. The developed network achieved classification accuracies of 81.08% and 74.55% for resting state and task state data, respectively.

Dia et al. proposed a supervised learning method to classify Autism Spectrum Disorder and to assess emotion levels among autistic children [8]. To evaluate the performance of the proposed approach, they used YouTube video frames of autistic children exhibiting typical autistic behaviors in unconstrained environments and conditions, as well as images of neurotypical people. They also proposed an extended version of a dataset containing additional influence labels corresponding to the influence levels of autistic children. Experiments were conducted using different models to determine the optimal performance of their architecture.

In 2024, researchers explored how the use of AI (Artificial Intelligence), particularly ML and deep learning (DL), can improve ASD detection [9]. They used natural language processing (NLP) to analyze Twitter posts, aiming to identify linguistic patterns associated with ASD. Various models, including decision trees, XGBOOST (eXtreme Gradient Boosting), k-nearest neighbors (KNN), Recurrent neural network (RNN), long short-term memory (LSTM), bidirectional long short-term memory (Bi-LSTM), and BERT (Bidirectional Encoder Representations from Transformers)-based models, were tested on a dataset of 404,627 tweets. BERTweet achieved the highest accuracy of 87.7%, demonstrating AI's potential in ASD diagnosis.

Researchers reviewed AI-based methodologies for ASD detection through computer vision techniques in 2024 [10]. They studied ML models such as SVM, decision trees, and gradient boosting, alongside deep learning models like CNNs, RNNs, LSTMs, and Transformer-based approaches. They proposed a binary image classifier using the Xception CNN model trained on facial images of children aged 2 to 8 years. With a dataset of 23,000 images, the model achieved an accuracy of 88.87%, highlighting the effectiveness of facial analysis in ASD detection.

Loganathan et al. developed a hybrid ensemble model combining ResNet101 and BiGRU networks optimized with the CHGSO algorithm for ASD detection using EEG signals [11]. The hybrid ensemble model shows superior performance in ASD detection compared to existing methods such as DNN (Deep Neural Networks), SVM (Support Vector Machine), KNN, and MGOA-RF. Their Hybrid ensemble model reaches Sensitivity of 98%, 99% higher Specificity, 98% F1-Score, MCC of 99%, Accuracy of 98%, and Precision of 99%.

ML for ASD detection faces challenges such as limited datasets, symptom variability, and model interpretability. Small sample sizes can lead to overfitting, while diverse symptom presentations make pattern recognition difficult. Additionally, selecting relevant features and ensuring model reliability remain key hurdles. Addressing these issues requires robust preprocessing and validation techniques.

3 Methods

In this section, the datasets, clustering techniques and machine algorithms used in the study are mentioned, and then detailed information about the proposed model is given.

3.1 Datasets

For this research purpose, 4 publicly available ASD datasets from Kaggle repository were used. Accordingly, the dataset for Toddler was taken from Kaggle [12], and the datasets for Children, Adult and Adolescent were taken from the UCI repository [13–15]. Detailed information about the datasets is given in Table 1.

While authorities create these datasets, ten behavioral traits (AQ-10) and different individual traits were used that have proven effective in detecting cases of ASD from behavioral science controls. In addition, there are two classes in all datasets [12–15], ASD and non-ASD. Therefore, binary classification was performed in all the methods used. When Table 1 is examined, it is seen that there is an imbalanced distribution in the Toddler and Adult datasets. For children and adolescents, it is seen that the data numbers of the classes are more balanced. Additionally, special attention should be given to the "Number of data in classes column", which indicates the class distribution in the datasets. Notably, a significant observation emerges in Table 1; the Toddler dataset exhibits a much larger sample size for ASD compared to non-ASDs, whereas the Adult dataset presents the opposite scenario. This is definitely a situation that will affect the classification because of the imbalance dataset problem [16].

Table 1: Information about datasets

Dataset name	Alias for dataset name	Feature type	Number of features	Number of data in classes	Number of data
Autism screening data for toddlers	Toddler	Categorical, continuous and binary	18	no: 326, yes: 729	1054
Autistic spectrum disorder screening data for children	Children	Categorical, continuous and binary	21	no: 151, yes: 141	292
Autism screening adult	Adult	Categorical, continuous and binary	21	no: 515, yes: 189	704
Autistic spectrum disorder screening data for adolescent	Adolescent	Categorical, continuous and binary	21	no: 41, yes: 63	104

3.2 Used Techniques

In this section, the methods used in the two phases of the study are mentioned.

3.2.1 Clustering Algorithms

This section provides information about the clustering algorithms employed in the study [17–20]. These algorithms, chosen for their popularity and ease of use, are as follows: K-means, Agglomerative Clustering, DBSCAN, MeanShift, Spectral Clustering, and Birch [17–20].

3.2.2 ML Algorithms

In this section, brief information about the ML algorithms utilized in the study is provided. The following algorithms were examined [16,21,22]: Extra Trees Classifier (ETC) [23], Gaussian Process Classifier (GPC) [24], KNN [16], Linear Support Vector Machines (LSVC) and Support Vector Machines (SVM) [25], Logistic Regression (LR), Multi-Layered Perceptron (MLP), Passive Aggressive Classifier (PAC), Ridge Classifier (RC) [26], and Stochastic Gradient Descent (SGDC) [27].

