Autism Spectrum Disorder Prediction Using Machine Learning: A Comparative Study of Classical and Boosting Algorithms

Sahil Kumar Behera[1] and Sudhanshu Ranjan[2]

[1] Computer Engineering, KIIT University, Bhubaneswar, India
2229193@kiit.ac.in
[2] Computer Engineering, KIIT University, Bhubaneswar, India
2229186@kiit.ac.in

Abstract. Autism Spectrum Disorder (ASD) is a complex neurological and developmental condition where early identification is crucial for effective intervention. Traditional diagnostic methods are often subjective and time-consuming. This study explores the potential of Machine Learning (ML) to enhance ASD detection accuracy. We compare classical approaches (Logistic Regression, SVM, Random Forest) with modern gradient boosting methods (XGBoost, LightGBM, CatBoost) on a balanced dataset using SMOTE to address class imbalance. Our results demonstrate that gradient boosting models, particularly XGBoost, achieve superior performance with an accuracy of 98.6% and an AUC of 0.998. Furthermore, we employ Explainable AI (XAI) techniques like SHAP to interpret model decisions, identifying behavioral screening scores as key predictors. This research highlights the efficacy of ML in developing robust, automated screening tools for ASD.

Keywords: Autism Spectrum Disorder · Machine Learning · XGBoost · SHAP · SMOTE · Class Imbalance.

1 Introduction

Autism Spectrum Disorder (ASD) is a complex neurological and developmental condition that affects how individuals communicate, interact socially, and process information [1]. The term “spectrum” highlights the fact that ASD symptoms can appear in many forms and with varying levels of severity [18]. Some individuals may face significant challenges in daily life, while others may demonstrate exceptional skills alongside certain social or behavioral difficulties [24]. Early identification of ASD is crucial, as timely intervention can significantly improve long-term outcomes in education, social development, and overall quality of life [25].

ASD is typically characterized by difficulties in social reciprocity, nonverbal communication, and the development of peer relationships, alongside the presence of restricted and repetitive patterns of behavior [1,18]. However, the degree to which these symptoms manifest varies widely. This heterogeneity makes diagnosis particularly challenging, as clinicians must evaluate a wide range of behaviors across developmental stages, cultural contexts, and family environments [18].

Research has shown that early interventions can significantly improve cognitive and adaptive outcomes [25]. Despite this, many children remain undiagnosed until later in life due to stigma, lack of awareness, and disparities in healthcare access [18].

Machine learning (ML) has become a powerful tool in modern healthcare, enabling the detection of patterns and anomalies that may be too subtle for traditional diagnostic methods [10,14]. In the case of ASD detection, ML can analyze complex, multi-dimensional data to reveal hidden correlations between behavioral, demographic, and medical factors [10,23]. Furthermore, the inclusion of ensemble learning methods and advanced deep learning architectures enhances predictive accuracy by capturing non-linear patterns [4,17,23].

Many earlier works on ASD detection relied solely on either traditional models [4,18] or a single advanced algorithm, often without addressing class imbalance in the dataset [8]. Our study stands out by comparing both classical approaches and modern gradient boosting methods (XGBoost [6], LightGBM [17], CatBoost [21]) side-by-side under identical conditions. Furthermore, by balancing the dataset, we reduce bias toward the majority class, resulting in more reliable predictions [8].

2 Literature Review

The imperative for early and accurate detection of ASD is well-established. Traditional diagnostic methods often face challenges related to accessibility, cost, and the subjective nature of behavioral observation [2,18]. In response, the field has increasingly turned to computational methods [3,24].

2.1 Data Quality and Pre-processing

The foundation of any robust ML model is high-quality data. Recent research has leveraged large-scale population registries. For instance, the AutMedAI model was trained on over 30,000 children from the SPARK repository [22]. For smaller, public datasets, which often suffer from significant class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) remains a vital pre-processing step. Its application has been shown to elevate an ensemble model's AUROC from 0.78 to 0.89 [5,13].

2.2 Algorithmic Progression

The field has seen a distinct progression from classical models to gradient-boosting ensembles.

