

Analysis and Detection of Autism Spectrum Disorder Using ML Techniques

Shantanu Mane, Sagar More, Aman Mulani, Prof. Dnyandeo Khemnar

Department of Information Technology GHRCEM Pune, India

DOI: <https://doi.org/10.51583/IJLTEMAS.2025.1407000042>

Abstract—Diagnosing Autism Spectrum Disorder (ASD) is challenging due to its complexity and the diverse symptoms it presents. In this study, we focus on applying machine learning techniques, specifically the Random Forest algorithm, for identifying ASD. Utilizing a comprehensive dataset that encompasses both behavioral and demographic information, we perform thorough preprocessing, feature selection, and model evaluation. The study examines the Random Forest classifier's effectiveness in differentiating between individuals with and without ASD. The results are encouraging and highlight the algorithm's predictive capabilities. By concentrating solely on this method, we gain insights into its strengths and limitations, which are critical for enhancing ASD diagnostic processes. This research underscores the potential of Random Forest in advancing early ASD detection and improving intervention strategies in clinical practice.

Index Terms - Autism Spectrum Disorder (ASD), Machine learning, Random Forest, Demographic data, Model performance, Early detection.

I. Introduction

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition that manifests as challenges in social interactions, communication, and repetitive behaviors. With approximately 1 in 54 children in the U.S. impacted by ASD, the demand for effective detection and intervention methods is high. Early diagnosis is key, as it can greatly enhance treatment outcomes and improve the quality of life for individuals with ASD. Despite advancements in research, early identification and intervention remain challenging, highlighting the need for ongoing investigation in this field.

Autism Spectrum Disorder (ASD) has historically been linked to a wide range of symptoms, such as difficulties with behavior, facial expressions, emotional regulation, and mental health. Researchers have explored various methods for diagnosing ASD, including medical imaging techniques like structural magnetic resonance imaging, as well as advanced approaches such as artificial intelligence and machine learning. However, significant obstacles remain in creating accurate and reliable diagnostic models, particularly concerning issues of accessibility, privacy, and data security. One major challenge is the limited availability and reach of data, which hampers the development of dependable intelligent models for ASD identification.

In this study, we focus on the application of a specific machine learning algorithm—the Random Forest classifier—for identifying ASD using a comprehensive dataset that captures both behavioral and demographic attributes. Our primary objective is to assess whether the Random Forest algorithm can accurately detect individuals with ASD through careful feature selection, preprocessing, and thorough model evaluation. We contribute to the ongoing research on ASD diagnosis by highlighting the strengths and limitations of this approach and offering insights that may assist in the development of improved early diagnosis and intervention strategies in clinical settings.

II. Literature Review

The complexities of diagnosing and managing Autism Spectrum Disorder (ASD) have led researchers to explore a wide range of machine learning techniques to improve identification and prediction. Raj and Masood [1] utilized machine learning methods to analyze and detect ASD, contributing to the expanding research on computational approaches for ASD diagnosis. Vakadkar et al. [2] emphasized the importance of early intervention in their study of machine learning techniques used to diagnose ASD in children.

Heinsfeld et al. [3] leveraged deep learning alongside the ABIDE dataset, demonstrating the potential of advanced computational techniques to enhance diagnostic capabilities. Omar et al. [4] showed the effectiveness of facial features as diagnostic markers by using a machine learning model based on facial processing deficiencies to predict ASD in children. Chowdhury and Iraj [6] explored the predictive power of machine learning classifiers for ASD, shedding light on the capacity of computational models in this domain. Mashudi et al. [7] contributed to understanding ASD in adults through the application of machine learning techniques for classification across various age groups.

The ASD prediction model developed by Alwidian et al. [8] highlighted the significance of intelligent algorithms in aiding clinical diagnoses. Hossain et al. [9] carried out experimental research on ASD diagnosis across different age groups, demonstrating the flexibility of machine learning methods in addressing developmental variations. In terms of data processing, categorical variables such as “gender,” “ethnicity,” and “jaundice” were transformed using techniques like label encoding or one-hot encoding to make them suitable for machine learning models.

Chaitra et al. [15] introduced a diagnostic prediction model based on complex network measures within a machine learning framework, offering a fresh perspective on ASD detection. Overall, the existing literature showcases the extensive range of machine

learning techniques, from deep learning and complex network metrics to traditional classifiers, that are being applied to ASD detection. These approaches provide promising opportunities to enhance early diagnosis and intervention for individuals with ASD.