3.3 Performance Metrics

Classification stands as a fundamental challenge in the field of ML, encompassing the task of forecasting the class labels of given input data. To gauge the performance of such models, the accuracy score emerges as a widely employed evaluation measure. It quantifies the proportion of accurate predictions made by the model in relation to the total number of predictions conducted. In addition, accuracy score, F1-score, ROC (Receiver Operating Characteristic)/AUC (Area Under the Curve), and values are calculated [16].

3.4 Proposed Model

In recent years, substantial efforts have been dedicated to enhancing ASD classification, and this research continues to progress. Upon examining the studies, it becomes evident that both image and text-based datasets were utilized.

However, it was observed that these datasets lack sufficient data, particularly in the case of text-based datasets, as they require user trust in the clinical environment, resulting in fewer attributes and data entries. Taking all of these factors into consideration, this study aims to investigate the impact of ML approaches, which is a popular topic today, on improving ASD detection with a limited amount of data. In this manner, the research consists of two phases:

1. Clustering the dataset, which is called "Data Preprocessing".
2. Applying selected ML algorithms on both clustered datasets, which is called "Classifier Training".

Fig. 1 illustrates the phases of the proposed model and provides further details, which will be discussed in depth in the following section.

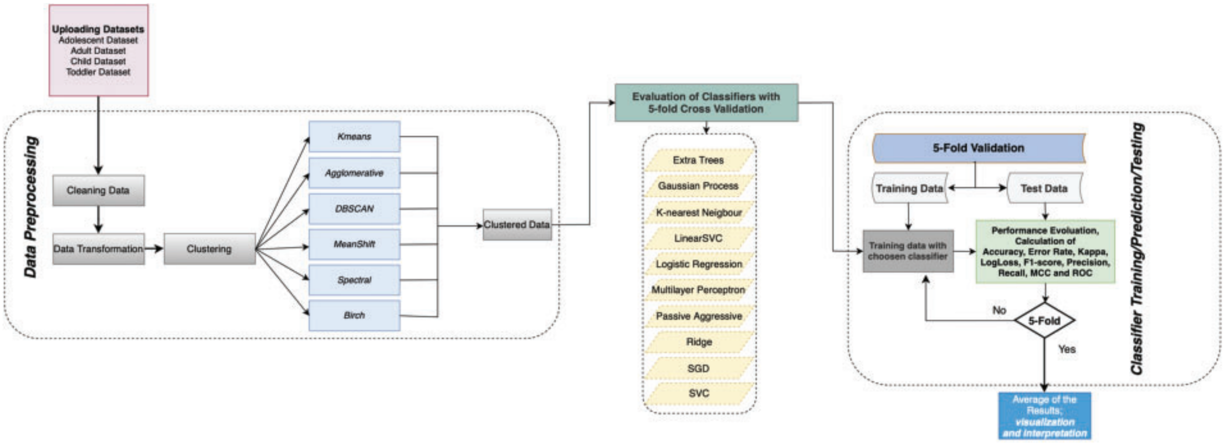

Figure 1: Flowchart of the proposed model

3.4.1 Environment and Development

One of the most crucial aspects in artificial intelligence studies is the development environment and techniques employed. Providing information about the development environments in research studies aims to guide researchers and prevent any inaccuracies. The working environment utilized for the proposed model is outlined in Table 2. Additionally, the Python programming language was employed for data preprocessing and applying ML techniques. Python has become a frequently preferred language for artificial learning approaches in recent years, offering a delightful development experience. Its availability of various artificial learning modules facilitates operations with ease. In addition, these modules can be customized according to user requirements, making them conducive to further development. Throughout this study, all the ML algorithms used retained their default Python values. In other words, no changes were made to any parameters of the algorithms, and they were run as defined in Python.

Table 2: Development environment

Hardware	Properties
CPU	Intel(R) Core(Tm) I7-8750H Cpu @2.20 GHz, 6 Cores
Op. Syst.	64 bits, Windows 11
Graphic Card	GTX 1650
L1/L2/L3Cache	384 KB/1.5 MB/9.0 MB
RAM	16.00 GB
Python version	3.9 64-bit

3.4.2 Proposed Algorithm

In this section, detailed information about the proposed hybrid method is given and how to implement the algorithm is explained step by step.

All operations were executed in a consistent manner for all four datasets.

