Journal of Engineering and Technological Advances (JETA)
https://jeta.segi.edu.my/index.php/segi
2024 Vol. 9 No. 1

TRENDS IN AUTISM SPECTRUM DISORDER PREDICTION USING MACHINE LEARNING: A REVIEW

[1]Saat, A.E., [1]Mohd Azwan, N.A., [1]Azman, A.S., [1]Kamarudzaman, M.A.A. [1]Ismail, N.A. [1]Ridzuan, F.*

[1] Faculty of Data Science and Computing, University Malaysia Kelantan, 16100 Kota Bharu, Kelantan, Malaysia.

* Corresponding Author: fakhitah.r@umk.edu.my TEL: (609)- 7717179

Received: 6 June 2024; Accepted: 24 June 2024; Published: 30 June 2024
doi: 10.35934/segi.v9i1.103

Highlights:

- *Support Vector Machine and Logistic Regression showed promise for ASD prediction.*
- *To achieve high accuracy in ASD prediction, proper handling of missing data, normalization, and feature selection is essential.*
- *This review explores machine learning techniques for ASD prediction, including data preparation, modeling, and evaluation.*

Abstract: Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that significantly affects social, linguistic, and cognitive skills. Early diagnosis is crucial for improving long-term outcomes, yet traditional diagnostic methods are time-consuming and expensive. This review aims to explore the potential of machine learning techniques in enhancing the accuracy and efficiency of ASD prediction and diagnosis. By examining ten studies, the review evaluates the various machine learning (ML) algorithms used, pre-processing techniques employed, and datasets analysed. Key findings indicate that pre-processing techniques such as handling missing values, normalization, and feature selection are vital for improving model accuracy. Support Vector Machine and Logistic Regression consistently demonstrated high accuracy in predicting ASD across various datasets. The conclusion underscores the importance of pre-processing in developing reliable machine learning models for ASD prediction and highlights the need for future research to address challenges related to data accessibility, model interpretability, and validation across diverse populations. The responsible integration of ML technologies into clinical practice could revolutionize early diagnosis and intervention strategies for ASD.

Keywords: ASD; data mining; machine learning; autism; prediction

1. Introduction

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition that affects individuals in various ways, primarily characterized by challenges in social communication, restrictive interests, and repetitive behaviours (Hodges *et al.*, 2020). The prevalence of ASD has been increasing globally, necessitating enhanced understanding, early diagnosis, and effective intervention strategies to improve the quality of life for those affected. The Centres for Disease Control and Prevention (CDC) in the United States estimate that 1 in 68 children are affected with autism. This implies that about 9000 Malaysian babies are born with autism each year (NASOM, 2022).

Traditional diagnostic methods, although thorough, often involve lengthy evaluations by specialized professionals, which can delay intervention and support. The diagnosis and treatment of ASD is a complicated developmental disease that impacts behaviour and communication (Okoye *et al.*, 2023), can be substantially aided by data-driven insights. The integration of machine learning techniques into the medical diagnostics field presents a promising avenue for addressing these challenges. Data mining is a practice of obtaining valuable information from massive data (Shu & Ye, 2023). It involves applying a variety of methods from database systems, machine learning, and statistics to find correlations, patterns, and trends in the data. Machine learning, a subset of artificial intelligence, involves algorithms that enable computers to learn from and make predictions based on data (Sarker, 2021). By leveraging machine learning, these algorithms can potentially streamline the diagnostic process, offering quicker, more consistent, and possibly earlier detection of the disorder.

Data mining can help healthcare professionals identify the most important factors for diagnosing a disease (Saleh & Rabie, 2023). This can lead to more accurate and efficient diagnoses, especially for complex conditions. Besides, it can help to identify and eliminate outliers in medical data, which can improve the accuracy of diagnostic models (Mehbodniya *et al.*, 2022; Saleh & Rabie, 2023). This can save time and resources by preventing the use of bad training data.

Data mining offers healthcare professionals a powerful set of tools to improve diagnoses, eliminate inefficiencies, and ultimately deliver better patient care. Therefore, this paper aims to explore the current state of research on using machine learning for the prediction and diagnosis of ASD. This review includes the studies that have applied various machine learning algorithms to ASD datasets, highlighting their methodologies and findings.