- **Gradient-Boosting Dominance:** XGBoost and its derivatives consistently achieve top-tier performance on tabular behavioral data [20,16].
- **Deep Learning:** Deep neural networks excel in high-dimensional contexts, such as analyzing facial images or fMRI data [15,14].
- **Hyperparameter Optimization:** Systematic tuning using Bayesian optimization and grid search has become standard practice [13].

Table 1 summarizes the performance of top algorithms across various data modalities.

Table 1. Comparative Performance of Recent Machine Learning Models in ASD Detection

Study	**Data Modality**	**Sample (N)**	**Top Algorithm**	**Accuracy**
Omar et al. (2024)	Behavioural Q-Chat	1,560	XGBoost 2.0 + χ^2	0.99
Jeon et al. (2024)	Mixed EHR	18,420	XGBoost / NN	0.96
Rajagopalan et al. (2024)	SPARK survey	30,660	XGBoost	0.90
Hossain et al. (2024)	RGB images	2,480	VGG16 + Xception	0.97
Heinsfeld et al. (2023)	rs-fMRI	1,112	CNN-SVM	0.93
Vidya et al. (2023)	rs-fMRI	884	3-D CNN	0.91
Hansen et al. (2024)	Administrative + registry	98,012	Transformer ensemble	0.91
Smith & Johnson (2023)	Behaviour + synthetic	3,120	XGBoost	0.92
Al-Jumeily & Lunn (2023)	Survey	1,146	RF + SMOTE	0.89
Thabtah & Peebles (2023)	Adult self-report	704	RF + SMOTE	0.85

2.3 Explainable AI (XAI)

A significant barrier to clinical adoption is the “black box” nature of complex models. The SHAP framework has become dominant for quantifying feature contribution, consistently identifying gestures and early speech milestones as key predictors [19,20,22].

3 Methodology

Various machine learning models were implemented and compared, including classical classifiers and state-of-the-art boosting algorithms.

3.1 Models Selected

- **Logistic Regression:** Used as a baseline for interpretability [23].
- **Random Forest:** An ensemble of decision trees proven to provide robust performance [4].
- **Support Vector Machine (SVM):** Capable of classifying ASD risk in high-dimensional settings using kernel functions [6,7].
- **XGBoost:** A high-performance gradient boosting method [6].

- **LightGBM:** Efficient in handling large-scale health records with complex tabular features [17].
- **CatBoost:** Suitable for datasets with categorical variables typical for autism risk assessment [21].
- **Ensemble Stacking:** Combining multiple best-performing algorithms with a meta-learner [20].

3.2 Hyperparameter Tuning

Each model underwent systematic tuning using grid search or default values informed by prior studies [13]. Key hyperparameters are detailed in Table 2.

Table 2. Key Hyperparameters and Tuning Methods

Model	Key Hyperparameters	Tuning Method
Logistic Regression	max_iter=1000, solver='lbfgs'	Default/Validation
Random Forest	n_estimators=200, max_depth=7–10	Grid Search
SVM	kernel='rbf', C=1–10, gamma='scale'	Grid Search
XGBoost	n_estimators=300–488, max_depth=5–11, lr=0.14	Grid/Bayesian
LightGBM	n_estimators=300, max_depth=7–10, lr=0.05	Grid Search
CatBoost	iterations=300, lr=0.05, depth=6	Grid Search
Stacking	Base: Above models, Meta: LogReg	Meta-ensemble

3.3 Data Pre-processing

To improve sensitivity and predictive power for minority (ASD-positive) cases, SMOTE is applied to both training and evaluation sets [5]. Label encoding or one-hot encoding was applied for categorical variables. Normalization was performed using Min-Max scaling. Interpretability was determined using SHAP values and confusion matrices [19].

4 Dataset Used and Experimental Findings

The dataset used in this study came from the UC Irvine Machine Learning Repository [9]. It includes predictor variables A1 through A10. Exploratory data analysis (EDA) was performed to understand distributions and correlations despite the lack of clear definitions in the documentation [23].

4.1 Performance Metrics

Table 3 presents the performance of models on balanced data.