1. The dataset has been transformed into a numerical format for mathematical operations. During these conversion processes, categorical data were organized, missing data were addressed, and labels were converted into numerical values.
    - The dataset was converted into a numerical format to facilitate mathematical operations required for ML models. This process included structuring categorical data, handling missing values, and transforming labels into numerical representations to ensure consistency across all features.
    - Since all categorical variables in the dataset were nominal (i.e., they do not have an inherent order or ranking), no specialized encoding techniques such as ordinal encoding were necessary. Instead, these categorical values were directly transformed into numerical representations while preserving their original properties.
    - Some columns contained an unique categorical entries, particularly ethnicity, country of residence, and relation. To manage this effectively, Label Encoding was used instead of One-Hot Encoding. This decision was made to avoid a significant increase in feature dimensionality, which could lead to excessive sparsity and computational inefficiencies. Label Encoding assigned each category a unique numerical value while preserving the dataset's structure and preventing unnecessary expansion of features.
    - Certain columns in the dataset contained missing values, represented by the symbol '?'. These missing entries were systematically addressed to maintain data integrity. Depending on the nature of the missing data, appropriate techniques such as imputation (e.g., replacing missing values with the mode or median) or row-wise removal were applied to ensure the dataset remained complete and suitable for ML analysis.
2. After the numerical operations were performed on the dataset, clustering was done separately with the selected clustering algorithms. Then, this clustered dataset was trained with ML algorithms.
3. In addition, a normality test was conducted using hypothesis tests for the data sets presented in Table 1. Specifically, the Shapiro-Wilk test was applied individually to each dataset. According to this hypothesis test, if the p-value of the examined data is equal to or above 0.05, it follows a normal distribution; otherwise, it does not.

The second phase of the study involves applying ML algorithms to the clustered dataset and examining the determined "Performance Metrics". Although 25 ML algorithms were initially applied, only 10 of them were ultimately selected. These selected algorithms are briefly summarized under the title of "ML Algorithms". The reason for choosing these specific algorithms is that their performance ratios remained consistent regardless of clustering. Remarkable improvements were observed in the 10 algorithms examined and proposed in the study.

Additionally, the 5-fold cross-validation method was employed during the training and testing phases of the study. This approach ensured more robust testing and estimation processes.

The decision to perform the cross-validation process five times is based on recommendations in the literature for ML algorithms [16]. This number is considered optimal for obtaining reliable results. Algorithm 1 shows the pseudo code of the proposed method.

Algorithm 1: Hybrid ASD detection model

```
Input: Four ASD datasets (Toddler, Children, Adolescent, Adult)
Output: Classification results with performance metrics
Step 1: Preprocessing
for each dataset do
    Convert categorical data using LabelEncoding
    Handle missing values ('?')
end for
Step 2: Clustering
for each dataset do
    Apply clustering (k-means, agglomerative, DBSCAN, mean-shift, spectral, birch)
end for
Step 3: Normality Test
for each dataset do
    Perform Shapiro-Wilk test (Check $p$-value for normality)
end for
Step 4: Model Training & Evaluation
for each dataset do
    Train 10 ML models on clustered data
    Evaluate using performance metrics
end for
Step 5: 5-Fold Cross-Validation
for each model do
    Perform 5-fold cross-validation
    Compute average scores
end for
Step 6: Compare & Report Results
Return: best model performances and insights
```

4 Result

In this section, the proposed approach for the study was implemented, and all experiments were conducted in the environment specified under the title "Environment and Development".

The results were evaluated separately for the clustered dataset. In the subsequent section, the outcomes obtained for each performance metric will be presented and thoroughly analyzed. Various clustering and

ML approaches were considered in the study, but only the most prominent ones were included. Algorithms such as random forest and decision tree, which are commonly used in the literature, were excluded from the study, as there are already sufficient studies available about these algorithms. Additionally, algorithms with low accuracy were not included in the study to focus on the most effective ones.

As in all ML studies, the accuracy rate was first calculated in this study. The accuracy rate results are given in Table 3.