2. Machine Learning Trends in ASD

ASD remains a complex and time-consuming process, often leading to delays in critical interventions. However, the field of machine learning is rapidly transforming the landscape of ASD research. A recent publication trend analysis on Scopus using the keywords "Autism Spectrum Disorder" or "ASD" and "machine learning" or "ML" returned 2,526 results. This clearly shows that there is great potential for ML to revolutionize the diagnosis of ASD. **Figure 1** shows the number of documents by year related to machine learning and ASD. From the figure, it is clearly seen that the number of documents related to both machine learning and ASD has been increasing over time.

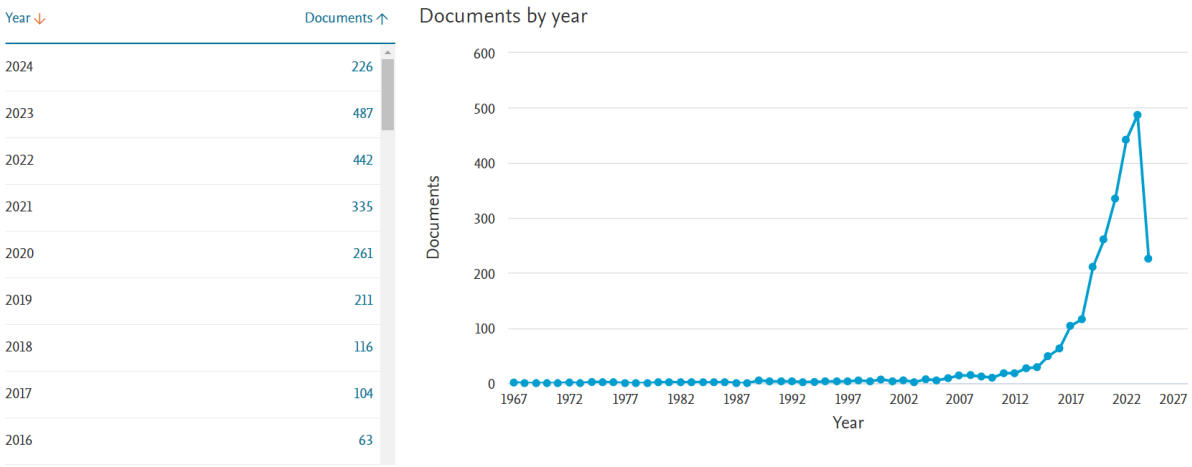

Figure 1. Number of documents by year (Scopus, 2024)

Table 1 shows the list of previous research related to ASD. Based on the review, most of the research are directed to the utilization of machine learning techniques to enhance the diagnosis and prediction of ASD. Early detection and intervention are crucial in managing ASD effectively, yet conventional diagnostic methods are often time-consuming and costly. Therefore, researchers are turning to ML algorithms to streamline the diagnostic process and improve accuracy.

Table 1. List of previous research from the literature related to ASD

No.	Title	Author	Objective	Result

1.	Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques	(Raj & Masood, 2020)	Exploring machine learning techniques for the prediction and analysis of ASD problems in children, adolescents, and adults.	Convolutional Neural Networks (CNN) outperformed other models with accuracy of 99.53% for the ASD screening in adult's dataset, 98.30% for the ASD screening in children dataset and 96.88% for the ASD screening in adolescent's dataset.
2.	Detection of Autism Spectrum Disorder (ASD) in Children and Adults Using Machine Learning	(Farooq *et al.*, 2023)	This research aimed to develop a novel ASD detection system using for early diagnosis, particularly focusing on children	The proposed model achieved high accuracy with 98% accuracy in detecting ASD in children, and 81% accuracy in detecting ASD in adults.
3.	Predicting Autism Spectrum Disorder Using Machine Learning Classifiers	(Chowdhury & Iraj, 2020)	This paper aims to measure and identify ASD	Support Vector Machine (SVM) classifier with the Gaussian Radial Kernel achieves 95% accuracy.
4.	An Evaluation of Machine Learning Approaches for Early Diagnosis of Autism Spectrum Disorder	(Rasul *et al.*, 2024)	Evaluate different machine learning approaches for the early diagnosis of ASD using classification and clustering	The SVM and Logistic Regression (LR) models achieve the highest accuracy of 100% for the children dataset. The LR model produces the highest accuracy of 97.14% for the adult dataset.