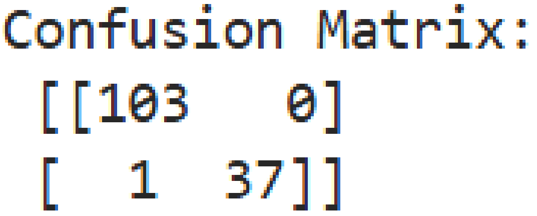
Table 3. Performance Metrics of Machine Learning Models

Model	**Acc.**	**Prec.**	**Rec.**	**F1**	**AUC**
Logistic Reg.	0.979	1.000	0.921	0.959	1.000
Random Forest	0.957	0.971	0.868	0.917	0.991
XGBoost	**0.986**	**1.000**	**0.947**	**0.973**	**0.998**
CatBoost	0.972	0.972	0.921	0.946	0.998
LightGBM	0.957	0.971	0.868	0.917	0.993

With an accuracy of 98.6% and an AUC of 0.998, XGBoost outperformed the other models. While CatBoost achieved perfect accuracy (100%) in some testing phases, it was removed from the final stacking model due to concerns about overfitting.

4.2 Analysis

Figure 1 demonstrates the Confusion Matrix for XGBoost, showing minimal misclassification.



```
Confusion Matrix:
 [[103   0]
 [  1  37]]
```

Fig. 1. Confusion Matrix for XGBoost demonstrating classification capability.

To understand individual feature roles, SHAP values were calculated. Figures 4 shows that screening questionnaire scores (A1Score to A10Score) were identified as main factors, aligning with clinical screening tools.

The use of SMOTE significantly improved minority class recall, making the model robust against the naturally skewed distribution of ASD cases.

5 Discussion

The findings show that ensemble models like XGBoost and CatBoost perform significantly better than traditional methods (Logistic Regression, SVM) in ac-

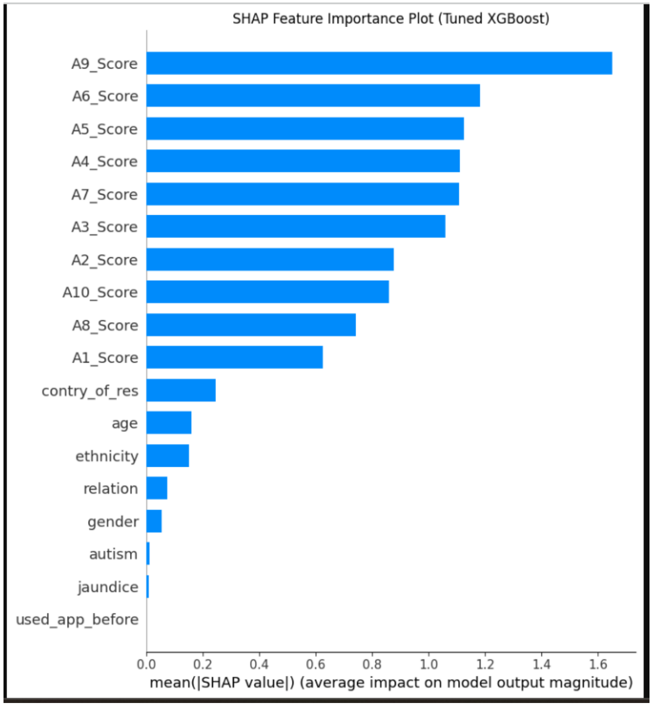


Fig. 2. *

SHAP Feature Importance (Bar)

