



(RESEARCH ARTICLE)



Predicting autism spectrum disorder through machine learning

Parwateeswar Gollapalli, Sana Tabasum *, Sai Kumar Ganta, Sidhartha Tadaboina and Aishwarya Gottipamula

Department of CSE (Data Science), ACE Engineering College, Hyderabad, Telangana, India.

World Journal of Advanced Research and Reviews, 2025, 25(02), 448-455

Publication history: Received on 25 December 2024; revised on 31 January 2025; accepted on 02 February 2025

Article DOI: <https://doi.org/10.30574/wjarr.2025.25.2.0380>

Abstract

Since social interaction, speech, and behaviour are all impacted by autism spectrum disorder (ASD), early detection is essential for prompt intervention. Through the analysis of behavioural and demographic data, this study creates a machine learning-based model to predict ASD in children. The system effectively classifies ASD cases using the Random Forest and XGBoost algorithms, making it more accessible than conventional diagnostic techniques. Model training, feature selection, and data preprocessing are all part of the methodology, and accuracy, precision, and recall measures are used to assess performance. In order to improve early diagnosis and intervention for improved cognitive and social development, the model seeks to offer an objective, scalable, and user-friendly screening tool.

Keywords: XGBoost; Random Forest; Machine Learning; Early Diagnosis; Autism Spectrum Disorder (ASD).

1. Introduction

The early identification of autism spectrum disorder (ASD) is essential in today's digital age to enable prompt interventions and enhance the lives of those impacted. However, the time-consuming nature of modern diagnostic methods like ADOS and ADI-R, which frequently require multiple visits and specialist tests, limits accessibility in disadvantaged areas. This challenge highlights the need for an efficient, data-driven approach to ASD prediction that enhances early detection and intervention.

To overcome these constraints, our project, Predicting Autism Spectrum Disorder, uses machine learning methods—more especially, the Random Forest and XGBoost algorithms to create a predictive model that can examine demographic and behavioural data. Our methodology improves diagnostic accuracy and offers an objective screening tool that reduces assessment time by identifying important ASD signals. More accessible and effective screening is made possible by this data-driven strategy, increasing the possibility of early intervention and support.

Early prediction of ASD has a profound effect on the fields of medicine and education. It facilitates early therapy methods by helping clinicians identify high-risk cases. In education, it assists parents and teachers in identifying behavioural patterns and learning challenges, enabling them to provide children with individualised support. This initiative advances objective, scalable, and effective diagnostic tools that benefit patients, families, and healthcare providers by using machine learning into ASD screening.

Machine learning-based ASD detection enhances early diagnosis and intervention by integrating advanced technology with traditional methods. The system ensures reliability through data preprocessing, feature selection, and evaluation using accuracy, precision, and recall. By addressing challenges like data variability and model interpretability, it supports timely and effective diagnosis.

* Corresponding author: Sana Tabasum

2. Related Work

Researchers have increasingly explored machine learning as a tool for detecting Autism Spectrum Disorder (ASD), aiming to complement traditional diagnostic methods such as behavioral assessments and clinical evaluations. While these conventional approaches are effective, they often demand substantial time and specialized expertise, prompting interest in automated solutions powered by artificial intelligence.

Various machine learning techniques, including Support Vector Machines (SVM), Decision Trees, and Neural Networks, have been applied to ASD classification. These models analyze behavioral patterns and responses from screening questionnaires, often achieving high accuracy. Additionally, feature selection methods help refine these models by minimizing redundant or irrelevant data, thereby improving predictive performance.

Publicly available datasets, such as those from the UCI Machine Learning Repository, have been instrumental in training ASD detection models. Some research efforts have also incorporated diverse data sources, including genetic and neuroimaging data, to enhance diagnostic accuracy. These advancements illustrate the expanding role of AI in autism research.

However, challenges persist in ensuring these models perform well across different populations. Many studies rely on small or imbalanced datasets that may not adequately represent variations in ASD symptoms across age groups and demographics. To address these issues, further research is needed on diverse and inclusive datasets, as well as explainable AI techniques that enhance model interpretability and reliability.

