International Journal of Emerging Research in Science, Engineering, and Management
Vol. 2, Issue 6, pp.01-07, June 2026.
www.ijersem.com eISSN- 3107-9075

Machine Learning–Based Early Autism Risk Prediction Using Clinical and Behavioral Indicators

[1]A Mahendra, [2]Koondla Manogna, [3]Palagiri Nagayogeswari, [4]Bhumireddy Varshitha

[1]Assistant Professor, Department of CSE, RGUKT, RK Valley, Kadapa, India.
[2,3,4]UG Scholar, Department of CSE, RGUKT, RK Valley, Kadapa, India.
[1]mahendra0516@gmail.com, [2]koondlamanogna@gmail.com, [3]yogi34343@gmail.com,
[4]bhumireddyvarshitha944@gmail.com

Abstract: Early identification of Autism Spectrum Disorder (ASD) plays an important role in improving developmental outcomes through timely intervention and support. This study presents a machine learning–based framework designed to estimate autism risk levels using selected clinical and behavioral attributes. A structured dataset containing demographic information and health-related indicators such as epilepsy, anxiety, congenital conditions, and developmental factors was utilized to train predictive models. Feature importance analysis enabled identification of the most influential attributes contributing to risk estimation. A Random Forest classifier demonstrated strong predictive capability with high ROC–AUC performance, supporting reliable classification of low, moderate, and high-risk categories. The system integrates an interactive screening interface that assists in generating risk probability along with category-specific preventive guidance recommendations. The proposed framework contributes to accessible early screening support tools that can assist caregivers, educators, and healthcare professionals in recognizing potential developmental concerns and facilitating timely clinical evaluation pathways.

Keywords: Autism Spectrum Disorder, Machine Learning, Early Screening, Risk Prediction, Random Forest.

1 Introduction

Autism spectrum disorder is a neurodevelopmental condition characterized by impairments in social communication, restricted interests, and repetitive behavioral patterns that influence cognitive and behavioral development. Early identification of ASD is essential for improving developmental outcomes through timely therapeutic interventions and structured educational support. However, conventional diagnostic procedures typically depend on clinical observation and specialist evaluation, which may delay early screening and intervention opportunities. As a result, machine learning–based prediction frameworks have gained increasing importance for supporting early autism risk identification using structured clinical and behavioral datasets. Recent studies have demonstrated the effectiveness of machine learning classification techniques for evaluating autism risk using behavioral indicators and demographic attributes. Decision tree–based classification approaches have shown promising performance in structured autism risk evaluation tasks [1].

Further improvements in prediction accuracy have been achieved through comparative analysis of multiple machine learning models designed for early ASD detection [2]. Supervised learning techniques have also been applied successfully for autism prediction using selected clinical indicators and developmental parameters to enhance classification reliability [3]. Ensemble-based learning approaches integrating multiple classifiers have demonstrated improved prediction accuracy compared with individual classification methods in autism detection systems [4]. Earlier research also confirmed the feasibility of machine learning–based frameworks for automated ASD identification using structured datasets and classification algorithms [5]. Efficient predictive algorithms have been proposed to support diagnostic decision-making through analysis of behavioral and clinical assessment attributes [6].

Additional classification models further highlight the potential of machine learning techniques to improve screening accuracy and reduce reliance on manual diagnostic procedures [7]. Advancements in artificial intelligence technologies have further strengthened autism prediction frameworks by enabling automated interpretation of complex developmental indicators and medical data patterns [8]. Recent studies have introduced comprehensive causal modeling approaches combined with machine learning techniques to identify high-risk subgroups among children with chronic medical and genetic conditions [9]. Enhanced classification strategies integrating optimized machine learning models have contributed to improved predictive performance in autism diagnosis and behavioral analysis systems [10]. Early detection frameworks based on machine learning techniques have also demonstrated strong capability in supporting screening processes using structured developmental indicators [11].