III. Methodology

A. Dataset

The dataset used in this study includes a broad range of clinical and demographic characteristics linked to individuals diagnosed with Autism Spectrum Disorder (ASD). The dataset includes variables such as age, gender, ethnicity, history of jaundice, family history of autism, residence location, prior screening history, and results from the 10-item Autism Spectrum Quotient (AQ) screening tool. Each patient is identified by an ID, and the target variable, "Class/ASD," indicates whether the individual has an ASD diagnosis (1) or not (0).

A1 Score to A10 Score: Represent scores from various assessment questions used to evaluate aspects of behavior or development related to autism.

Age: Represents the child's age.

Gender: Specifies the child's gender.

Jaundice: Indicates whether the child had jaundice at birth.

Autism: Binary target variable, where 1 indicates the child has been diagnosed with autism, and 0 indicates otherwise.

Relation: Binary field that might show the relationship of the individual providing the data (0 or 1), indicating, for instance, parent or caregiver.

B. Data Cleaning

Data cleaning played a crucial role in preparing the dataset for analysis and modeling. Several measures were taken to ensure the data's integrity and suitability for machine learning.

Handling Missing Values: Missing values were present in categories like "age," "gender," and "ethnicity." These values were either removed or imputed to ensure completeness and reliability.

Normalization: Numerical attributes, including "score" and "age," were normalized to bring them to a common scale, preventing skewed analysis due to varying feature ranges.

Encoding Categorical Variables: Attributes such as "gender," "ethnicity," and "jaundice" were converted using one-hot or label encoding to make them compatible with machine learning models.

Removing Redundant Attributes: Redundant features like "ID" and "age description" were removed to improve modeling speed, computational efficiency, and to reduce the risk of overfitting.

C. Model Selection

Selecting the right machine learning model was essential for accurately identifying ASD. For this study, the **Random Forest Classifier** was chosen due to its robustness, flexibility, and ability to handle high-dimensional data with complex feature interactions.

Random Forest Classifier:

The Random Forest algorithm is an ensemble method that constructs multiple decision trees during training and aggregates their predictions. For classification tasks, it calculates the mode of the predicted classes. By building different trees using various subsets of data, this method helps reduce overfitting. Randomness is introduced in feature selection at each node, which further improves the generalization performance by reducing correlation between trees. This method is well-suited for ASD detection because it efficiently handles large datasets, can automatically select important features, and is resilient to noisy data.

D. Backend Development

The backend of the ASD detection system was developed using the Python web framework Flask. The primary objective was to create an efficient, user-friendly web application capable of predicting ASD characteristics based on user inputs.

The core backend logic was implemented in the app.py script, which integrated the trained machine learning model. The **predict()** method processed input data from the user, scaled it with a pre-trained scaler, and applied the Random Forest model to generate predictions. The results were then displayed on the user interface.

An API endpoint (/predict api) was also developed to handle JSON-formatted requests, enabling easy integration with other services. The API returned predictions in a structured format for further processing by client applications.

E. Frontend Development

The frontend of the ASD detection system was designed to be intuitive and user-friendly, enabling users to input relevant data and receive predictions regarding potential ASD symptoms in children.

HTML Templates: The primary template file (index.html) featured a form to collect data like age, gender, ethnicity, jaundice history, family history of autism, and screening test results. The form was clearly structured to ensure ease of use.

CSS Styling: The `styles.css` file was used to enhance the visual appeal and usability of the interface, with stylistic guidelines for buttons, input fields, text, and overall layout. The goal was to create a visually cohesive, engaging user interface.

Flask Templating Engine: Dynamic rendering of content based on server-side data allowed for smooth integration between the frontend and backend. Upon submission of the input form, asynchronous HTTP requests were made to the backend for processing, and the results were displayed dynamically on the frontend based on the predictions received.

