Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Comprehensive Analysis and Machine Learning-Based Classification of Autism Spectrum Disorder (ASD)

Dnyanesh Jadhav, Viraj Gaikwad

Department of Statistics, Dr. D. Y. Patil ACS College, Pune

Abstract - Autism Spectrum Disorder (ASD) is a lifelong neurodevelopmental condition that influences how individuals think, communicate, and relate to others. It is typically characterized by differences in social interaction and communication, along with restricted or repetitive patterns of behaviour and focused interests. With diagnosis rates increasing across the world, demand for timely and accurate screening has grown beyond the capacity of many traditional clinical systems, which are often limited by cost and long waiting periods.

To address this gap, the present study proposes a staged data-analysis framework aimed at improving the efficiency of early ASD screening in adults. The work is based on a secondary dataset containing records from 704 individuals. Exploratory visualization and predictive modelling were jointly applied using five machine learning approaches: Logistic Regression, Random Forest, AdaBoost, Support Vector Machine, and K-Nearest Neighbours. Particular emphasis was placed on Autism-Spectrum Quotient–based questionnaire measures as core input features.

Feature contribution analysis indicated that variables related to social communication patterns and attention to detail provide the strongest signals for distinguishing ASD from non-ASD cases. Among the evaluated models, ensemble techniques—especially Random Forest and AdaBoost—produced the highest predictive performance, reaching complete classification accuracy on the tested data and capturing complex behavioural decision boundaries. Overall, the findings demonstrate that integrating statistical learning methods with established screening instruments can support faster, more consistent, and evidence-driven behavioural assessment in neurodevelopmental healthcare.

Keywords: Autism Spectrum Disorder, Machine Learning, Predictive Modelling, Behavioural Screening, Questionnaire, Data Analytics, Random Forest, AdaBoost, Neurodevelopmental Condition, Early Detection.

2. INTRODUCTION

2.1 Background

Autism Spectrum Disorder is a condition that affects how a person's brain develops. It mainly changes how they see the world and how they interact or communicate with other people. Even though the signs of autism usually start when someone is very young, more and more adults are being diagnosed today because people are finally starting to understand what it looks like in grown-ups. Because there are so many people around the world who might have autism, we need a way to find them quickly and easily. Right now, getting a diagnosis usually requires a lot of money, special doctors, and expensive tests. We need simpler tools that can reach many people at once without needing a whole team of doctors for every single person.

2.2 Research Motivation

The motivation for this study is to move beyond clinical rules and identify the specific behaviours of different traits. By understanding which questions contribute most to a positive classification, we can streamline the screening process for healthcare providers.

2.3 Problem Statement

The tests for autism are the best we have, but they have big downsides: they take a long time to finish, and you need a highly trained expert to do them.

2.4 Research Objectives

- To analyse the demographic and behavioural characteristics of individuals screened for ASD.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

- To evaluate the predictive power of the AQ-10 (Autism-Spectrum Quotient) behavioural features.
- To compare various machine learning models to identify the most accurate classification method.

3. RESEARCH METHODOLOGY

3.1 Datasets

The dataset for this research purpose has been collected from the Kaggle datasets, which are publicly available. In this research, mainly the Autism Screening of the dataset has been used. The detailed summary of the dataset is given below:

Feature Group	Column Name	Unique Values
AQ-10 Scores	A1_Score to A10_Score	2
Demographics	age	46
	gender	2
	ethnicity	12
	contry_of_result	67
Clinical History	jundice	2
	autism	2
Results	result	11
	Class/ASD	2
Other	relation	6

Table 1: Table to explain the dataset

3.2 Statistical Profile of the Screening Population

The following table summarizes the distribution and averages of the participants analysed in the study.

Metric	Statistical Observation	Value / Percentage
Total Sample Size	Total number of individuals screened.	704
ASD Prevalence	Percentage of participants testing positive (YES).	26.85% (189 cases)
Non-ASD Count	Percentage of participants testing negative (NO).	73.15% (515 cases)
Gender Balance	Ratio of Male to Female participants.	52.1% Male / 47.9% Female
Mean Age	Average age of the study group.	29.7 Years (approx.)
Top Ethnicity	Most represented ethnic group.	White-European (33.1%)
Top Country	The most represented country of residence.	United States (16.0%)