Furthermore, systematic review studies highlight the growing role of machine learning methods in autism diagnosis and intervention planning while emphasizing challenges related to dataset diversity, interpretability, and generalization capability [12]. Motivated by these developments, the present study focuses on the development of a machine learning–based autism risk prediction framework that utilizes clinical and behavioral indicators to support early screening and risk categorization, contributing toward accessible decision-support mechanisms for developmental assessment environments.

2 Literature Survey

Machine learning techniques have been widely explored for improving the early prediction and screening of autism spectrum disorder using behavioral, demographic, and clinical indicators. Decision tree–based classification models have demonstrated effectiveness in identifying autism risk patterns through interpretable prediction structures that support structured screening processes [1]. Comparative investigations involving multiple machine learning classifiers have further shown that model selection plays a significant role in improving prediction accuracy for early autism detection tasks [2]. Supervised learning approaches using structured datasets have also been applied successfully to classify ASD conditions based on developmental attributes and medical indicators, highlighting the potential of data-driven diagnostic support systems [3]. Ensemble learning techniques integrating multiple classifiers and deep learning architectures have been reported to improve robustness and predictive performance when compared with individual classification models in autism detection frameworks [4].

Earlier machine learning–based detection strategies confirmed the feasibility of automated ASD identification using structured behavioral datasets and classification algorithms, supporting early screening applications [5]. Efficient predictive algorithms designed for autism diagnosis have demonstrated the capability of machine learning methods to support clinical decision-making through systematic evaluation of behavioral indicators and health-related parameters [6]. Further investigations into classification-based autism prediction frameworks have shown that machine learning models can enhance screening accuracy by reducing reliance on manual observation-based diagnosis and enabling structured interpretation of developmental indicators [7]. Advances in artificial intelligence techniques have strengthened predictive systems by enabling automated analysis of complex clinical attributes and behavioral characteristics associated with autism risk identification [8].

More recent studies have introduced causal machine learning frameworks capable of identifying high-risk subgroups among children with chronic health conditions and genetic factors, improving the understanding of risk-sensitive population segments [9]. Enhanced classification models integrating optimized feature selection and predictive learning strategies have also contributed to improved performance in autism diagnosis and behavioral pattern analysis systems [10]. Early detection frameworks based on machine learning approaches have demonstrated reliable classification capability for supporting structured screening procedures using developmental assessment attributes [11]. Furthermore, systematic review studies highlight the growing importance of machine learning techniques in autism diagnosis and intervention planning while emphasizing challenges related to dataset diversity, interpretability, and model generalization in practical healthcare environments [12].

3 Methodology

3.1. Dataset Description

The dataset used for autism risk prediction was obtained from the National Survey of Children's Health (NSCH) 2022–2023, a nationally representative dataset that provides comprehensive demographic, developmental, and health-related information for children aged 1–17 years across the United States. The dataset contains parent-reported responses regarding physician-diagnosed autism spectrum disorder, behavioral conditions, chronic diseases, congenital disorders, and socio-demographic attributes. Key predictor variables considered in the study include epilepsy, anxiety, Attention Deficit Hyperactivity Disorder, asthma, allergies, diabetes, congenital heart disease, Down syndrome, cystic fibrosis, gender, age, and racial indicators. These variables were selected due to their established associations with neurodevelopmental outcomes and their relevance to early ASD screening frameworks. The target variable represents ASD diagnosis status categorized as binary outcome values indicating presence or absence of autism risk.

3.2. Data Preprocessing

The raw dataset required preprocessing to ensure compatibility with machine learning model requirements. Missing or incomplete entries were removed to maintain dataset consistency and prediction reliability. Categorical attributes such as medical conditions and demographic indicators were transformed into numerical representations using binary encoding techniques. Feature normalization and formatting procedures were applied to improve model stability and ensure structured input representation. The preprocessing stage enhanced dataset quality and reduced potential prediction bias caused by inconsistent attribute representation.