Table 3: Accuracy scores with and without clustering using ML algortihms

		Algorithms									
Dataset	**Clustering**	**ETC**	**GPC**	**KNN**	**LSVC**	**LR**	**MLP**	**PAC**	**RC**	**SGD**	**SVC**
Children	Agglomerative	98.29	100.00	75.00	95.89	100.00	81.51	**81.16**	96.92	**80.48**	81.85
	Birch	**98.63**	100.00	75.00	94.18	100.00	**84.59**	**82.53**	96.92	76.37	81.85
	DBSCAN	97.60	100.00	75.00	**96.58**	100.00	**85.96**	**83.90**	96.92	**79.79**	81.51
	KMeans	98.29	100.00	75.00	93.84	100.00	80.82	**72.60**	96.92	78.77	81.51
	MeanShift	98.29	100.00	75.00	94.86	100.00	80.82	**75.00**	96.92	69.18	81.51
	NONE	98.29	100.00	75.00	95.89	100.00	82.88	72.26	96.92	79.11	81.85
	Spectral	**98.63**	100.00	75.00	92.12	100.00	85.27	**72.60**	96.92	**87.33**	81.85
Toddler	Agglomerative	98.86	100.00	97.72	100.00	100.00	**98.67**	**95.64**	94.97	99.62	99.43
	Birch	98.77	100.00	97.53	100.00	100.00	97.82	**96.58**	95.35	98.01	99.43
	DBSCAN	**99.34**	100.00	97.91	100.00	100.00	**99.81**	93.83	95.54	98.01	99.43
	KMeans	98.96	100.00	**98.01**	100.00	100.00	98.20	**96.39**	95.16	96.87	99.43
	MeanShift	99.05	100.00	97.82	100.00	100.00	**98.67**	**94.88**	95.35	99.53	99.43
	NONE	99.15	100.00	97.82	100.00	100.00	98.48	94.40	95.54	99.81	99.43
	Spectral	99.24	100.00	97.82	99.81	100.00	**98.58**	**96.39**	95.35	98.39	99.43
Adolscent	Agglomerative	95.19	94.23	82.69	**82.69**	94.23	75.96	**81.73**	87.50	82.69	81.73
	Birch	95.19	94.23	82.69	**83.65**	94.23	**87.50**	78.85	87.50	83.65	81.73
	DBSCAN	94.23	94.23	82.69	81.73	94.23	**81.73**	77.88	89.42	78.85	81.73
	KMeans	94.23	94.23	81.73	80.77	94.23	**83.65**	72.12	89.42	81.73	81.73
	MeanShift	94.23	94.23	81.73	81.73	93.27	**85.58**	80.77	87.50	**84.62**	81.73
	NONE	97.12	94.23	82.69	81.73	94.23	79.81	80.77	89.42	83.65	81.73
	Spectral	94.23	94.23	82.69	80.77	94.23	**85.58**	**83.65**	87.50	81.73	81.73
Adult	Agglomerative	**98.44**	73.15	77.70	91.05	**99.86**	**93.47**	**76.56**	94.89	**79.83**	83.66
	Birch	**98.15**	73.15	77.70	**92.47**	**99.86**	**92.61**	**78.55**	94.89	**86.51**	83.66
	DBSCAN	**98.30**	73.15	77.70	**92.90**	98.58	**87.64**	**78.69**	94.89	75.99	83.66
	KMeans	**98.86**	73.15	77.70	**93.89**	98.44	**89.49**	67.61	**95.03**	**86.65**	83.66
	MeanShift	**98.15**	73.15	77.70	**93.89**	**99.72**	**91.90**	**81.25**	94.89	**81.68**	83.66
	NONE	97.87	73.15	77.70	91.90	99.43	84.23	75.57	94.89	79.40	83.66
	Spectral	**98.58**	73.15	77.70	89.63	99.43	84.09	**78.84**	94.89	**82.95**	83.66

Table 3 displays the datasets in the leftmost column, followed by the clustering methods and their accuracy rates. Initially, each dataset was classified without clustering. Notably, some values are written in bold to emphasize the increase in accuracy rate, which will also be applied in other tables. A careful examination of Table 3 reveals that most of the clustering methods enhance the performance of algorithms. Particularly, almost all algorithms with Spectral demonstrated an increase in accuracy rate across all datasets. Furthermore, there is a noticeable difference in accuracy improvement between the Toddler and Adult datasets, as indicated in the “Datasets” title. In the Adult dataset, clustering improved accuracy rates for all algorithms, with Extra Trees Classifier (ETC) showing particularly notable success. This is because the number of non-ASD samples is higher in the Adult dataset. Continuing the analysis of Table 3, the most successful algorithms were found to be MLP, PAC, and ETC. These algorithms either maintained or increased accuracy rates across all clustering methods. The ROC curve is a very important performance measure for classification problems. ROC is a probability curve, and the area under it, AUC, represents the degree or measure of separability. For this reason, these values are very important in solving classification problems. Fig. 2 shows the performance of classification algorithms in detecting ASD with and without clustering.