5.	The Classification of Autism Spectrum Disorder by Machine Learning Methods on Multiple Datasets for Four Age Groups	(Khudhur & Khudhur, 2023)	Develop a machine learning model to predict ASD across different age groups (toddler, child, adolescent, and adult)	Decision Tree (DT), LR and Random Forest (RF) are the most effective model that achieved higher accuracy.
6.	Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques	(Vakadkar *et al.*, 2021)	To develop machine learning model to help detect children who may be at risk for ASD	LR models provide higher accuracy
7.	Effective Autism Spectrum Disorder Prediction to Improve the Clinical Traits Using Machine Learning Techniques	(Surendiran *et al.*, 2022)	To develop ASD prediction framework that support behavioural aspect-based analysis in children and adolescents	ANN produces the highest accuracy of 97.68%, outperforming other methods.
8.	Machine Learning-Based Models for Early-Stage Detection of Autism Spectrum Disorders	(Akter *et al.*, 2019)	Apply machine learning for predicting ASD across different developmental stages (toddlers, children, adolescents and adults) to identify significant ASD risk factors.	SVM showed the best performance for the toddler dataset, while Adaboost gave the best results for the children and adult datasets and Glmboost provided the best results for the adolescent dataset.
9.	Prior Prediction and Management of Autism in Child Through Behavioural	(Rahman & Mamun, 2022)	To analyse machine learning algorithm to determine set of conditions predictive of ASD	Weighted Voting Classifier (WVC) achieved the highest accuracy (97%) on the dataset

	Analysis Using Machine Learning Approach			
10.	Autism Spectrum Disorder Prediction in Children Using Machine Learning	(Abdelwahab *et al.*, 2024)	Use machine learning technique to improve ASD diagnosis across children, adolescents, and adults.	LR demonstrated the highest accuracy for the selected dataset

3. Dataset Characteristics

The datasets used by other researchers come from publicly available repositories, primarily the UCI Machine Learning Repository and Kaggle. The main types of datasets utilized include ASD Screening Data for Adults, Children, Adolescents, and Toddlers. Each dataset consists of categorical, continuous, and binary attributes.

The attributes common across these datasets include patient age, sex, nationality, history of jaundice, family history of pervasive developmental disorders, experimental fulfilment, and screening-related information. Additionally, screening application usage and test type were recorded, along with screening scores derived from responses to ten questions, providing a comprehensive basis for ASD analysis and prediction across different age groups. This comprehensive dataset allows for a thorough analysis and prediction modelling of ASD across different age groups.

Several studies have utilized these datasets, each focusing on different age groups. Raj and Masood (2020) and Abdelwahab *et al*. (2024) focused on adults, children, and adolescents. Meanwhile, Khudur and Khudur (2023), Akter *et. al*. (2019), and Farooq *et al*. (2023) employed ASD screening data for toddler, children, adolescence and adult. On the other hand, Chowdury and Iraj (2021) focused on adults, while Vakadkar *et al.* (2021) concentrated on toddlers. Rasul *et al*. (2024) used child and adult datasets from the same sources and added a combined dataset for further analysis. Overall, it is evident that most studies utilized the same datasets, but the results and algorithms used varied significantly. These differences can be attributed to the distinct pre-processing techniques employed in each study.

4. Pre-processing Techniques

Pre-processing techniques are crucial in preparing the dataset for effective machine learning model training, ensuring that the data is clean and appropriately formatted. Common pre-processing steps include handling missing values, normalizing or standardizing the data, and encoding categorical variables. The choice of pre-processing method can significantly influence the performance and accuracy of the machine learning models, leading to different outcomes across studies using the same dataset (Alshdaifat *et al.*, 2021).

Various pre-processing methods have been employed by different researchers to prepare their datasets for machine learning analysis, significantly impacting the results and algorithm performance. Rasul *et al.* (2024) utilized missing value handling, one-hot encoding, feature scaling, and also performed feature selection to refine the dataset. Similarly, Khudur and Khudur (2023) and Abdelwahab *et al.* (2024) incorporated missing value handling, encoding, normalization, and feature selection in their study. Raj and Masood (2020) emphasized transforming raw data into a meaningful format by handling missing values through imputation, detecting outliers, data discretization, and data reduction. Farooq *et al.* (2023) focused on data cleaning, noise removal, and normalization to ensure a high-quality dataset. Chowdury and Iraj (2021) primarily addressed missing values, while Vakadkar *et al.* (2021) included handling missing values, outliers, noise removal, encoding, normalization, and feature selection. Akter *et. al* (2019) dealt with noisy and missing records by replacing them with mean values, and also applied encoding techniques. Each of these pre-processing strategies aimed to enhance the dataset's quality, ultimately contributing to more accurate and reliable machine learning models for ASD prediction. **Table 2** summarizes the pre-processing technique used by other researchers.