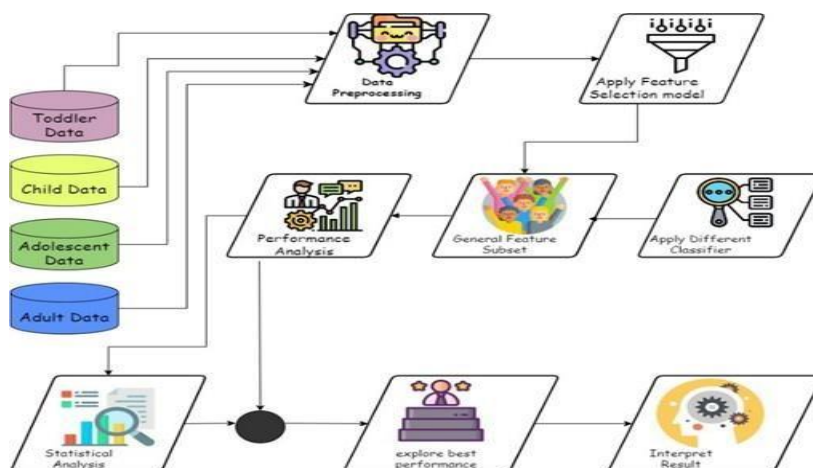
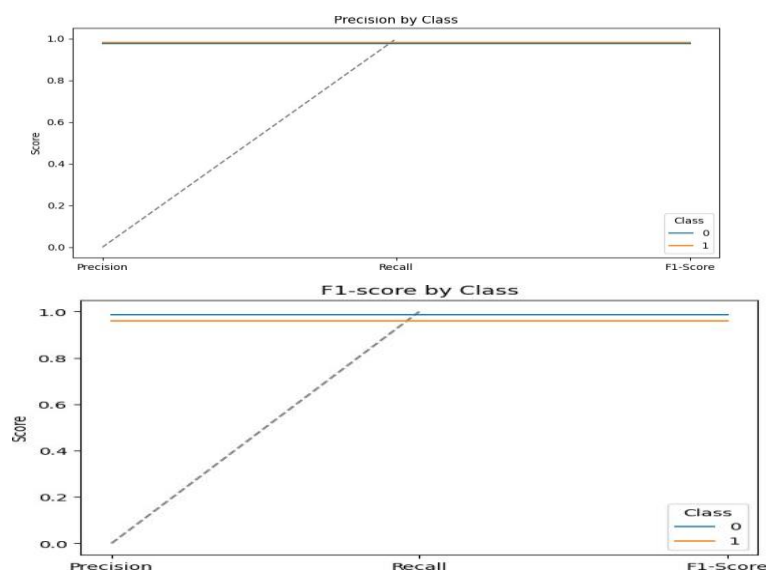


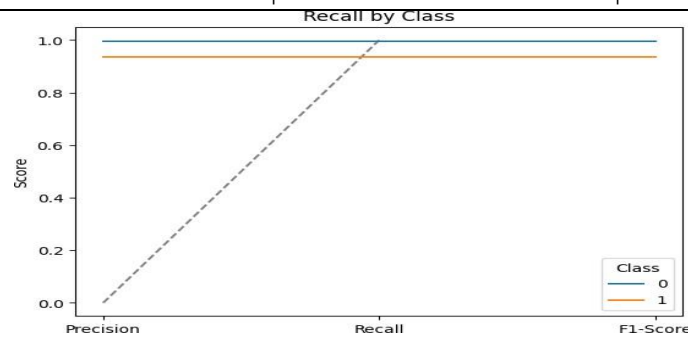
Fig. 1. Flow Diagram for Proposed System

IV. Results and Evaluation

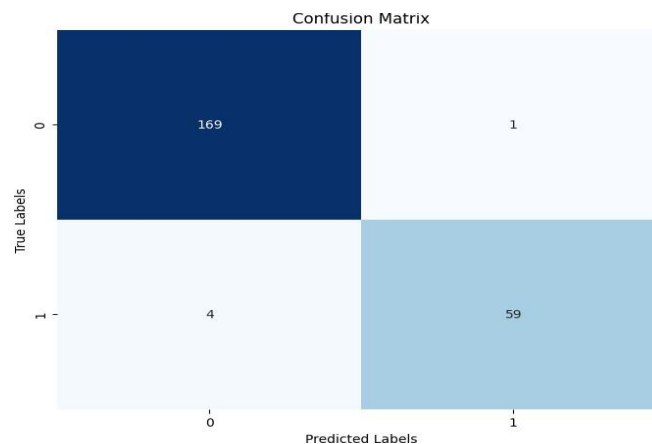
Utilizing demographic and behavioral characteristics, the Random Forest classifier—a widely recognized machine learning algorithm—was employed to identify children diagnosed with Autism Spectrum Disorder (ASD). The model achieved a remarkable accuracy rate of 97.85% on the test dataset, showcasing the Random Forest algorithm's efficacy in accurately classifying individuals with and without ASD.