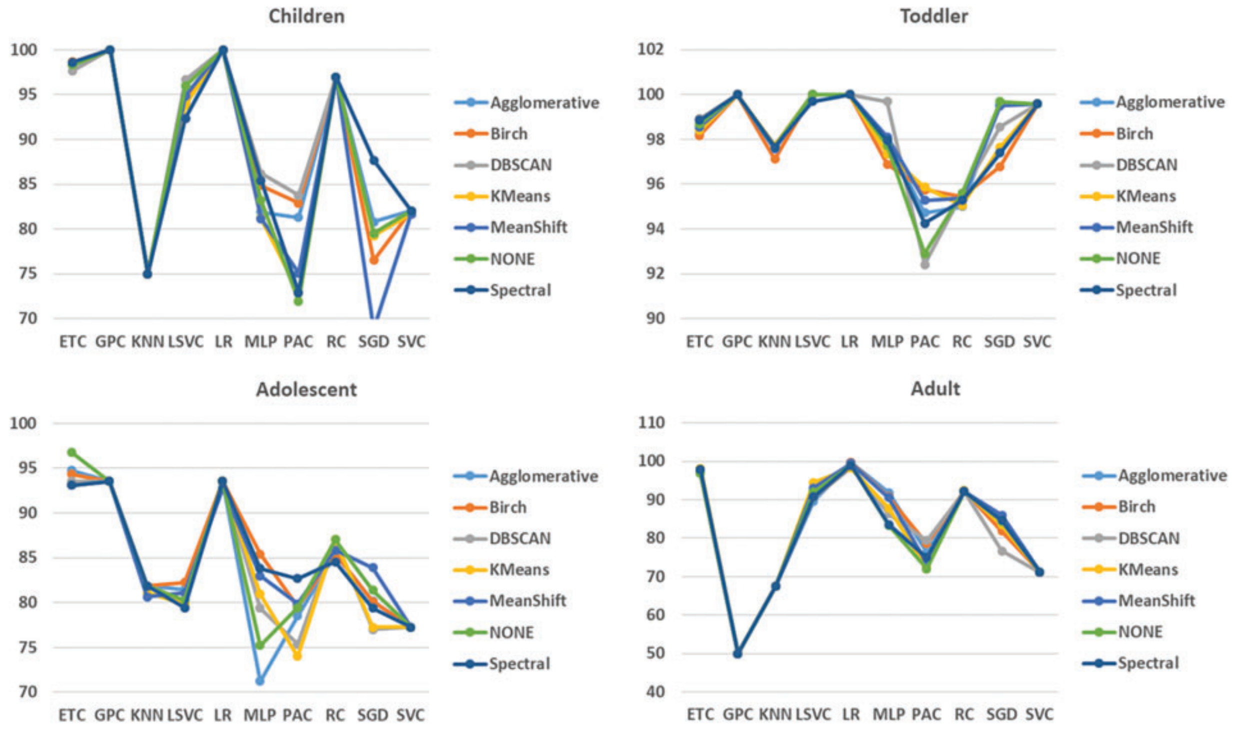

Figure 2: AUC/ROC scores for all datasets with and without clustering

For the model, the ROC curve was also drawn for each execution. However, they are not given here because they seem too complex and are too numerous. Table 4 shows the F1-score of classification algorithms in detecting ASD with and without clustering.

Previously given metrics cannot give a complete result for imbalanced datasets. Basically, it is the MCC criterion that evaluates by looking at the correlation (phi-coefficient) relationship between the actual data and the predicted data. Since the Toddler and Adult dataset has an uneven distribution, the clustering algorithms and classification methods vary. However, the result still does not change. In ASD detection, ETC, MLP, PAC, and RC algorithms, together with the Spectral and DBSCAN algorithms, show great performance and make accurate detection.

Table 4: F1-scores with and without clustering using ML algorithms

		Algorithms									
Dataset	**Clustering**	**ETC**	**GPC**	**KNN**	**LSVC**	**LR**	**MLP**	**PAC**	**RC**	**SGD**	**SVC**
Children	Agglomerative	98.23	100.00	74.39	95.89	100.00	82.91	81.36	96.86	81.79	82.27
	Birch	98.59	100.00	74.39	93.99	100.00	85.34	83.71	96.86	76.61	82.27
	DBSCAN	97.54	100.00	74.39	96.55	100.00	86.47	82.78	96.86	77.74	81.88
	KMeans	98.23	100.00	74.39	93.66	100.00	82.28	74.36	96.86	80.86	81.88
	MeanShift	98.25	100.00	74.39	94.74	100.00	82.05	74.91	96.86	65.38	81.88
	NONE	98.23	100.00	74.39	95.83	100.00	83.87	68.73	96.86	80.76	82.27
	Spectral	98.58	100.00	74.39	92.41	100.00	85.42	74.19	96.86	88.1	82.27
Toddler	Agglomerative	99.18	100.00	98.34	100.00	100.00	99.05	96.85	96.31	99.73	99.59
	Birch	99.11	100.00	98.21	100.00	100.00	98.43	97.54	96.59	98.58	99.59
	DBSCAN	99.52	100.00	98.48	100.00	100.00	99.86	95.56	96.73	98.54	99.59
	KMeans	99.25	100.00	98.56	100.00	100.00	98.71	97.39	96.46	97.68	99.59
	MeanShift	99.32	100.00	98.42	100.00	100.00	99.04	96.21	96.59	99.66	99.59
	NONE	99.38	100.00	98.41	100.00	100.00	98.91	95.98	96.73	99.86	99.59
	Spectral	99.45	100.00	98.41	99.86	100.00	98.98	97.45	96.59	98.85	99.59
Adolscent	Agglomerative	96.06	95.31	85.71	85.94	95.31	82.52	86.13	90.23	86.57	86.71
	Birch	96.12	95.31	85.71	86.82	95.31	90.23	81.36	90.23	87.77	86.71
	DBSCAN	95.31	95.31	85.71	85.27	95.31	85.71	82.71	91.85	83.08	86.71
	KMeans	95.38	95.31	84.8	84.13	95.31	87.41	73.87	91.85	86.71	86.71
	MeanShift	95.38	95.31	85.04	84.8	94.49	88.89	84.13	90.08	87.30	86.71
	NONE	97.64	95.31	85.71	85.27	95.31	85.31	84.38	91.85	87.22	86.71
	Spectral	95.38	95.31	85.71	84.38	95.31	88.55	86.61	90.51	85.71	86.71
Adult	Agglomerative	97.04	74.83	52.28	83.89	99.73	87.89	63.41	90.11	71.49	59.36
	Birch	96.50	74.83	52.28	86.38	99.73	86.32	66.37	90.11	74.11	59.36
	DBSCAN	96.77	74.83	52.28	86.63	97.37	78.41	67.11	90.06	63.66	59.07
	KMeans	97.86	74.83	52.28	89.38	97.11	81.12	57.78	90.36	75.52	59.36
	MeanShift	96.5	74.83	52.28	88.95	99.47	85.35	61.85	90.06	73.62	59.36
	NONE	95.98	74.83	52.28	85.93	98.93	73.51	58.65	90.06	71.29	59.36
	Spectral	97.33	74.83	52.28	82.82	98.93	73.58	62.84	90.06	73.57	59.36