Table 2. List of pre-processing technique used

Authors	Missing Value	Encoding	Outliers	Feature Scaling	Feature Selection	Data Reduction	Data Discretization	Noise Removal	Normalization	Data Cleaning
(Raj & Masood, 2020)	√		√			√				
(Farooq *et al.*, 2023)								√	√	√
(Chowdhury & Iraj, 2020)	√									
(Rasul *et al.*, 2024)	√	√		√	√					

(Khudhur & Khudhur, 2023)	√	√			√					√	
(Vakadkar *et al.*, 2021)	√	√	√		√				√		
(Akter *et al.*, 2019)	√								√		
(Abdelwahab *et al.*, 2024)	√	√			√					√	

5. Machine Learning Algorithm

After data pre-processing, the refined data is then fed into various machine learning algorithms for prediction. This step is critical as it enables the development of models that can accurately identify and predict ASD, ultimately aiding in early detection and intervention. **Table 3** shows the list of algorithms used by other researchers.

Table 3. List of algorithm used

Authors	**NB**	**SVM**	**LR**	**KNN**	**ANN**	**CNN**	**DT**	**RF**	**AdaBoost**	**GImBoost**	**WVC**
(Raj & Masood, 2020)	√	√	√	√	√	√					
(Farooq *et al.*, 2023)		√	√								
(Chowdhury & Iraj, 2020)		√									
(Rasul *et al.*, 2024)		√	√		√						
(Khudhur & Khudhur, 2023)	√	√	√	√			√	√			
(Vakadkar *et al.*, 2021)	√	√	√	√				√			
(Surendiran *et al.*, 2022)		√		√	√		√	√			
(Akter *et al.*, 2019)		√							√	√	

(Rahman & Mamun, 2022)	√	√	√	√	√		√	√	√		√
(Abdelwahab *et al.*, 2024)	√	√	√	√			√	√			

NB = Naïve Bayes; SVM = Support Vector Machine; LR = Logistic Regression; KNN = K-nearest Neigbour, ANN = Artificial Neural Network; CNN = Convolutional Neural Network; DT = Decision Tree; RF = Random Forest; WVC = Weighted Voting Classifier

Based on the data presented in **Table 2**, SVM emerges as the most frequently utilized algorithm, appearing in ten research studies. Following closely behind is LR, employed seven times. KNN in the third position, with six research studies. SVM are favoured for machine learning in ASD since it performs effectively in high-dimensional spaces (Pisner & Schnyer, 2020), which is essential since ASD datasets often encompass numerous features related to behavioural, physiological, and demographic factors. Additionally, SVM is robust to overfitting (Boateng *et al.*, 2020), a crucial feature when dealing with the noisy and intricate datasets often found in this field. These characteristics collectively make SVM a powerful and versatile tool for ASD prediction, capable of handling the complexities inherent in medical and behavioural data.

Similarly, LR and KNN are commonly employed in machine learning for ASD prediction due to their distinct advantages and suitability for different types of datasets. LR is favoured for its simplicity and interpretability (Dumitrescu *et al.*, 2022), making it particularly useful in clinical settings where understanding the model's decisions is crucial. The coefficients in LR clearly indicate the weight of each feature, helping clinicians and researchers identify which factors are most predictive of ASD. Additionally, LR provides probabilistic outputs, offering not just the prediction but also the probability of belonging to a particular class (Cemiloglu *et al.*, 2023). This probabilistic nature aids in assessing the degree of risk and making informed clinical decisions. On the other hand, KNN classifies new instances based on the majority class of their nearest neighbours, making it easy to implement and understand (Boateng *et al.*, 2020). Its non-parametric nature means that KNN makes no assumptions about the underlying data distribution, allowing it to model complex relationships in the data effectively. This flexibility is particularly useful for ASD datasets, which can be diverse and non-linear. KNN is also adaptable to various types of data, including numerical and categorical features, and can be

tailored to different distance metrics based on the data's characteristics (Wazirali, 2020). **Figure 2** shows the summary of algorithm used by other researchers.