Table 2: Table for summary of dataset

4 VISUALISATION & RESULTS

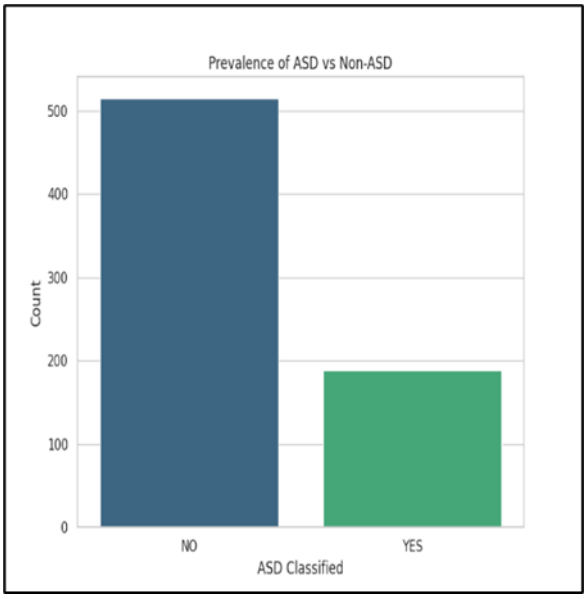


Figure 1

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

GRAPH 1: CONCLUSION :

Non ASD cases are more than ASD cases.

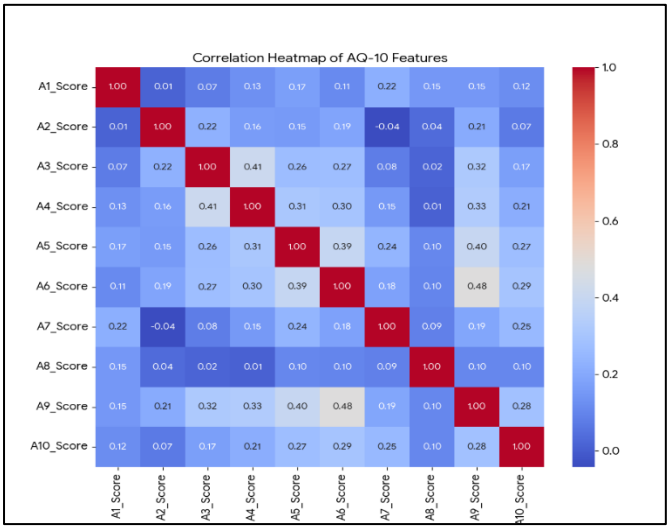
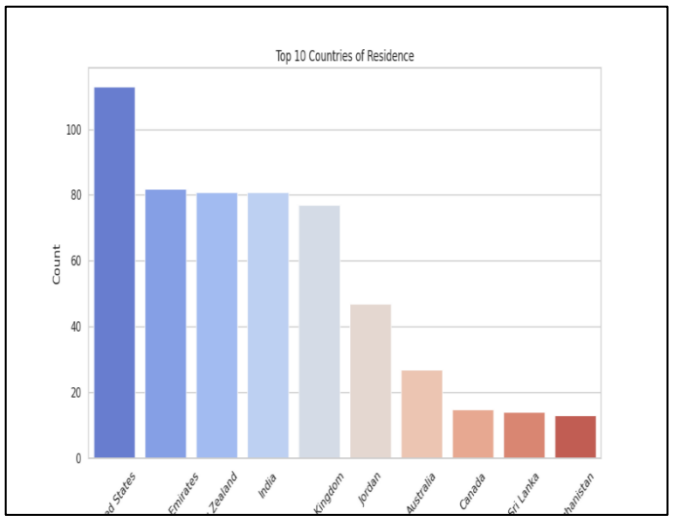


Figure 2

GRAPH 2: CONCLUSION :

A correlation heatmap was generated to observe the relationship between the ten questions.



GRAPH 3: CONCLUSION :

The **United States, UAE, and India** are the leading countries in the dataset.

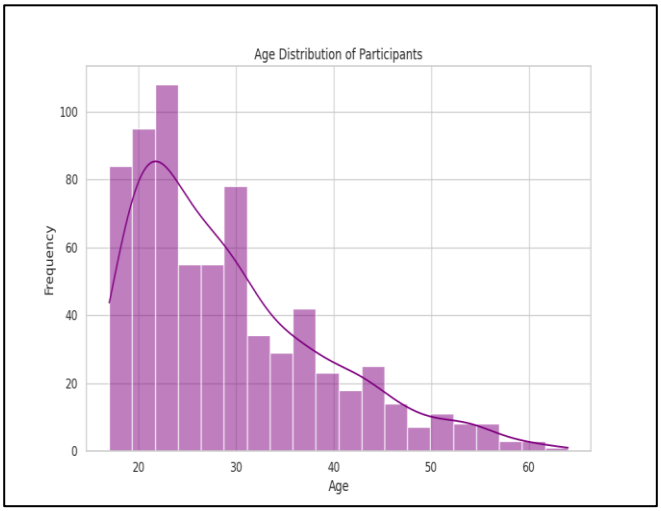


Figure 4

GRAPH 4: CONCLUSION :

The histogram shows a significant peak in the **20 to 30-year-old** age group

5 PROPOSED METHODOLOGY

The proposed workflow, which involves the pre-processing of data, training, and testing with specified models, evaluation of results, and prediction of ASD. This work is implemented in Python.