5 Discussion

When the literature and existing papers are examined, it is seen that many researchers tend to solve the issue of ASD detection. The main goal of the study is to design a model that will increase the detection performance without interfering with the number of features in a small sample size. This study aims to design an ASD detection system for people of different age groups. Since diagnosis is very important for ASD, studies in this area are very important. Detection of ASD is very difficult, especially in age groups with small data and limited number of features.

The study examined the effect of six clustering methods on ASD datasets and the rate of improvement in classification. For this, the selected datasets were clustered and then evaluated with the specified performance metrics. The classification results were analyzed, the changes in detection results after clustering were observed, and a hybrid model was proposed. Fig. 3 shows total time of clustering and classification for all algorithms.

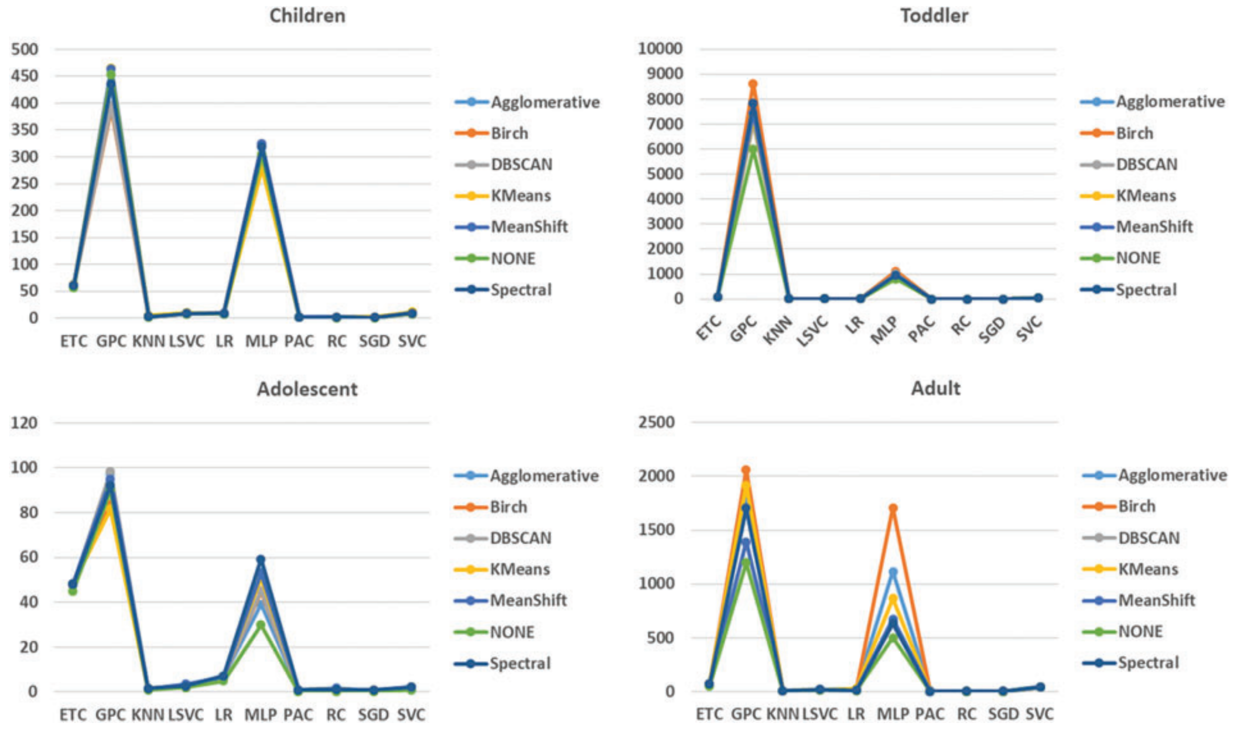

Figure 3: Time for all datasets with and without clustering using ML algortihms

It is seen that almost every clustering method works successfully in certain algorithms with every dataset. However, the Spectral method stands out in this sense. The values obtained as a result of clustering the ASD datasets with Spectral increased the detection of ASD in all datasets compared to the unclustered state. Spectral Clustering and DBSCAN have shown superior performance over other clustering algorithms, particularly in small datasets, due to their ability to capture complex data structures. Spectral leverages graph-based techniques, transforming data into a similarity matrix before applying clustering.

