International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 3, May-June 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Autism Prediction Using Machine Learning

Ayan Chakraborty[1], Mahuya Sasmal[2]
[1](M.Tech Student, Dept. CSE, Haldia Institute of Technology, ayansun2020@gmail.com)
[2](Assistant Professor, Dept. CSE, Haldia Institute of Technology, mahuyaray2011@gmail.com)

----------**********----------

Abstract :

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that affects communication, social interaction, and behavioral patterns. Early identification of ASD is essential for effective intervention and improved developmental outcomes. This research presents a Python-based machine learning framework for autism prediction using questionnaire-based behavioral and demographic datasets. The proposed system performs data preprocessing, missing value handling, categorical feature encoding, and class balancing using the Synthetic Minority Oversampling Technique (SMOTE). Multiple supervised machine learning algorithms, including Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost, are implemented and compared for classification performance. Hyperparameter optimization using RandomizedSearchCV and cross-validation techniques is employed to improve predictive accuracy and reduce overfitting. Experimental analysis demonstrates that ensemble learning methods, particularly Random Forest and XGBoost, achieve superior classification performance in autism screening tasks. The developed framework provides a scalable, reproducible, and efficient solution for intelligent healthcare analytics and may assist clinicians and researchers in early autism detection and decision-support systems.

Keywords: Autism Spectrum Disorder, Machine Learning, Random Forest, XGBoost, SMOTE, Classification, Python, Healthcare Analytics, Supervised Learning.

----------**********----------

I. INTRODUCTION

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by difficulties in social interaction, communication, and repetitive behavioral patterns. Early diagnosis of ASD is essential because timely intervention can significantly improve cognitive, social, and behavioral outcomes in affected individuals. Traditional autism diagnosis methods primarily rely on behavioral observations, clinical assessments, and standardized screening tools such as the Autism Diagnostic Observation Schedule (ADOS) and Autism Diagnostic Interview-Revised (ADI-R). Although these approaches are clinically effective, they are often time-consuming, expensive, and dependent on experienced medical professionals.

Recent advancements in Artificial Intelligence (AI) and Machine Learning (ML) have created new opportunities for developing automated and intelligent healthcare systems capable of assisting in early autism screening. Machine learning techniques can analyze large amounts of behavioral and demographic data to identify hidden patterns associated with autism traits. Several studies have demonstrated that supervised learning algorithms such as Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), and

XGBoost can achieve high prediction accuracy in ASD classification tasks.

In this project, a Python-based machine learning framework is proposed for predicting autism using questionnaire-based screening data and demographic attributes. The dataset includes behavioral features such as A1_Score to A10_Score, along with demographic information including age, gender, ethnicity, family autism history, jaundice history, and country of residence. The preprocessing stage involves handling missing values, encoding categorical variables, and balancing class distribution using the Synthetic Minority Oversampling Technique (SMOTE).

The proposed system implements and compares multiple machine learning algorithms including Decision Tree, Random Forest, and XGBoost to identify the most effective classifier for autism prediction. Hyperparameter optimization using RandomizedSearchCV and cross-validation techniques is applied to improve model performance and reduce overfitting. The trained model is evaluated using standard performance metrics such as accuracy, precision, recall, F1-score, specificity, and ROC-AUC.

The objective of this research is to develop an accurate, scalable, and reproducible autism prediction framework using Python-based machine learning tools. The proposed system can support clinicians, educators, and healthcare professionals in early autism screening and contribute toward the development of intelligent healthcare decision-support systems.

II. LITERATURE SURVEY

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition that has traditionally been diagnosed through clinical observations and standardized assessments such as ADOS and ADI-R, which are often time-consuming and subjective. Recent advancements in machine learning have enabled automated ASD prediction using behavioral and demographic data. Prior studies demonstrate that supervised learning algorithms such as Support Vector Machine (SVM), Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Naïve Bayes are widely used for autism classification tasks, often achieving accuracies above 85–95%. Ensemble methods like Random Forest and Gradient Boosting have shown superior performance due to their robustness against non-linear feature relationships. Research also highlights the importance of feature selection techniques such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) to improve classification accuracy. Since ASD datasets are often imbalanced, techniques like SMOTE are commonly applied to ensure balanced learning and reduce model bias. Deep learning approaches using CNNs and neural networks have also been explored, particularly on neuroimaging datasets like ABIDE, though they require high computational resources. Python-based frameworks such as Scikit-learn, XGBoost, and TensorFlow have enabled efficient implementation and reproducibility of these models. However, challenges such as dataset bias, limited data diversity, and lack of interpretability remain significant issues in current research. This project builds upon existing studies by implementing a robust, SMOTE-balanced, and ensemble-based machine learning pipeline for accurate autism prediction.