3. Existing System

Autism Spectrum Disorder (ASD) is presently diagnosed by clinical exams, behavioral assessments, and standardized screening instruments such as the Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R). These diagnostic procedures incorporate systematic observations, parent interviews, and examinations by qualified specialists to examine an individual's communication abilities, social interactions, and behavioral patterns. Despite being widely used and clinically approved, these techniques necessitate specialised knowledge, numerous testing sessions, and a substantial amount of time, which frequently makes early identification difficult.

The time-consuming and resource-intensive nature of the conventional ASD diagnostic procedure is one of its main drawbacks. Long interviews, structured play-based evaluations, and comprehensive questionnaires are all common components of assessments, and completing them might take hours or even many sessions. Furthermore, because behavioural observations are subjective, the results may differ based on the interpretation and experience of the clinician. This subjectivity can occasionally result in delayed or incorrect diagnoses of ASD, particularly when symptoms are modest or resemble those of other developmental disorders.

Furthermore, access to ASD diagnostic services is limited in many regions, particularly in rural or underdeveloped areas where there is a shortage of trained specialists. The requirement for in-person assessments creates logistical challenges, increasing wait times for diagnosis and delaying early intervention. Families often struggle to access timely evaluations due to long appointment waitlists, financial constraints, or geographic barriers. As a result, many children do not receive an official diagnosis until later in childhood, reducing the effectiveness of early therapeutic interventions.

Given these challenges, there is a growing need for automated, data-driven approaches that can assist healthcare professionals in diagnosing ASD more efficiently and objectively. Machine learning and artificial intelligence have the potential to revolutionize ASD screening by analyzing behavioral, demographic, and genetic data, providing faster and more accessible diagnostic tools. These advanced techniques can help overcome the limitations of the existing system by reducing subjectivity, increasing scalability, and enabling earlier detection, ultimately improving outcomes for individuals with ASD.

4. Proposed Model

The proposed ASD diagnosis system utilizes machine learning algorithms such as Random Forest and XGBoost to enable precise and automated screening. By analyzing behavioral characteristics and demographic data, the system effectively classifies individuals as either ASD-positive or ASD-negative. These algorithms excel in handling complex patterns within datasets, allowing for an objective and data-driven approach to ASD detection. Unlike traditional diagnostic methods that

rely on subjective assessments, this system ensures a more standardized and efficient evaluation process. Through feature selection and encoding, optimised preprocessing methods guarantee data integrity, and performance indicators like accuracy, precision, and recall confirm the model's efficacy.

The system uses optimised preprocessing methods, such as feature selection and encoding, to preserve data integrity and improve prediction reliability. By determining the most pertinent features and transforming categorical data into a format appropriate for machine learning models, these techniques improve the dataset. To evaluate the model's efficacy and make sure it consistently produces reliable results, performance metrics like accuracy, precision, and recall are used. These preprocessing procedures help create a screening tool that is more accurate and comprehensible by reducing noise and enhancing data quality.

This AI-powered system's capacity to lessen reliance on specialised experts is one of its main benefits, increasing accessibility and scalability for ASD screening. Access to qualified specialists is frequently limited in areas with inadequate medical resources, which causes diagnoses to be delayed. This system's automation of the screening process makes it possible to identify ASD early and provide timely therapies that can greatly enhance a person's developmental outcomes.

Additionally, scalability is improved by integrating machine learning into ASD diagnosis, enabling extensive screenings across a range of demographics. Over time, the system's predicted accuracy will increase due to its adaptability, which allows it to be continuously improved with new datasets. This strategy uses AI-driven insights to close the gap between current technological developments and conventional diagnostic limits, ultimately encouraging early diagnosis and improved assistance for people with ASD.

5. Methodology

The ASD prediction system's technique is centred on gathering a variety of behavioural and diagnostic data, creating an effective machine learning model, and facilitating precise ASD classification. To improve model performance, the procedure entails preprocessing the data, which includes feature selection and encoding. The system's flexibility for early diagnosis and intervention in the actual world is guaranteed by ongoing optimisation and assessment.

An explanation of the methods and algorithms used in machine learning.

Selected: The imported libraries are:

```
In [1]: # Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.metrics import (
    accuracy_score,
    precision_score,
    recall_score,
    f1_score,
    confusion_matrix,
    classification_report
)
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
```

Figure 1 Libraries imported

5.1. Data Collection

Data collection is crucial for building a reliable ASD prediction model. This step ensures access to diverse and representative data for accurate classification.