A detailed examination of the model's performance metrics, presented in the classification report, reveals that the Random Forest model excels in precision and recall for both ASD-positive and ASD-negative cases. The model recorded recall rates of 94% for non-ASD cases and 99% for ASD cases, indicating its capability to reduce false negatives and false positives. Furthermore, it achieved precision rates of 98% for both ASD and non-ASD classifications, reinforcing its reliability in making accurate predictions.

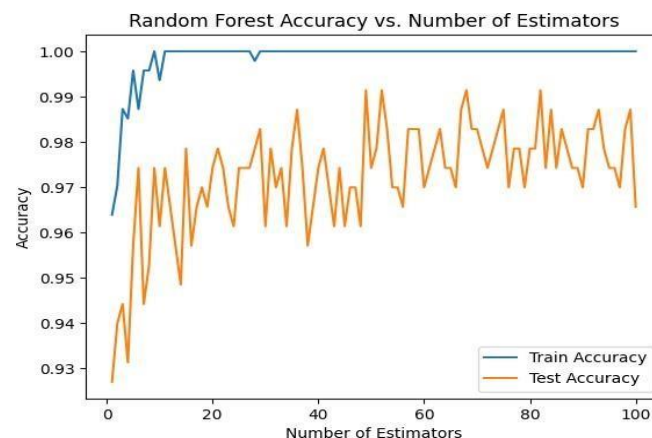




The confusion matrix, which illustrates the distribution of true positives, false positives, true negatives, and false negatives, further validates the robustness of the Random Forest model. There are minimal instances of misclassification, with most predictions aligning closely with the diagonal.



Additionally, analyzing the accuracy of the Random Forest model in relation to the number of estimators (trees) provides valuable insights into its behavior as model complexity varies. This examination helps in understanding how the model's performance evolves with the addition of more trees, allowing for better optimization and fine-tuning of the model to achieve the best results.



The graph depicting accuracy as a function of the number of estimators allows for the identification of the optimal balance between model complexity and performance. This analysis facilitates the determination of the ideal number of trees needed to minimize computational resource usage while maximizing accuracy, thereby aiding in the effective tuning of the model for improved outcomes.

V. Discussion

The use of the Random Forest classifier for identifying Autism Spectrum Disorder (ASD) yields positive results, demonstrating its potential as a valuable tool in clinical practice. Our study reveals a high level of accuracy and reliable performance metrics, underscoring the effectiveness of this machine learning approach in accurately classifying individuals with ASD. This highlights its significance as a diagnostic and therapeutic resource for autism spectrum disorders (ASDs), offering healthcare professionals a dependable means to identify and promptly address ASD-related challenges.

However, it is essential to acknowledge the inherent limitations and challenges associated with implementing machine learning algorithms in clinical settings. While the Random Forest model exhibits remarkable accuracy, its interpretability may be restricted

compared to simpler models like logistic regression. This lack of interpretability can hinder healthcare professionals from understanding the factors influencing the model's predictions, potentially impacting their decision-making in clinical practice. Future research should aim to enhance the interpretability of Random Forest models by exploring feature importance analysis and visualization techniques to clarify the underlying reasons behind the model's predictions.

Moreover, considerable attention must be given to the scalability and generalizability of the Random Forest classifier across diverse demographics and communities. Since our study relies on a specific dataset, it may not adequately capture the complexity and variation of ASD presentations across different populations. Consequently, further investigation with more extensive and diverse datasets is necessary to validate the model's effectiveness across various demographic groups. Additionally, exploring alternative feature selection methods and pre-processing techniques may enhance the model's applicability in different clinical environments and optimize its performance. By addressing these challenges and limitations, we can maximize the potential of machine learning algorithms like Random Forest in ASD identification, ultimately improving clinical outcomes and interventions for individuals with ASD.

VI. Conclusion

This paper concludes that the Random Forest classifier demonstrates high accuracy and robust performance metrics, establishing it as a promising tool for the detection of Autism Spectrum Disorder (ASD). We have shown that machine learning algorithms can significantly assist healthcare professionals in the early diagnosis and intervention of ASD through meticulous pre-processing, feature selection, and model validation. The encouraging results of this study emphasize the necessity for ongoing research and development in this field to enhance and refine ASD diagnostic techniques.