The results of the study are summarized in the following section:

1. Using clustering before classification, the dataset for ASD detection generally improves performance. The quality and size of the dataset are crucial factors for building an effective prediction model. Clustering has played a significant role in improving dataset quality, ultimately leading to the creation of more successful prediction models, through the increased availability of larger datasets.
2. Particularly in ASD imbalanced datasets, the large number of non-ASD samples enhances the model's success rate.
3. Clustering the data allows it to be brought within a certain range, leading to more consistent model performance. In this regard, the Spectral and MLP methods can be preferred as an option for the classification of ASD datasets.
4. In the study, it was observed that the prediction rate was increased through correct pre-processing in datasets that did not have a normal distribution.

Overall, the study highlights the importance of clustering techniques in improving the detection of ASD and identifies specific algorithms that perform exceptionally well in different dataset scenarios.

Table 5 has the accuracy rate comparison with similar ASD studies found in the literature. The values for [28] and [29]are based on the reference [29], and the values for [30] represent the overall average accuracy achieved by their proposed models.

Table 5: Comparison of accuracy with other studies

	Accuracy scores (%)			
Datasets	[28]	[29]	[30]	**Proposed study**
Toddler		98.77		99.34
Children	97.80	97.20	96.04	98.60
Adolescent	94.23	93.89	99.95	87.50
Adult	99.85	98.36	97.32	99.86

The study has several limitations. First, while four different ASD datasets were used, the results may not fully generalize to other datasets or populations. The sample size in certain age groups remained relatively small, which could limit the model's performance in real-world, larger datasets. Additionally, the datasets were imbalanced, with more non-ASD samples than ASD samples, potentially influencing the model's ability to accurately detect ASD. Although Spectral Clustering and DBSCAN showed strong performance, their effectiveness may vary with different datasets, which limits the broader applicability of the findings. The preprocessing steps played a significant role in the model's success, but these methods may not be equally effective for datasets with different distributions. Finally, while the classification algorithms demonstrated improved performance with certain clustering methods, the results may not be consistent across all algorithms or datasets.

6 Conclusion and Future Work

Detection of ASD is a critical area of research in psychology, especially with the rise of technology and artificial learning approaches. This study aimed to improve ASD detection in different age groups, particularly focusing on performance enhancement with small sample sizes. A hybrid model was proposed that integrates clustering and classification techniques, evaluating various clustering methods on six different ASD datasets. The results show that clustering significantly improved the performance of the 10 ML algorithms tested, with Spectral Clustering and the ETC, MLP, PAC, and RC algorithms yielding the most prominent improvements. Importantly, the study demonstrated that clustering on limited data could enhance estimation performance without reducing any features.

This work makes several key contributions: the development of a hybrid model for ASD detection, the application of clustering methods to small datasets, and the identification of algorithms that perform particularly well in these conditions. Moreover, the proposed model outperformed previous studies for three of the four age groups (Toddler, Children, and Adult), indicating its potential for improved detection in these groups.

However, the study does have limitations, such as the reliance on small datasets, which may have influenced the results, particularly for the Adolescent group. Future work will focus on integrating larger and more diverse datasets to validate the model's effectiveness. Collaboration with clinical psychologists will also be crucial to evaluate the clinical applicability and robustness of the model. Additionally, exploring other feature selection methods and testing the model on new datasets can help refine the approach and further enhance ASD detection performance.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Gozde Karatas Baydogmus; analysis and interpretation of results: Gozde Karatas Baydogmus, Onder Demir; draft manuscript preparation: Gozde Karatas Baydogmus, Onder Demir. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All data that support the findings of this study are included within the article.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Centers for Disease Control and Prevention. Data & statistics on autism spectrum disorder. 2023 [cited 2025 May 19]. Available from: https://www.cdc.gov/autism/data-research/index.html.
2. Magán-Maganto M, Bejarano-Martín Á, Fernández-Alvarez C, Narzisi A, García-Primo P, Kawa R, et al. Early detection and intervention of ASD: a European overview. Brain Sci. 2017;7(12):159. doi:10.3390/brainsci7120159.
3. Farooq MS, Tehseen R, Sabir M, Atal Z. Detection of autism spectrum disorder (ASD) in children and adults using machine learning. Sci Rep. 2023;13(1):9605. doi:10.1038/s41598-023-35910-1.
4. Uddin MJ, Ahamad MM, Sarker PK, Aktar S, Alotaibi N, Alyami SA, et al. An integrated statistical and clinically applicable machine learning framework for the detection of autism spectrum disorder. Computers. 2023;12(5):92. doi:10.3390/computers12050092.
5. Abdelwahab MM, Al-Karawi KA, Hasanin E, Semary H. Autism spectrum disorder prediction in children using machine learning. J Disab Res. 2024;3(1):20230064. doi:10.57197/jdr-2023-0064.
6. Abbas RT, Sultan K, Sheraz M, Chuah TC. A comparative analysis of automated machine learning tools: a use case for autism spectrum disorder detection. Information. 2024;15(10):625. doi:10.3390/info15100625.
7. Xu Y, Yu Z, Li Y, Liu Y, Li Y, Wang Y. Autism spectrum disorder diagnosis with EEG signals using time series maps of brain functional connectivity and a combined CNN–LSTM model. Comput Methods Programs Biomed. 2024;250(12):108196. doi:10.1016/j.cmpb.2024.108196.
8. Dia M, Khodabandelou G, Sabri AQM, Othmani A. Video-based continuous affect recognition of children with Autism Spectrum Disorder using deep learning. Biomed Signal Process Control. 2024;89(2):105712. doi:10.1016/j.bspc.2023.105712.
9. Rubio-Martín S, García-Ordás MT, Bayón-Gutiérrez M, Prieto-Fernández N, Benítez-Andrades JA. Enhancing ASD detection accuracy: a combined approach of machine learning and deep learning models with natural language processing. Health Inf Sci Syst. 2024;12(1):20. doi:10.1007/s13755-024-00281-y.
10. Pandey R, Maurya N, Maurya P, Saxena P. Predictive approach for Autism detection using computer vision and deep learning. In: 2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon). Piscataway, NJ, USA: IEEE; 2024. p. 1–6. [cited 2025 May 19]. Available from: https://ieeexplore.ieee.org/document/10575142.
11. Loganathan S, Geetha C, Nazaren AR, Fernandez MHF. Autism spectrum disorder detection and classification using chaotic optimization based Bi-GRU network: an weighted average ensemble model. Expert Syst Appl. 2023;230(1):120613. doi:10.1016/j.eswa.2023.120613.