5.1.1. Dataset Selection

Use datasets containing behavioral and diagnostic information, such as those from clinical assessments or surveys. These datasets cover a range of features related to ASD symptoms and behavioral patterns.

5.1.2. Data Preprocessing:

Normalize numerical values and encode categorical variables for compatibility with machine learning models. Handle missing data and perform feature selection to retain the most relevant attributes. Data augmentation techniques, such as synthetic data generation, can be used to enhance dataset diversity and improve model robustness.

5.2. System Architecture

The system architecture for ASD prediction outlines the key components and their interactions, enabling efficient classification of individuals as ASD-positive or ASD-negative. It integrates data preprocessing, feature selection, and machine learning models to ensure accurate predictions and support early diagnosis.

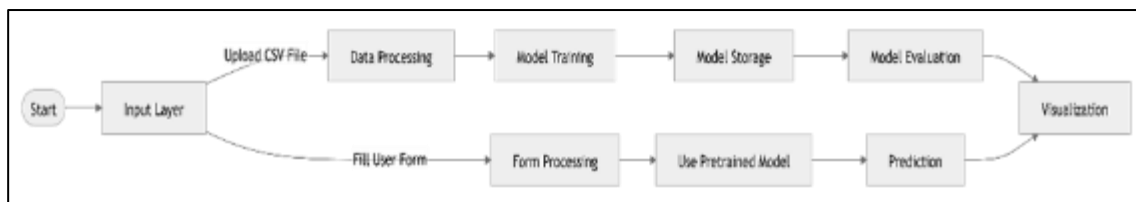


Figure 2 System Architecture

It illustrates how input data flows through various processing layers, resulting in meaningful outputs.

5.2.1. System Architecture Components

- **Input Module:** Captures behavioral and diagnostic data through surveys, clinical assessments, or sensors, ensuring a continuous flow of data for analysis.
- **Data Preprocessing Module:** Normalizes numerical data and encodes categorical variables, preparing the dataset for machine learning model training.
- **Machine Learning Module:** Uses algorithms like Random Forest and XGBoost to classify individuals as ASD-positive or ASD-negative based on the processed data.
- **Performance Evaluation Module:** Evaluates the model's accuracy, precision, and recall, ensuring reliable predictions with confidence levels.
- **User Interface (UI):** Displays real-time results, including predicted ASD status, confidence scores, and relevant insights in an accessible format.
- **Output:** Provides predictions in real-time with visual results, including ASD Classification and Confidence percentages.

5.3. Model Development

The model is the core component of the system, responsible for learning and predicting ASD.

5.3.1. Define Machine Learning Model Architecture:

Use algorithms like Random Forest and XGBoost to classify individuals based on behavioral and diagnostic features. Implement feature selection techniques to retain the most relevant attributes and improve accuracy. Train and validate the model to optimize performance and ensure reliable predictions.

5.3.2. Validation:

The data into training, validation, and test sets. Monitor metrics such as accuracy and recall on the validation set to identify overfitting. Use techniques like cross-validation and hyperparameter tuning to improve model generalization and performance.

```

In [ ]: def initialize_models(X, y):
    numerical_columns = X.select_dtypes(include=['int64', 'float64']).columns
    categorical_columns = X.select_dtypes(include=['object']).columns

    # Create and fit preprocessor
    preprocessor = ColumnTransformer(
        transformers=[('num', StandardScaler(), numerical_columns),
                     ('cat', OneHotEncoder(handle_unknown='ignore'), categorical_columns)])
    X_processed = preprocessor.fit_transform(X)

    # Train test split
    X_train, X_test, y_train, y_test = train_test_split(X_processed, y, test_size=0.2, random_state=42)

    # Train models
    rf_model = RandomForestClassifier(random_state=42).fit(X_train, y_train)
    xgb_model = XGBClassifier(use_label_encoder=False, eval_metric='logloss', random_state=42).fit(X_train, y_train)

    return preprocessor, rf_model, xgb_model, X_test, y_test