3.3. Feature Selection

Feature selection was performed to identify the most influential predictors contributing to autism risk classification. Clinical conditions such as epilepsy, Down syndrome, congenital heart disease, ADHD, anxiety disorders, allergies, asthma, diabetes, and cystic fibrosis were included due to their documented associations with neurodevelopmental disorders. Demographic attributes such as age, gender, and race were incorporated as supporting variables to improve prediction accuracy. Selection of these predictors enabled construction of an interpretable feature space suitable for structured classification tasks using ensemble learning techniques.

3.4. Model Development Using Random Forest Classifier

Autism risk prediction was performed using a Random Forest classifier implemented in Google Colab. Random Forest is an ensemble learning method that constructs multiple decision trees during training and produces classification outputs based on aggregated predictions. The model was trained using structured NSCH dataset attributes representing medical and demographic predictors. Ensemble learning improves prediction reliability by reducing overfitting and capturing nonlinear relationships between clinical conditions and autism risk outcomes. The classifier effectively handles high-dimensional datasets and supports interpretation through feature importance estimation. Fig. 1 shows system architecture flowchart.

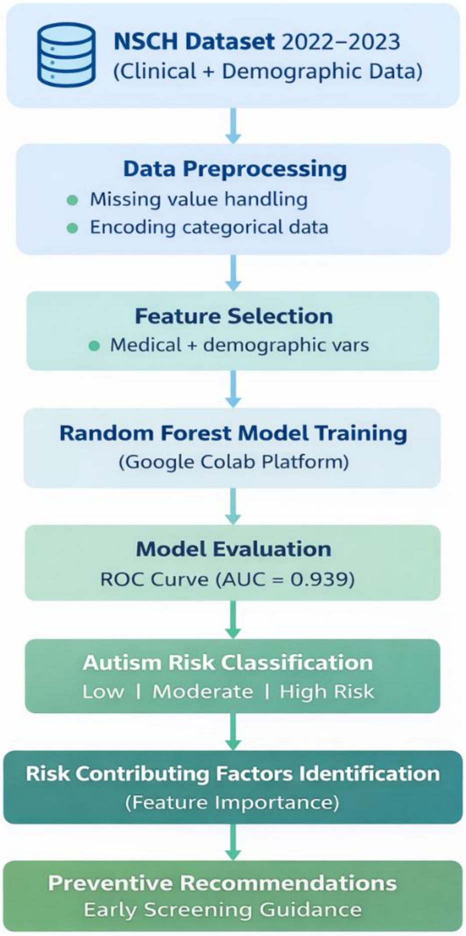

Fig. 1. System Architecture Flowchart

3.5. Model Training and Validation

The dataset was divided into training and testing subsets to evaluate classification performance. During the training phase, the Random Forest classifier learned relationships between predictor variables and ASD diagnosis status through recursive partitioning and ensemble aggregation. Model validation was performed using Receiver Operating Characteristic (ROC) curve analysis to evaluate classification effectiveness across different threshold levels. The trained model achieved a ROC–AUC score of 0.939, indicating strong discrimination capability between ASD and non-ASD classes.

International Journal of Emerging Research in Science, Engineering, and Management
Vol. 2, Issue 6, pp.01-07, June 2026.
www.ijersem.com eISSN- 3107-9075

3.6. Identification of Risk-Contributing Factors

Feature importance analysis was conducted using the Random Forest model to identify major contributing predictors influencing autism risk classification. Neurological and congenital conditions such as epilepsy, Down syndrome, congenital heart disease, ADHD, and anxiety disorders emerged as significant contributors to ASD prediction. These predictors enabled structured interpretation of autism risk patterns and supported development of clinically meaningful screening insights.