12. Thabtah F. Autism screening data for toddlers. 2018 [cited 2025 May 19]. Available from: https://www.kaggle.com/datasets/fabdelja/autism-screening-for-toddlers.
13. Thabtah F. Autistic spectrum disorder screening data for children. UCI Machine Learning Repository; 2017. doi:10.24432/C5659W.
14. Thabtah F. Autism screening adult. UCI Machine Learning Repository; 2017. doi:10.24432/C5F019.
15. Thabtah F. Autistic spectrum disorder screening data for adolescent. UCI Machine Learning Repository; 2017. doi:10.24432/C5V89T.
16. Karatas G, Demir O, Sahingoz OK. Increasing the performance of machine learning-based IDSs on an imbalanced and up-to-date dataset. IEEE Access. 2020;8:32150–62. doi:10.1109/access.2020.2973219.
17. Xu D, Tian Y. A comprehensive survey of clustering algorithms. Ann Data Sci. 2015;2:165–93. doi:10.1007/s40745-015-0040-1.
18. Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, Costa LdF, et al. Clustering algorithms: a comparative approach. PLoS One. 2019;14(1):e0210236. doi:10.1371/journal.pone.0210236.
19. Hussain I, Nataliani Y, Ali M, Hussain A, Mujlid HM, Almaliki FA, et al. Weighted multiview K-means clustering with L2 regularization. Symmetry. 2024;16(12):1646. doi:10.3390/sym16121646.
20. Yang MS, Hussain I. Unsupervised multi-view K-means clustering algorithm. IEEE Access. 2023;11(6):13574–93. doi:10.1109/access.2023.3243133.
21. Das K, Behera RN. A survey on machine learning: concept, algorithms and applications. Int J Innovat Res Comput Commun Eng. 2017;5(2):1301–9.
22. Alzubi J, Nayyar A, Kumar A. Machine learning from theory to algorithms: an overview. In: Journal of Physics: Conference Series. Bangalore, India: IOP Publishing; 2018. Vol. 1142.
23. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006;63(1):3–42. doi:10.1007/s10994-006-6226-1.
24. Gibbs MN, MacKay DJ. Variational Gaussian process classifiers. IEEE Transact Neural Netw. 2000;11(6):1458–64. doi:10.1109/72.883477.
25. Tang Y. Deep learning using linear support vector machines. arXiv:1306.0239. 2013.
26. Singh A, Prakash BS, Chandrasekaran K. A comparison of linear discriminant analysis and ridge classifier on Twitter data. In: 2016 International Conference on Computing, Communication and Automation (ICCCA). Piscataway, NJ, USA: IEEE; 2016. p. 133–8. [cited 2025 May 19]. Available from: https://ieeexplore.ieee.org/document/7813704.391.
27. Amari Si. Backpropagation and stochastic gradient descent method. Neurocomputing. 1993;5(4–5):185–96. doi:10.1016/0925-2312(93)90006-o.
28. Thabtah F, Peebles D. A new machine learning model based on induction of rules for autism detection. Health Inform J. 2020;26(1):264–86. doi:10.1177/1460458218824711.
29. Akter T, Satu MS, Khan MI, Ali MH, Uddin S, Lio P, et al. Machine learning-based models for early stage detection of autism spectrum disorders. IEEE Access. 2019;7:166509–27. doi:10.1109/access.2019.2952609.
30. Raj S, Masood S. Analysis and detection of autism spectrum disorder using machine learning techniques. Procedia Comput Sci. 2020;167(12):994–1004. doi:10.1016/j.procs.2020.03.399.