```

Figure 3 Model Building

5.3.3. Training:

Train the model using batch processing, dividing the dataset into smaller subsets for efficient learning. Apply gradient boosting techniques to minimize the loss function and update model parameters for improved accuracy.

```

In [ ]: # Train models
rf_model = RandomForestClassifier(random_state=42).fit(X_train, y_train)
xgb_model = XGBClassifier(use_label_encoder=False, eval_metric='logloss', random_state=42).fit(X_train, y_train)

```

Figure 4 Model Training

5.3.4. Testing:

Evaluate the model on unseen test data to ensure its ability to generalize to new cases. Use performance metrics such as accuracy, precision, and recall to assess the model's effectiveness in predicting ASD.

5.4. Real-Time Data Processing

This step ensures the system operates effectively in real-time environments for ASD prediction.

5.4.1. Input Capture:

Collect real-time behavioral and diagnostic data through surveys, clinical assessments, or sensors. Integrate tools for seamless data collection and ensure consistent input for analysis.

5.4.2. Data Preprocessing:

Apply normalization and encoding techniques to preprocess the data in real-time, ensuring compatibility with the machine learning model.

5.4.3. Feature Extraction:

Identify key features from the input data, such as behavioral patterns and diagnostic attributes, to focus on relevant indicators for ASD classification, improving prediction accuracy.

5.5. ASD Detection

This step involves predicting ASD status based on the extracted features.

5.5.1. Model Prediction

The proposed ASD diagnosis system employs trained machine learning models, specifically Random Forest and XGBoost, to classify individuals as ASD-positive or ASD-negative. These models analyze a combination of behavioral characteristics, demographic data, and diagnostic indicators to identify patterns associated with ASD. By leveraging decision trees and boosting techniques, the system effectively distinguishes between individuals at risk and those not affected. The predictive model undergoes extensive training using labeled datasets to enhance its accuracy, ensuring it can generalize well to new cases. Additionally, real-time classification allows for quick and efficient screening, which is particularly useful in large-scale applications, such as healthcare settings and community-based ASD detection programs. The automated classification process not only reduces the burden on medical professionals but also minimizes human error, leading to more consistent and objective results.

5.5.2. Confidence Scoring

To enhance reliability, the system generates confidence scores for each prediction, providing an estimate of how certain the model is in its classification. These scores, typically derived from probability distributions, offer valuable insights into the model's certainty and trustworthiness in determining ASD risk. High-confidence predictions indicate a strong likelihood of accuracy, whereas lower confidence scores signal cases that may require further evaluation by healthcare professionals. This feature enables clinicians and caregivers to prioritize cases needing more thorough assessments, reducing false positives and negatives. Moreover, confidence scoring plays a crucial role in personalized decision-making, allowing medical professionals to weigh model predictions alongside traditional diagnostic methods. By integrating this mechanism, the system improves transparency, making AI-driven ASD screening more dependable, interpretable, and clinically applicable in real-world settings.

5.6. Visualization

Delivering results in a user-friendly manner is crucial for system usability.

5.6.1. Output Display:

Develop a graphical user interface (GUI) using tools like Streamlit to display the ASD prediction results. Show the ASD classification (positive/negative), confidence scores, and relevant insights in an easily interpretable format.