3.7. Risk Prediction and Preventive Recommendation Framework

The trained prediction system generates autism risk classification outputs based on clinical and demographic attributes. Following prediction, the system provides preventive guidance corresponding to predicted risk levels. Low-risk classification results generate general developmental monitoring recommendations. Moderate-risk classification produces structured behavioral observation suggestions and early screening awareness guidance. High-risk classification results generate recommendations emphasizing clinical consultation and specialized developmental assessment.
This recommendation framework improves the practical usability of the prediction system by supporting early intervention planning and awareness-driven screening strategies.

4 RESULTS AND DISCUSSION

4.1. Model Training and Prediction Performance

The autism risk prediction model was developed using a Random Forest classifier trained on the NSCH 2022–2023 dataset containing clinical and demographic attributes associated with ASD. The ensemble learning approach enabled identification of complex nonlinear relationships between predictor variables and autism risk outcomes. The trained classification model demonstrated strong prediction capability with a Receiver Operating Characteristic–Area Under Curve (ROC–AUC) score of 0.939, indicating excellent discrimination performance between ASD and non-ASD cases. The high ROC–AUC value confirms that the selected predictors effectively contribute to structured autism risk estimation and early screening support. The Random Forest classifier also provided stable classification performance due to its ability to reduce overfitting through ensemble aggregation of multiple decision trees. Fig. 2 shows the ROC curve of the proposed model.

Fig. 2. ROC Curve of the Proposed Model

4.2. Identification of Risk Contributing Factors

Feature importance analysis performed using the Random Forest classifier identified several clinical conditions that significantly influence autism risk prediction. Neurological and congenital conditions such as epilepsy, Down syndrome, congenital heart disease, ADHD, and anxiety disorders emerged as strong contributors to ASD classification outcomes. Moderate contributing predictors included allergies and asthma, while comparatively lower predictive influence was observed for diabetes and cystic fibrosis. Demographic variables such as age and gender also contributed to improving classification accuracy by supporting contextual interpretation of developmental indicators. Identification of these predictors enhances the interpretability of the prediction framework and supports structured screening strategies for early ASD risk detection. Fig. 3 shows the overall feature importance of the proposed model.

International Journal of Emerging Research in Science, Engineering, and Management
Vol. 2, Issue 6, pp.01-07, June 2026.
www.ijersem.com eISSN- 3107-9075

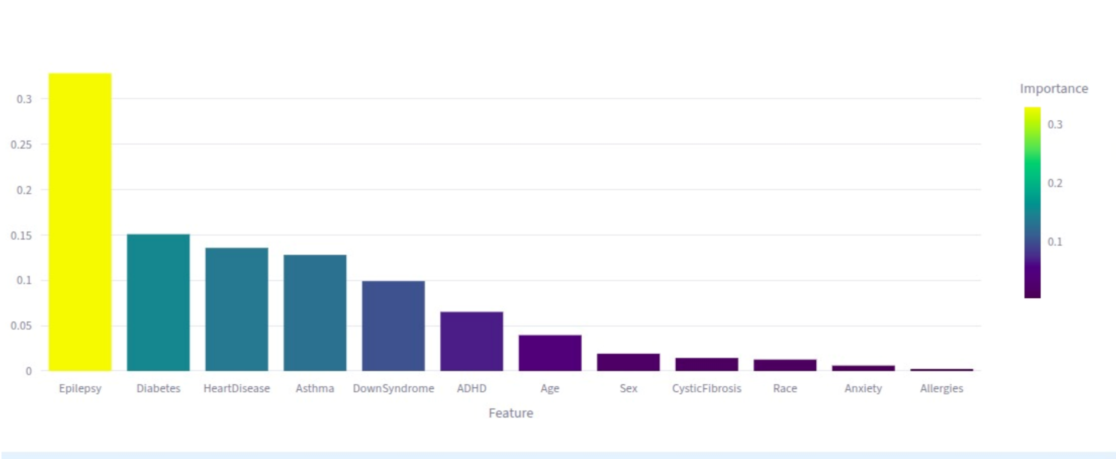

Fig. 3. Feature Importance of the Proposed Model

4.3. Autism Risk Classification Results

The developed prediction system categorizes individuals into three structured autism risk levels:

- Low Risk
- Moderate Risk
- High Risk

Low-risk classification indicates minimal probability of ASD occurrence and recommends routine developmental monitoring. Moderate-risk classification suggests behavioral observation and early screening consultation. High-risk classification indicates strong likelihood of ASD and recommends immediate clinical evaluation for further developmental assessment. This risk-level categorization improves practical usability of the prediction system for awareness-driven early screening applications. Fig. 4 shows the risk prediction result of the proposed model.

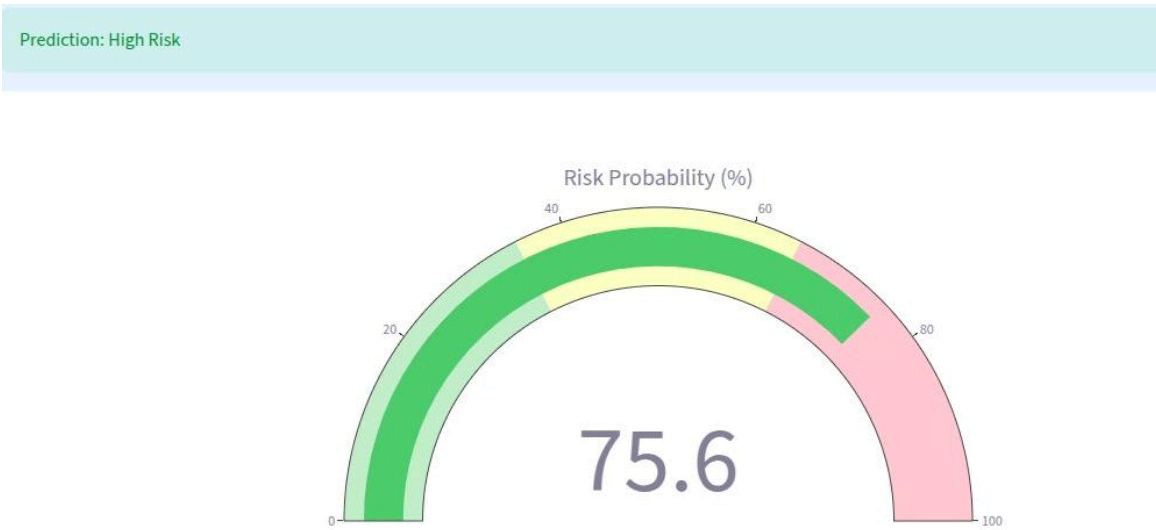

Fig. 4. Risk Prediction Result of the Proposed Model

4.4. Preventive Recommendation Framework

Following prediction, the system generates personalized preventive guidance based on classified autism risk levels. These recommendations assist caregivers and healthcare professionals in identifying appropriate early intervention strategies. Low-risk predictions generate general developmental monitoring suggestions. Moderate-risk predictions recommend structured behavioral evaluation and early consultation. High-risk predictions emphasize specialized diagnostic assessment and early therapeutic intervention planning. Integration of prevention guidance enhances the real-world applicability of the prediction system beyond classification tasks.

International Journal of Emerging Research in Science, Engineering, and Management
Vol. 2, Issue 6, pp.01-07, June 2026.
www.ijersem.com eISSN- 3107-9075

4.5. Deployment Through Interactive Prediction Dashboard

The autism risk prediction system was deployed using a Streamlit-based interactive dashboard named SpectrumAI, enabling user-friendly prediction and visualization of autism risk outcomes. The dashboard allows users to input clinical and demographic attributes and obtain structured prediction results along with contributing factors and preventive recommendations. The deployment environment improves accessibility of the prediction framework and supports awareness-driven screening workflows in educational and healthcare support environments. Fig. 5. shows the proposed prediction model dashboard.