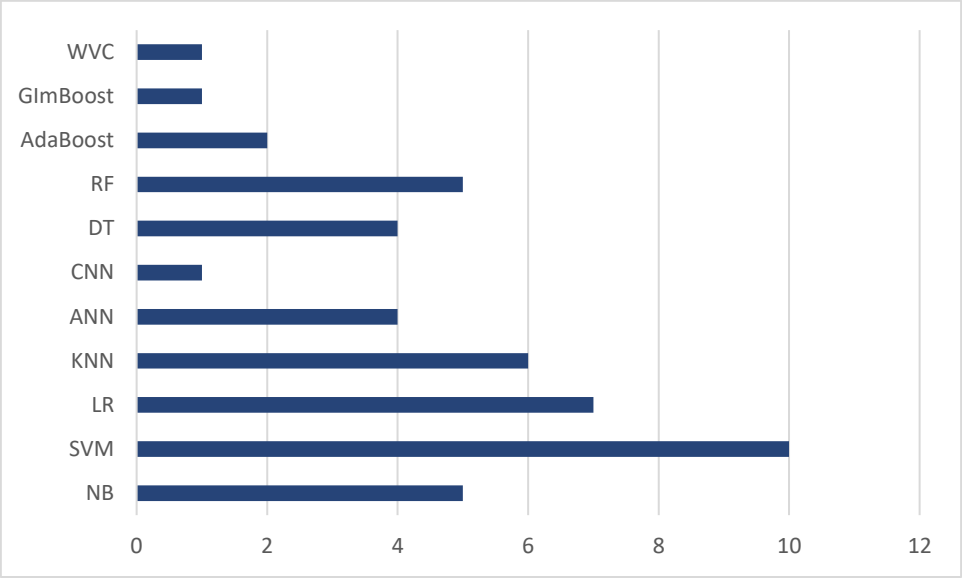

Figure 2. List of algorithms used by other researchers

Based on the study, it is evident that machine learning algorithms show promise in accurately predicting and analysing ASD. For example, LR consistently demonstrates high accuracy across different datasets and age groups. This algorithm's effectiveness suggests its potential utility in clinical settings for early ASD detection, which could significantly impact intervention strategies and long-term outcomes for individuals with ASD

Moreover, other algorithms such as SVM, RF, and Adaboost have also shown promising results in certain studies. These algorithms offer different strengths and may be more suitable for specific datasets or age groups. For instance, SVM has shown excellent performance in predicting ASD in toddlers, while Adaboost outperforms other classifiers for children and adults.

6. Conclusion

In conclusion, this review highlights the immense potential of machine learning for improving ASD prediction and diagnosis accuracy and efficiency. By leveraging advanced algorithms and robust data pre-processing techniques, researchers can develop increasingly sophisticated

models that facilitate early detection and intervention, ultimately leading to better long-term outcomes for individuals with ASD. Data pre-processing plays a critical role, as it ensures the quality and usability of data for machine learning models. However, addressing challenges like data accessibility, model interpretability, and ensuring generalizability across diverse populations remains critical. Future research should explore advanced pre-processing methods, evaluate models on broader demographics, and prioritize ethical considerations surrounding data privacy in ASD diagnosis.

Acknowledgement

The authors would like to thank Universiti Malaysia Kelantan for supporting the research of this study.

Credit Author Statement

Writing—original draft preparation, Saat, A.E., Mohd Azwan, N.A., Azman, A.S. and Kamarudzaman, M.A.A.; writing—review and editing, Ridzuan, F. and Ismail, N.A.; supervision, Ridzuan, F.

Conflicts of Interest

The authors declare no conflict of interest.

References

Abdelwahab, M.M., Al-Karawi, K.A., Hasanin, E.M., & Semary, H.E. (2024). Autism spectrum disorder prediction in children using machine learning. *Journal of Disability Research*, 3(1), 20230064.

Akter, T., Satu, M.S., Khan, M.I., Ali, M.H., Uddin, S., Lio, P., Quinn, J.M.W., & Moni, M.A. (2019). Machine learning-based models for early stage detection of autism spectrum disorders. *IEEE Access, 7*, 166509–166527.