6. Results and Discussion

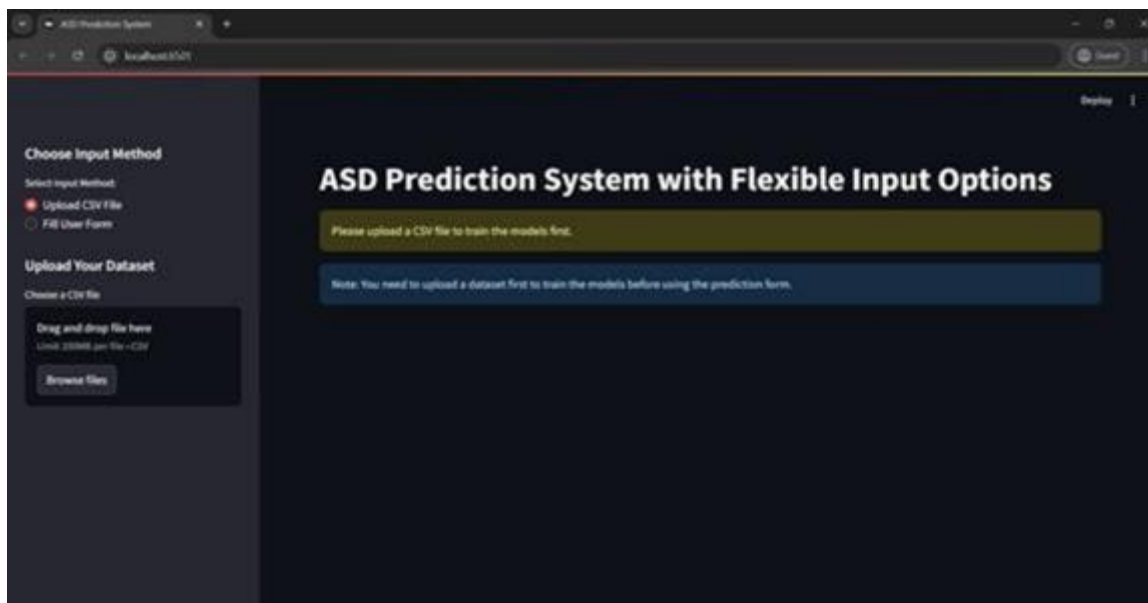


Figure 5 User Interface

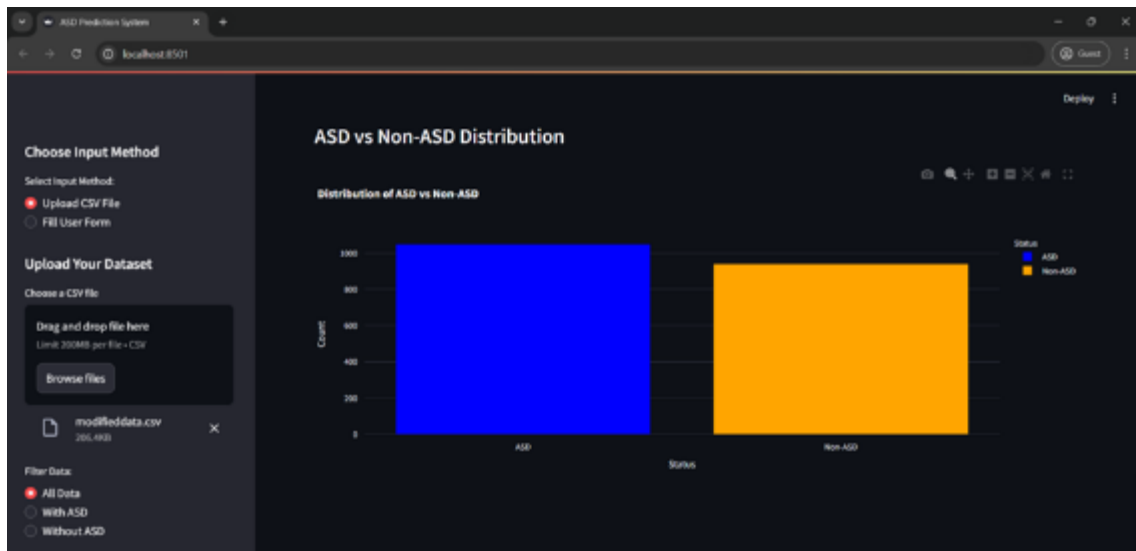


Figure 6 Autism Prediction

7. Conclusion

In conclusion, the proposed ASD detection system enhances traditional diagnostic methods by incorporating machine learning algorithms for more accurate and efficient early screening. By leveraging algorithms like Random Forest and XGBoost, the system effectively classifies individuals as ASD-positive or ASD-negative based on key behavioral and diagnostic features. The model's high accuracy, combined with real-time predictions, offers significant support to healthcare professionals, enabling more objective assessments. This system has the potential to improve early intervention and accessibility, particularly in resource-limited regions, helping to bridge the gap in early ASD detection. The adaptable, scalable nature of the system ensures its utility across diverse applications, contributing to better outcomes for individuals affected by ASD.

Compliance with ethical standards

Disclosure of conflict of interest





No conflict of interest to be disclosed.