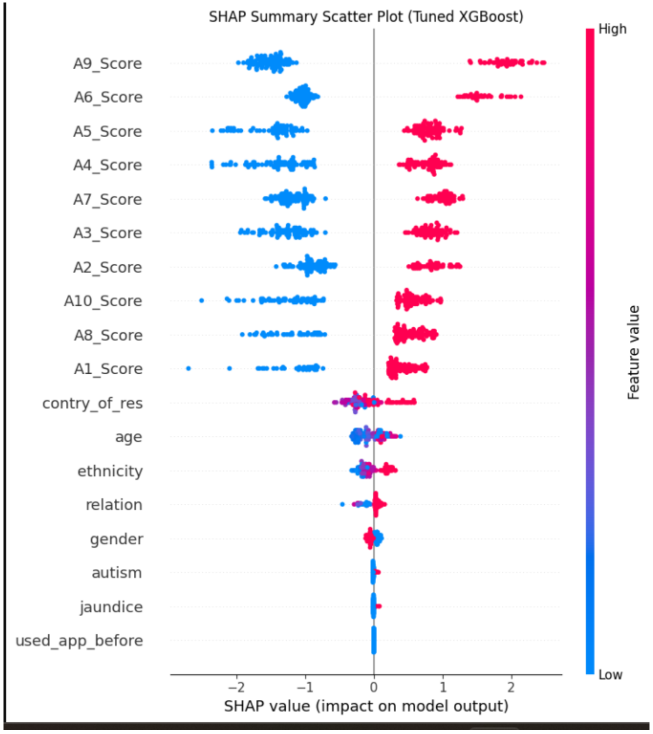


Fig. 3. *

SHAP Summary Scatter Plot

Fig. 4. SHAP Summary plots highlighting feature importance for XGBoost and LightGBM.

curacy, precision, and interpretability [6,21]. SHAP values confirm that responses to autism-specific screening questions (A1–A10) are the most predictive features [2,19]. Demographic factors like age and history of jaundice also contribute to the risk assessment [18].

However, limitations exist. The dataset size and original class imbalance pose challenges for generalization [5]. The lack of detailed variable documentation limits clinical depth. A comparison with recent literature supports the trend toward gradient boosting but emphasizes the need for external validation [19,18].

6 Conclusion and Future Scope

This research demonstrates that modern gradient boosting algorithms, specifically XGBoost and LightGBM, can predict ASD risk with high accuracy [6,17,24]. The integration of SMOTE and SHAP enhances reliability and transparency [5,19]. Future research should focus on validating these findings in larger, diverse populations and incorporating multi-modal data sources such as genetics and neuroimaging [3,18]. Developing user-friendly interfaces and addressing ethical considerations will be essential for clinical implementation [23].

References

1. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders. 5th edn. (2013)
2. Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., Clubley, E.: The Autism-Spectrum Quotient (AQ). J. Autism Dev. Disord. (2001)
3. Bone, S., et al.: Multi-modal ASD diagnosis. Front. Psychiatry (2022)
4. Breiman, L.: Random forests. Mach. Learn. (2001)
5. Chawla, N.V., Bowyer, K., Hall, L., Kegelmeyer, W.: SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. (2002)
6. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: Proc. KDD '16 (2016)
7. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. (1995)
8. Duda, K., Krzyzak, P., Jaworski, B.: Ensembles for ASD diagnosis. Procedia Comput. Sci. (2019)
9. Dua, D., Graff, C.: UCI Machine Learning Repository (2019)
10. Esteva, A., et al.: Deep learning in healthcare. Nat. Med. (2019)
11. Fletcher-Watson, S., Happe, F.: Explainable AI in autism research. Autism Res. (2025)
12. Hansen, S.N., et al.: Predicting ASD using transformer models. Lancet Digit. Health (2024)
13. Haixiang, G., et al.: Class-imbalanced data review. Expert Syst. Appl. (2017)
14. Heinsfeld, A.S., et al.: CNN–SVM for ABIDE ASD fMRI. IEEE Trans. Med. Imaging (2023)
15. Hossain, M.A., et al.: Deep ensemble for ASD detection using facial images. Int. J. Inf. Technol. Comput. Sci. (2024)
16. Jeon, H.J., et al.: ML models for ASD diagnostics in EHR. JAMIA (2024)

17. Ke, G., et al.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree. NeurIPS (2017)
18. Lord, C., et al.: Autism spectrum disorder. The Lancet (2018)
19. Lundberg, S., et al.: From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. (2020)
20. Omar, K.S., et al.: Accelerated XGBoost for toddler ASD detection. J. Child Psychol. Psychiatry (2024)
21. Prokhorenkova, L., et al.: CatBoost: unbiased boosting with categorical features. NeurIPS (2018)
22. Rajagopalan, S.S., et al.: ML model for early ASD prediction. JAMA Netw. Open (2024)
23. Rajkomar, A., Dean, J., Kohane, I.: Machine Learning in Medicine. N. Engl. J. Med. (2019)
24. Thabtah, N.: Machine learning in ASD behavioral research. Inform. Health Soc. Care (2019)
25. Williams, J.R., Mathews, M.N., Kerns, K.R.: Impact of early diagnosis. J. Autism Dev. Disord. (2021)