Alshdaifat, E. A., Alshdaifat, D. A., Alsarhan, A., Hussein, F., & El-Salhi, S. M. D. F. S. (2021). The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance. *Data*, 6(2), 11.

Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: A review. *Journal of Data Analysis and Information Processing*, 8(4), 341-357.

Cemiloglu, A., Zhu, L., Mohammednour, A. B., Azarafza, M., & Nanehkaran, Y. A. (2023). Landslide

susceptibility assessment for Maragheh County, Iran, using the logistic regression algorithm. *Land*, 12(7), 1397

Chowdhury, K., & Iraj, M. A. (2020). Predicting autism spectrum disorder using machine learning classifiers. In 2020 *International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*, 324-327. IEEE.

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, 297(3), 1178-1192.

Farooq, M. S., Tehseen, R., Sabir, M., & Atal, Z. (2023). Detection of autism spectrum disorder (ASD) in children and adults using machine learning. *Scientific Reports*, 13(1), 9605.

Hodges, H., Fealko, C., & Soares, N. (2020). Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation. *Translational Pediatrics*, 9(Suppl 1), S55.

Khudhur, D. D., & Khudhur, S. D. (2023). The classification of autism spectrum disorder by machine learning methods on multiple datasets for four age groups. *Measurement: Sensors*, 27, 100774.

Mehbodniya, A., Khan, I. R., Chakraborty, S., Karthik, M., Mehta, K., Ali, L., & Nuagah, S. J. (2022). [Retracted] Data Mining in Employee Healthcare Detection Using Intelligence Techniques for Industry Development. *Journal of Healthcare Engineering*, 2022(1), 6462657.

NASOM. (2022). *Autism*. Retrieved from https://www.nasom.org.my/autism/

Okoye, C., Obialo-Ibeawuchi, C. M., Obajeun, O. A., Sarwar, S., Tawfik, C., Waleed, M. S., Wasim, A. U., Mohamoud, I., Afolayan, A. Y., & Mbaezue, R. N. (2023). Early diagnosis of autism spectrum disorder: A review and analysis of the risks and benefits. *Cureus*, 15(8).

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. *In Machine learning*, 101-121. Academic Press.

Rahman, M. M., & Mamun, S. A. (2022). Prior prediction and management of autism in child through behavioral analysis using machine learning approach. *In Rhythms in Healthcare*, 63-77. Singapore: Springer Nature Singapore.

Raj, S., & Masood, S. (2020). Analysis and detection of autism spectrum disorder using machine learning techniques. *Procedia Computer Science*, 167, 994-1004.

Rasul, R. A., Saha, P., Bala, D., Karim, S. R. U., Abdullah, M. I., & Saha, B. (2024). An evaluation of machine learning approaches for early diagnosis of autism spectrum disorder. *Healthcare Analytics*, 5, 100293.

Saleh, A. I., & Rabie, A. H. (2023). A new autism spectrum disorder discovery (ASDD) strategy using data mining techniques based on blood tests. *Biomedical Signal Processing and Control*, 81,

104419.

Sarker, I. H. (2021). Machine learning: algorithms, real-world applications and research directions. *SN Comput Sci*, 2:160.

Scopus. (2024). *Analyze Search Results*. https://www.scopus.com/term/analyzer.uri?sort=plf-f&src=s&sid=c69e7c88b07554f575cff99a0368026f&sot=a&sdt=a&sl=86&s=TITLE-ABS-KEY%28%28%22Autism+Spectrum+Disorder%22+OR+%22ASD%22%29+AND++%28%22machine+learning%22+OR+%22ML%22%29%29&origin=resultslist&count=10&analyzeResults=Analyze+results

Shu, X., & Ye, Y. (2023). Knowledge Discovery: Methods from data mining and machine learning. *Social Science Research*, 110, 102817.

Surendiran, R., Thangamani, M., Narmatha, C., & Iswarya, M. (2022). Effective autism spectrum disorder prediction to improve the clinical traits using machine learning techniques. *International Journal of Engineering Trends and Technology (IJETT)*, *70*(4), 343–359.

Vakadkar, K., Purkayastha, D., & Krishnan, D. (2021). Detection of autism spectrum disorder in children using machine learning techniques. *SN Computer Science*, 2, 1-9.

Wazirali, R. (2020). An improved intrusion detection system based on KNN hyperparameter tuning and cross-validation. *Arabian Journal for Science and Engineering*, 45(12), 10859-10873.