However, it is crucial to recognize the challenges and limitations associated with the use of machine learning algorithms in clinical settings. Despite its impressive accuracy, the interpretability issues of the Random Forest model could complicate clinical decision-making. Future research should focus on improving the interpretability of machine learning models to provide healthcare practitioners with a clearer understanding of the factors influencing the model's predictions. This includes exploring methods for feature importance analysis and visualization.

Furthermore, additional studies are needed to assess the generalizability and scalability of machine learning algorithms across diverse communities and demographics. While our research presents promising findings based on a specific dataset, further investigations should aim to validate the model's performance in various clinical scenarios and demographic groups. By addressing these challenges and limitations, we can advance the field of ASD detection and contribute to the development of more effective early diagnosis and intervention programs, ultimately improving outcomes for individuals with ASD and their families.

References

1. Raj, Suman, and Sarfaraz Masood. "Employing machine learning techniques for the analysis and detection of autism spectrum disorder." *Procedia Computer Science* 167 (2020): 994-1004.
2. Vakadkar, K., Purkayastha, D., and Krishnan, D. "Utilizing machine learning techniques for the detection of autism spectrum disorder in children." *SN Computer Science* 2 (2021): 1-9.
3. Heinsfeld, Anibal Solon, et al. "Deep learning approaches for identifying autism spectrum disorder using the ABIDE dataset." *NeuroImage: Clinical* 17 (2018): 16-23.
4. Omar, Kazi Shahrukh, Prodipta Mondal, Nabila Shahnaz Khan, Md Rezaul Karim Rizvi, and Md Nazrul Islam. "Predicting autism spectrum disorder through a machine learning framework." In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1-6. IEEE, 2019.
5. Liu, Wenbo, Ming Li, and Li Yi. "A machine learning framework for identifying children with autism spectrum disorder based on their face processing abnormalities." *Autism Research* 9, no. 8 (2016): 888-898.
6. Chowdhury, K., and Iraj, M.A. "Using machine learning classifiers to predict autism spectrum disorder." In *2020 International Conference on Recent Trends in Electronics, Information, and Communication Technology (RTEICT)*, pp. 324-327. IEEE, November 2020.
7. Mashudi, Nurul Amirah, Norulhusna Ahmad, and Norliza Mohd Noor. "A machine learning approach for classifying adult autism spectrum disorder." *IAES International Journal of Artificial Intelligence* 10, no. 3 (2021): 743.
8. Alwidian, J., Elhassan, A., and Ghnemmat, R. "Predicting autism spectrum disorder using machine learning techniques." *International Journal of Recent Technology and Engineering* 8, no. 5 (2020): 4139-4143.
9. Hossain, Md Delowar, Muhammad Ashad Kabir, Adnan Anwar, and Md Zahidul Islam. "An experimental analysis of autism spectrum disorder detection using machine learning techniques across toddler, child, adolescent, and adult datasets." *Health Information Science and Systems* 9 (2021): 1-13.
10. Stevens, E., Dixon, D. R., Novack, M. N., Granpeesheh, D., Smith, T., and Linstead, E. "Unsupervised machine learning for identifying and analyzing behavioral phenotypes in autism spectrum disorder." *International Journal of Medical Informatics* 129 (2019): 29-36.
11. Eman, D., and Emanuel, A.W. "A review of machine learning classifiers for autism spectrum disorder." In *2019 4th International Conference on Information Technology, Information Systems, and Electrical Engineering (ICITISEE)*, Nov 20 (pp. 255-260). IEEE.

12. Bhola, Jyoti, Rubal Jeet, Malik Mustafa Mohammad Jawarneh, and Shadab Adam Pattekari. "Analyzing and identifying autism spectrum disorder using machine learning techniques." In *Artificial Intelligence for Accurate Analysis and Detection of Autism Spectrum Disorder*, pp. 69-81. IGI Global, 2021.
13. Nogay, Hidir Selcuk, and Hojjat Adeli. "The application of machine learning (ML) in the diagnosis of autism spectrum disorder (ASD) using brain imaging." *Reviews in the Neurosciences* 31, no. 8 (2020): 825-841.
14. Hyde, Kayleigh K., Marlena N. Novack, Nicholas LaHaye, Chelsea Parlett-Pelleriti, Raymond Anden, Dennis R. Dixon, and Erik Linstead. "A review of supervised machine learning applications in autism spectrum disorder research." *Review Journal of Autism and Developmental Disorders* 6 (2019): 128-146.