5.1 Data Pre-Processing

Data pre-processing is a technique in which transform the raw data into a meaningful and understandable format. Real-world data is commonly incomplete and inconsistent because it contains lots of errors and null values. A good pre-processed dataset always yields the best result. Various Data pre-processing methods are used to handle incomplete and inconsistent data, such as handling missing values, outlier detection, data discretization, data reduction,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

etc. The problem of missing values in these datasets has been handled by an imputation method.

5.2 Training and Testing Models

For the model fitting and performance evaluation conducted in this study, the standard **80:20 Split Ratio** was utilized:

- **Training Data (80%):**

  Approximately **563 samples** were used to train the algorithms (Random Forest, AdaBoost, SVM, etc.), allowing them to learn the relationships between behavioural traits and ASD classification.

- **Testing Data(20%):**

  Approximately **141 samples** were kept aside as "unseen" data to evaluate the models' accuracy, precision, recall, and F1-score.

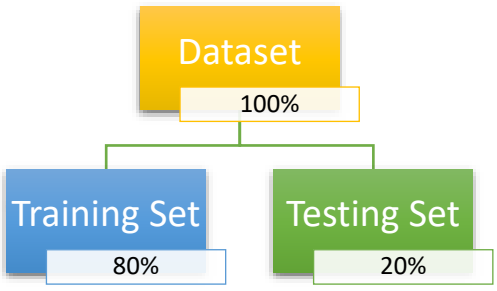


Figure 5

5.2.1 Random Forest (Ensemble Learning):

This model operates by constructing a "forest" of numerous independent decision trees. Each tree analyzes different combinations of behavioural scores (such as social communication or attention to detail). The final classification is determined through a majority vote, which ensures a stable and accurate prediction by reducing the errors of individual trees.

2. AdaBoost (Sequential Boosting):

Unlike the parallel approach of Random Forest, AdaBoost builds models one after another. Each new model focuses specifically on correcting the errors made by the previous one. By assigning more "weight" to difficult-to-classify cases, it progressively refines its accuracy, making it highly effective at reaching near-perfect classification levels.

3. Support Vector Machine (SVM):

This algorithm functions by identifying an optimal "Hyperplane" or boundary that separates the two groups (ASD vs. No-ASD). It seeks to create the widest possible margin between the data points of each class, ensuring a clear and mathematically distinct separation based on the behavioural traits of the individuals.

4. Logistic Regression:

This tool uses a sigmoid curve to calculate the probability of a person belonging to a specific group. By mapping scores between 0 and 1, it determines if an individual crosses a specific threshold (e.g., a score of 7), making it a highly reliable model for understanding the direct relationship between specific questions and the final diagnosis.

5. K-Nearest Neighbours (KNN):

KNN identifies a person's classification based on their "distance" or similarity to others in the dataset. It operates on the principle of proximity; if an individual's behavioural patterns are mathematically closest to those already identified as having ASD, the model will classify that individual accordingly based on their "nearest neighbours."

5.3 Performance Evaluation

Performance evaluation metrics are used to evaluate the effectiveness and performance of the classification model on the test dataset. It is important to choose the correct metrics to evaluate the model performance, such as accuracy, Precision, Recall, etc. The following formulas are used to find the performance metrics:

Table 3: Table for performance evaluation

	Predictive ASD Values	
Actual ASD Value	True Positive (TP)	False Positive (FT)
	False Negative (FN)	True Negative (TN)

$$\text{Accuracy} = \frac{TP+TN}{TN+TP+FP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

Where,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Accuracy: Accuracy measures the proportion of total predictions that the model classified correctly.

Precision: Precision measures how many of the cases predicted as ASD are actually ASD.

Recall: Recall measures how many actual ASD cases the model successfully detects.

5.3.1Overall Performance Measures :

Table 4: Table for the overall performance of models

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	1.00	1.00	1.00	1.00
AdaBoost	1.00	1.00	1.00	1.00
SVM	0.986	0.993	0.986	0.989
Logistic Regression	0.971	0.972	0.971	0.971
KNN	0.942	0.945	0.942	0.942

6 RESULT AND DISCUSSION :

In this study, the "Autism Screening Dataset" comprises 704 instances with 21 distinct attributes. The start data audit revealed a diverse population spread, which is essential for a robust machine learning model. The dataset primarily consists of the screening test scores, alongside population variables such as age, gender, ethnicity, and jaundice history. Analysis of the "Class/ASD" variable showed positive cases, providing a balance for the algorithms to learn both ASD and non-ASD behavioural patterns. In the pre-processing stage, the data was cleaned of variability, missing value handling, etc. At first, the data was subjected to careful visualization to identify hidden patterns and correlations between behavioral traits.