References

- [1] https://www.researchgate.net/publication/377762482_Predicting_Autism_Spectrum_Disorder_Using_Machine_Learning_Classifiers
- [2] <https://www.nature.com/articles/s41598-023-35910-1>
- [3] <https://www.ijcaonline.org/archives/volume186/number63/taleb-2025-ijca-924437.pdf>
- [4] Available: <https://www.kaggle.com/datasets/uppulurimadhuri/dataset>
- [5] M. P. D. S. V. Srinivasan and A. S. M. S. Jadhav, "Predictive Modeling for Autism Spectrum Disorder Using Machine Learning Algorithms," *Procedia Computer Science*, vol. 165, pp. 272- 280, 2019.
- [6] <https://ieeexplore.ieee.org/document/10593455>
- [7] <https://www.scienceopen.com/hosted-document?doi=10.57197/JDR-2023-0064>
- [8] KUMAR, P. ASHOK, GSatish KUMAR, and SETTI NARESH KUMAR. "Improve the Capacity of Uniform Embedding for Efficient JPEG Steganography Based on DCT." (2015).
- [9] Kumar, P. Ashok, B. Vishnu Vardhan, and Pandi Chiranjeevi. "Investigating Context-Aware Sentiment Classification Using Machine Learning Algorithms." In *XVIII International Conference on Data Science and Intelligent Analysis of Information*, pp. 13-26. Cham: Springer Nature Switzerland, 2023.

- [10] Sunkavalli, Jayaprakash, B. Madhav Rao, M. Trinath Basu, Harish Dutt Sharma, P. Ashok Kumar, and Ketan Anand. "Experimentation Analysis of VQC and QSVM on Sentence Classification in Quantum Paradigm." In *2024 International Conference on Computing, Sciences and Communications (ICCSC)*, pp. 1-5. IEEE, 2024.
- [11] Kumar, P. Ashok. "Event Based Time Series Sentiment Trend Analysis."
- [12] PANDI, CHIRANJEEVI, THATIKONDA SUPRAJA, P. ASHOK KUMAR, and RALLA SURESH. "A SURVEY: RECOMMENDER SYSTEM FOR TRUSTWORTHY."
- [13] Kumar, P. Ashok, B. Vishnu Vardhan, and Pandi Chiranjeevi. "Correction to: Investigating Context-Aware Sentiment Classification Using Machine Learning Algorithms." In *XVIII International Conference on Data Science and Intelligent Analysis of Information*, pp. C1-C1. Cham: Springer Nature Switzerland, 2023.

Author's short biography

<p>Mr Parwateeswar Gollapalli</p> <p>Parwateeswar Gollapalli, a Software Engineer and Educator, specializes in AI, Network Drivers, and Automation. With an M.Tech (Gold Medalist) from Hyderabad Central University and a B.Tech from JNTU (K), he has industry experience at One Convergence as a Developer and QA Analyst. He has taught AI, Data Structures, and Compiler Design as a GATE CS subject matter expert. A multiple-time GATE qualifier, he has also cleared UGC NET-JRF, ISRO ICRB, and BHEL,HAL exams.</p>	
<p>Sana Tabasum</p> <p>A final-year B.Tech student at ACE Engineering College, specializing in CSE (Data Science). With a strong interest in data science and programming, constantly exploring new technologies to enhance knowledge and skills. Focused on applying expertise effectively in real-world scenarios.</p>	
<p>Sai Kumar Ganta</p> <p>A final-year Computer Science (Data Science) student at ACE Engineering College, passionate about data science, programming, and emerging technologies. Exploring new concepts and refining skills is exciting, with a strong eagerness to apply knowledge to solve meaningful challenges.</p>	
<p>Sidhartha Tadaboina</p> <p>Final-year B.Tech student specializing in CSE (Data Science) at ACE Engineering College, driven by a passion for programming and data-driven solutions. Enthusiastic about learning innovative technologies and continuously developing skills to make a meaningful impact in the field</p>	
<p>Aishwarya Gottipamula</p> <p>As a final-year B.Tech student specializing in CSE (Data Science) at ACE Engineering College, I am passionate about programming and data-driven solutions. I enjoy learning about innovative technologies and continuously developing my skills to make a meaningful impact in the field.</p>	