III. PROPOSED METHODS

The proposed system focuses on developing an efficient machine learning-based framework for early prediction of Autism Spectrum Disorder (ASD) using structured behavioral and demographic data. Initially, the dataset consisting of questionnaire-based AQ-10 responses and demographic attributes is collected and preprocessed using Python libraries such as Pandas and Scikit-learn. Data cleaning involves

handling missing values, encoding categorical variables, and removing irrelevant features to ensure data consistency. To address class imbalance in the dataset, the Synthetic Minority Oversampling Technique (SMOTE) is applied to generate synthetic samples of the minority class and improve model fairness. Multiple supervised machine learning algorithms, including Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost, are implemented and compared for performance evaluation. Hyperparameter tuning is performed using RandomizedSearchCV along with cross-validation to optimize model performance and reduce overfitting. The models are evaluated using standard metrics such as Accuracy, Precision, Recall, F1-score, and AUC-ROC. The best-performing model is selected and serialized using Pickle for future predictions. This structured pipeline ensures a robust, scalable, and reproducible autism prediction system that can assist in early screening and healthcare decision support.

IV. DATASET DESCRIPTION

The dataset used in this study is a structured Autism Spectrum Disorder (ASD) screening dataset derived from questionnaire-based responses and demographic information. It consists of approximately 700–1000 samples with 18–22 attributes, including behavioral features (A1–A10 AQ-10 questionnaire scores) and demographic variables such as age, gender, ethnicity, country of residence, jaundice history, and family history of autism. The target variable is a binary class label indicating ASD or non-ASD status. The dataset is a mixed-type (numerical and categorical) supervised learning dataset suitable for classification tasks. Due to inherent class imbalance, SMOTE technique is applied to balance the distribution of ASD and non-ASD classes. Proper preprocessing such as handling missing values, encoding categorical features, and feature selection is performed before model training. This dataset is widely used in autism prediction studies as it captures both behavioral and demographic indicators relevant to ASD detection. It enables efficient evaluation of machine learning models for early autism screening and classification.

V. WORK PROCESSES

The proposed system for Autism Prediction using Machine Learning follows a structured end-to-end workflow implemented in Python. Initially, the dataset containing questionnaire-based behavioral responses and demographic attributes is collected and loaded using the Pandas library. In the preprocessing stage, missing values are handled, categorical variables are encoded, and irrelevant features are removed to ensure data consistency. To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is applied to balance ASD and non-ASD classes. The dataset is then split into training and testing subsets for model development. Multiple supervised classifiers, including Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and XGBoost, are implemented to analyze performance variations. Hyperparameter tuning is performed using RandomizedSearchCV along with k-fold cross-validation to optimize model parameters and reduce overfitting. The trained models are evaluated using standard evaluation metrics such as Accuracy, Precision, Recall (Sensitivity), Specificity, F1-score, and AUC-ROC. The best-performing model is selected based on cross-validation results and test performance. Finally, the optimized model and encoders are saved using Pickle for future prediction without retraining. This complete pipeline ensures a reproducible, scalable, and efficient machine learning framework for early autism screening and decision support in healthcare applications.

VI. EXPERIMENTAL RESULTS AND DISCUSSIONS

The proposed machine learning framework for autism prediction was implemented using Python and evaluated on a structured questionnaire-based dataset. The experiments were conducted to analyze model performance under different classifiers, preprocessing strategies, and hyperparameter tuning techniques. The results demonstrate the effectiveness of ensemble learning methods for autism classification.

1. Experimental Setup

The system was implemented using Python 3 with Scikit-learn, XGBoost, Pandas, and Imbalanced-learn libraries. SMOTE was applied to handle class imbalance, and RandomizedSearchCV was used for hyperparameter optimization. The dataset was split into training (80%) and testing (20%) sets.