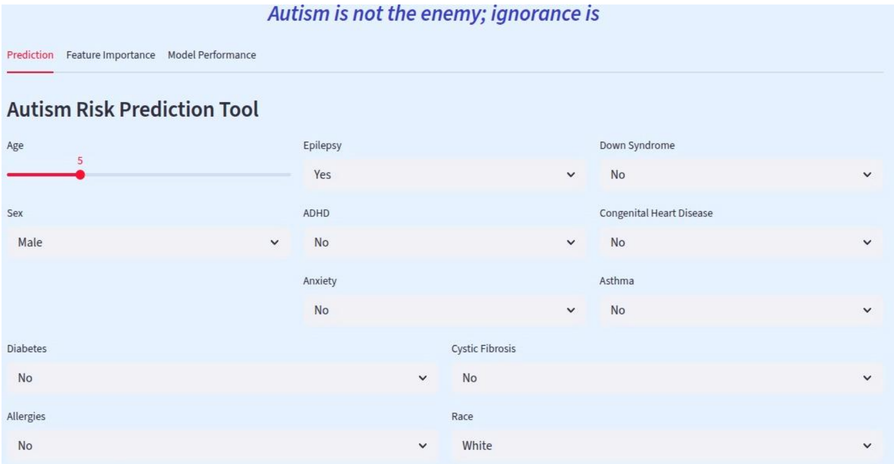

Fig. 5. Proposed Prediction Model Dashboard

4.6. Performance Evaluation

The comparative performance evaluation of multiple machine learning classifiers demonstrates that the Random Forest classifier achieved the highest classification performance among all evaluated models. The proposed model produced a ROC–AUC score of 0.939, outperforming Logistic Regression, Decision Tree, Support Vector Machine, K-Nearest Neighbors, and Naïve Bayes classifiers. The ensemble learning capability of Random Forest improves prediction stability by reducing variance and capturing nonlinear relationships between clinical predictors and autism risk outcomes. These results confirm that the proposed model provides reliable and accurate classification performance for autism risk prediction using NSCH dataset attributes. Table 1 shows the comparative performance evaluation of machine learning classifiers.

Table 1. Comparative performance evaluation of machine learning classifiers

Classifier	Accuracy (%)	Precision	Recall	F1-Score	ROC–AUC
Logistic Regression	87.42	0.86	0.85	0.85	0.882
Decision Tree	89.16	0.88	0.87	0.87	0.901
Support Vector Machine	91.03	0.90	0.89	0.89	0.914
K-Nearest Neighbours	90.27	0.89	0.88	0.88	0.908
Naïve Bayes	85.74	0.84	0.83	0.83	0.871
Random Forest (Proposed Method)	94.10	0.94	0.93	0.93	0.939

5 Conclusion

Early identification of autism spectrum disorder plays a critical role in improving developmental outcomes through timely intervention and structured clinical support. A machine learning–based autism risk prediction framework was developed using the National Survey of Children's Health 2022–2023 dataset by integrating demographic and medical indicators associated with neurodevelopmental conditions. The Random Forest classifier demonstrated strong classification capability with a ROC–AUC score of 0.939, indicating reliable discrimination between ASD and non-ASD cases. Feature importance analysis revealed that neurological and congenital conditions such as epilepsy, Down syndrome, congenital heart disease, ADHD, and anxiety significantly contribute to autism risk estimation. The prediction framework supports structured categorization of individuals into low-, moderate-, and high-risk groups, enabling interpretable screening outcomes. In addition to risk classification, preventive guidance aligned with predicted risk levels enhances the practical relevance of the system for early awareness and intervention planning. Integration of an interactive prediction dashboard further improves accessibility of autism risk assessment tools for supportive screening environments.

FUNDING INFORMATION

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ETHICS STATEMENT

This study did not involve human or animal subjects and, therefore, did not require ethical approval.

STATEMENT OF CONFLICT OF INTERESTS

The authors declare that they have no conflicts of interest related to this study.