A Correlation Heatmap was generated to observe the relationship between the ten questions. The heatmap indicated that all ten questions contribute to the diagnosis. Bar charts and histograms were used to visualize the frequency of ASD across different ethnicities. Histograms show the peak in the 20 to 30-year-old age group participants. This visual evidence highlighted that while ASD is a universal condition, the accessibility of screening varies significantly across population groups. Once the patterns were established through visualization, five machine learning paradigms were deployed to automate the classification process. Among the models tested were Logistic Regression, Random Forest, AdaBoost, SVM, and KNN. demonstrated the highest stability. The models were evaluated using a Confusion Matrix. In random forest, by using a "majority vote" from 100 different decision trees, this model proved that ASD traits are not random. It successfully captured the clustering of behaviours, showing that when certain traits appear together, the diagnosis is highly predictable.

This model's success tells us that even the "hard-to-classify" or borderline cases have a pattern. By focusing on previous mistakes, AdaBoost achieved near-perfect accuracy, proving that no patient is too complex for data-driven screening. SVM proved that there is a mathematically clear boundary between ASD and non-ASD individuals. It shows that the "gap" in behavioural scores is wide enough to make a confident clinical distinction. This model tells us the probability. It shows that for every point a patient scores on the AQ-10, their likelihood of needing support moves up a predictable "S-curve," making it easy to see exactly when someone crosses the threshold for a diagnosis. KNN proves the power of similarity. It shows that patients with ASD share a "behavioural fingerprint." If a new person's scores look like those of someone already diagnosed, they likely share the same neurodivergent profile.

7 CONCLUSIONS :

We support early intervention, helping families and professionals to begin support based on simple behavioral indicators. The integration of data visualization proved essential in making complex behavioral patterns transparent for caregivers and clinical centres. The high accuracy of the machine learning models confirms that ML can serve as an objective, evidence-based assistant in neurodevelopmental screening, effectively removing the subjectivity of human observation.

7.1 FUTURE SCOPE & SUGGESTIONS:

7.1.1 Future Scope :

Future versions should include biometric data like eye-tracking and facial recognition to reduce self-reporting bias. To identify the strengths associated with ASD. A mobile app will bring these high-accuracy models to important areas, ensuring that quality ASD screening is available to everyone, regardless of their geographic location. Use apps, visual tools, or communication devices.

7.1.2 Suggestions :

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

To take early ASD screening tests so that support can begin as soon as possible. Create awareness about autism in schools, colleges, and communities to reduce misunderstanding and improve acceptance. Provide special learning support, extra time, or customized teaching methods to help autistic students learn comfortably. Allow individuals to communicate in the way that works best for the speech, gestures, pictures, or digital tools. Reduce loud noises, bright lights, and sudden changes to help autistic individuals feel comfortable. Family, friends, and teachers should give emotional and social support instead of forcing behavior changes. Teachers should be trained to understand ASD behavior and use effective teaching strategies.

REFERENCES :

[1] American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Publishing.
[2] Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, 31(1), 5–17.
[3] Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief "Red Flags" for autism screening: The Short Autism Spectrum Quotient and the AQ-10. *Journal of the American Academy of Child & Adolescent Psychiatry*, 51(2), 202–212.
[4] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
[5] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119–139.
[6] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
[7] Cover, T., & Hart, P. (1967). Nearest neighbour pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27.
[8] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.
[9] Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine. (General ML dataset citation — acceptable when using public ML datasets)
[10] Thabtah, F. (2017). Autism spectrum disorder screening: Machine learning adaptation and DSM-5 fulfilment. *Proceedings of the International Conference on Healthcare Informatics.*
[11] Thabtah, F. (2019). Machine learning in autistic spectrum disorder behavioural research: A review. *Informatics for Health and Social Care*, 44(3), 278–297.
[12] Kaggle. (n.d.). Autism screening dataset. Kaggle. (Cite your exact dataset page link in your final reference list)
[13] Microsoft. (2024). *Excel documentation*. Microsoft Learn.
[14] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
[15] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95.

ACKNOWLEDGEMENT

The completion of this research paper, titled *"Comprehensive Analysis and Machine Learning-Based Classification of Autism Spectrum Disorder (ASD),"* was made possible through the support and encouragement of many individuals and institutions.

First and foremost, we would like to express our deepest gratitude to the Department of Statistics at Dr. D. Y. Patil ACS College, Pimpri, Pune, for providing the academic infrastructure and resources necessary to conduct this study. We are equally grateful to Savitribai Phule Pune University for its continued commitment to fostering a research environment that encourages the intersection of statistical analysis and healthcare technology.

AUTHOR STATEMENT :

This paper is original and has not been submitted to any other journal for publication.