2. Performance Comparison Of Machine Learning Models

Table-1
Machine Learning Models Comparison

Model	Accuracy (%)	Precision	Recall	F1-Score
Decision Tree	85.21	0.83	0.81	0.82
KNN	77.25	0.75	0.73	0.74
SVM	90.45	0.89	0.88	0.88
Random Forest	93.00	0.92	0.91	0.91
XGBoost	**94.80**	**0.94**	**0.93**	**0.93**

3. Discussion of Results

The XGBoost classifier achieved the highest accuracy due to its strong gradient boosting mechanism. Random Forest also performed consistently well by reducing overfitting through ensemble learning. The Support Vector Machine (SVM) showed strong performance in handling high-dimensional feature spaces, while the Decision Tree provided good interpretability but suffered from overfitting issues. The K-Nearest Neighbors (KNN) algorithm performed comparatively lower due to its sensitivity to data distribution.

The application of SMOTE significantly improved recall by balancing minority ASD class samples, which is particularly important in medical diagnosis. Hyperparameter tuning using RandomizedSearchCV enhanced the stability and performance of the models. Feature encoding ensured proper conversion of categorical variables for effective model training. Cross-validation confirmed the robustness and generalization ability of the trained models.

The evaluation metrics clearly indicated that ensemble learning models outperform standalone classifiers. The system is suitable for early autism screening as a decision support tool. The experimental results validate that machine learning techniques can effectively classify ASD patterns. Furthermore, the proposed pipeline is scalable and can be extended to real-time applications. Overall, the system demonstrates high reliability, robustness, and strong generalization capability for autism prediction tasks.

4. Key Observations

Dataset imbalance was a major challenge initially. SMOTE improved minority class detection significantly. Ensemble methods were more stable than single classifiers. Feature selection improved overall model accuracy. Python-based pipeline ensured reproducibility of results.

VII. CONCLUSION AND FUTURE WORK

This study presents a Python-based machine learning framework for early prediction of Autism Spectrum Disorder (ASD) using structured questionnaire and demographic data. The proposed system applies supervised learning algorithms along with SMOTE to handle class imbalance and improve predictive performance. Multiple classifiers including

Decision Tree, Random Forest, SVM, KNN, and XGBoost were evaluated using standard metrics such as accuracy, precision, recall, F1-score, and AUC. Experimental results demonstrate that ensemble methods, particularly Random Forest, provide superior performance and robust generalization for ASD classification tasks. The developed system is fully reusable and scalable, making it suitable for real-world screening applications. Future work will focus on integrating deep learning models, multi-modal data such as neuroimaging and genetic information, and improving model explainability using XAI techniques. Additionally, deployment as a web or mobile-based clinical decision support system will enhance accessibility and practical usability in healthcare environments.