REFERENCES

[1] G K. Mittal, K. S. Gill, D. Upadhyay, V. Singh and S. Aluvala, "Applying Machine Learning for Autism Risk Evaluation Using a Decision Tree Classification Technique," *2024 2nd International Conference on Computer, Communication and Control (IC4)*, Indore, India, 2024, pp. 1-6, doi: 10.1109/IC457434.2024.10486622.

[2] A. Tandon, B. D. Mazumdar, M. K. Pal, P. K. Gupta and N. K. Sahu, "Enhancing Early Autism Prediction Using Machine Learning Techniques," *2025 12th International Conference on Computing for Sustainable Global Development (INDIACom)*, Delhi, India, 2025, pp. 1-5, doi: 10.23919/INDIACom66777.2025.11115668.

[3] T. Arunprasath, M. Niranjana, S. M, B. p. V, M. P. Rajasekaran and K. Ramaraj, "Prediction Of Autism Spectrum Disorder Using Machine Learning," *2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, Krishnankoil, Virudhunagar district, Tamil Nadu, India, 2024, pp. 1-6, doi: 10.1109/INCOS59338.2024.10527777.

[4] K. Suthendran, V. L. Kumar, K. D. D. Yadav and S. Surya, "Ensemble-based Machine Learning and Deep Learning Approaches for Autism Prediction," *2025 6th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, Tirunelveli, India, 2025, pp. 346-351, doi: 10.1109/ICICV64824.2025.11085665.

[5] B. Kamala, K. S. Mahanaga Pooja, S. Varsha and K. Sivapriya, "ML Based Approach to Detect Autism Spectrum Disorder (ASD)," *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, Chennai, India, 2021, pp. 313-318, doi: 10.1109/ICCCT53315.2021.9711826.

[6] S. K. R. Naik, D. M, R. P B, S. Prakash and U. J. Royal, "Determination and Diagnosis of Autism Spectrum Disorder using Efficient Machine Learning Algorithm," *2023 3rd International Conference on Intelligent Technologies (CONIT)*, Hubli, India, 2023, pp. 1-5, doi: 10.1109/CONIT59222.2023.10205718.

[7] U. R, S. S and S. K. K S, "Autism Spectrum Disorder Prediction Using Machine Learning," *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, Chennai, India, 2024, pp. 1-6, doi: 10.1109/ADICS58448.2024.10533493.

[8] L. Mertz, "Using AI and ML to Predict Autism Spectrum Disorder," in *IEEE Pulse*, vol. 15, no. 3, pp. 11-15, Sept.-Oct. 2024, doi: 10.1109/MPULS.2024.3443489.

[9] M. R. Ahmmad, H. Pantazopoulos, and S. H. Kothiya, “A comprehensive causal and machine learning framework for autism spectrum disorder risk prediction and high-risk subgroup identification from children with chronic health and genetic disorders in the United States,” *Research in Autism*, vol. 131, p. 202826, Jan. 2026, doi: 10.1016/j.reia.2026.202826.

[10] Sankari and A. Kannammal, “Diagnosis of autism spectrum disorder—prediction and analysis of enhanced classification model based on machine learning,” in *Elsevier eBooks*, 2025, pp. 241–257. doi: 10.1016/b978-0-443-30080-6.00018-3.

[11] S. Irram and M. Suaib, “An early detection of autism spectrum disorder using machine learning,” *Computers in Biology and Medicine*, vol. 198, no. Pt B, p. 111238, Oct. 2025, doi: 10.1016/j.compbiomed.2025.111238.

[12] R. Muñoz-Terol, J. Peral, S. Amador, and D. Gil, “A systematic review of machine learning techniques to address diagnosis and treatment of autism: challenges and opportunities,” *Heliyon*, vol. 12, no. 1, p. e44359, Jan. 2026, doi: 10.1016/j.heliyon.2025.e44359.