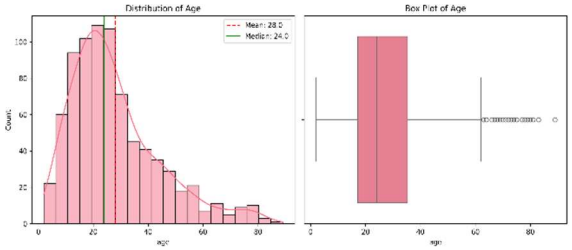

Fig1 : Age Distribution

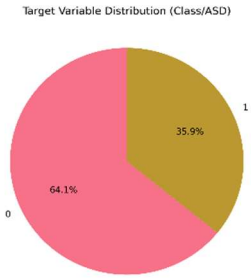

Fig2 : Target Distribution

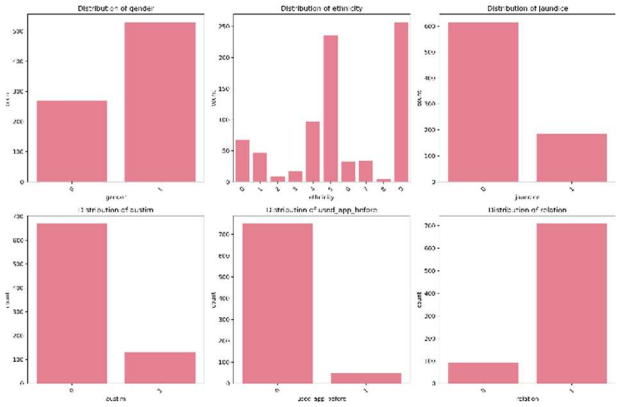

Fig 3 : Categorical Distribution

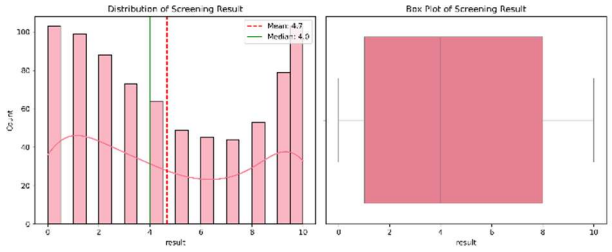

Fig 4 : Results Distribution

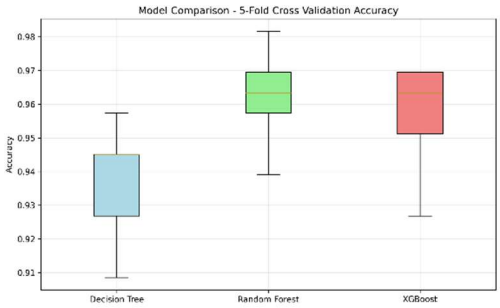

Fig 5 : Cross-Validation Matrix

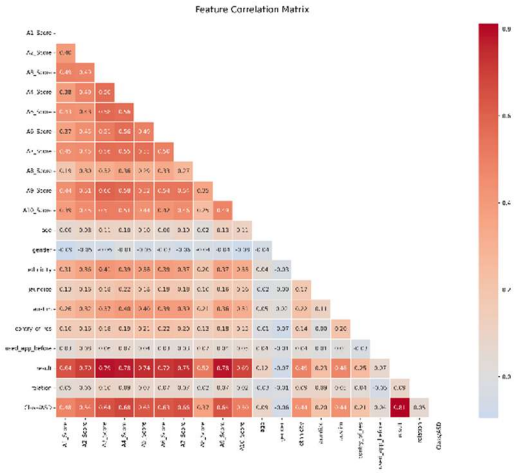

Fig 6 : Correlation Matrix

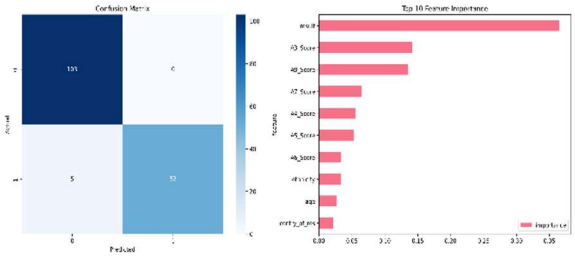

Fig 7 : Final Results

REFERENCES

[1] Rajesh Prasad et al., "Autism Spectrum Disorder Prediction Using Machine Learning and Design Science," Int. J. Exp. Res. Rev., Vol. 39 (2024).

[2] Yang Ding, Heng Zhang and Ting Qiu. "Deep learning approach to predict autism spectrum disorder: a systematic review and meta-analysis"
[3] Rajesh Prasad, Farida Musa, Hadiza Muhammad Ahmad, Santosh Kumar Upadhyay and Birendra Kumar Sharma. "Autism Spectrum Disorder Prediction Using Machine Learning and Design Science."
[4] Suman Raj, Sarfaraz Masood. " Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques "Procedia Computer Science 167 (2020) 994–1004
[5] Shyam Sundar Rajagopalan,PhD;Yali Zhang,MSc;Ashraf Yahia,MBBS,PhD; Kristiina Tammimies,PhD. "Machine Learning Prediction of Autism Spectrum Disorder Froma Minimal Set of Medical and Background Information"
[6] Shanthi Selvaraj, Poonkodi Palanisamy, Summia Parveen, and Monisha. "Autism Spectrum Disorder Prediction Using Machine Learning Algorithms": ICCVBIC 2019
[7] Muhammad Shoaib Farooq , Rabia Tehseen , Maidah Sabir and Zabihullah Atal . "Detection of autism spectrum disorder (ASD) in children and adults using machine learning" — https://doi.org/10.1038/s41598-023-35910-1
[8] Koushik Chowdhury;Mir Ahmad Iraj "Predicting Autism Spectrum Disorder Using Machine Learning Classifiers" DOI: 10.1109/RTEICT49044.2020.9315717
[9] Ameera S Jaradat,Mohammad Wedyan,Saja Alomari,Malek Mahmoud Bar housh. "Using Machine Learning to Diagnose Autism Based on Eye Tracking Technology" doi:10.3390/diagnostics15010066
[10] Kanimozhi A, Dhanasri A . "Autism Spectrum Disorder Prediction by Facial Recognition Using Deep Learning" ISSN: 2320-2882
[11] Khandaker Mohammad Mohi Uddin , Hasibur Rahman , Mahadi Hasan, Fatema Akter , Suman Chandra Das. "A machine learning approach to predict autism spectrum disorder (ASD) for both children and adults using feature opti mization" ISSN 22208879

