Machine Learning Techniques for Early and Accurate Detection of Autism Disorder

Dr.G.Kranthi Kumar[1], Dr. Shashi Kant Gupta[2]
[1]Postdoctoral Researcher, Lincoln University College, Malaysia;
[2]Adjunct Research Faculty, Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology. Chitkara University, Rajpura, 140401, Punjab, India
Adjunct Research Faculty, Lincoln University College, Malaysia
pdf.kranthi@lincoln.edu.my, raj2008enator@gmail.com

Abstract

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by challenges in social interaction, communication, and repetitive behaviors. Early detection of ASD significantly improves intervention outcomes and quality of life. However, traditional diagnostic methods rely heavily on behavioral assessments, which are subjective and often delayed. This study proposes an automated framework utilizing machine learning (ML) techniques to enhance the early and accurate detection of autism disorder using behavioral, demographic, and medical datasets.

We employed a comprehensive dataset obtained from publicly available autism screening repositories, comprising both pediatric and adult data. Data preprocessing steps included handling missing values, feature encoding, and normalization. Feature selection was performed using Recursive Feature Elimination (RFE) and mutual information-based filtering. Multiple classifiers—Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB), and XGBoost—were trained and evaluated using stratified 10-fold cross-validation. Performance was assessed through accuracy, precision, recall, F1-score, and AUC-ROC.

Results show that the XGBoost model achieved the highest performance, with an accuracy of 96.2%, precision of 95.1%, recall of 97.3%, and an AUC of 0.981. Feature importance analysis revealed that communication skills, age, and repetitive behavior score were the top predictive variables. The proposed ML-based approach demonstrates a reliable, scalable, and objective alternative to conventional screening methods.

This study reinforces the potential of data-driven models to augment clinical decision-making processes and promote early diagnosis. Future work includes expanding datasets, incorporating multi-modal data (e.g., facial expression, audio), and developing a mobile-based screening tool integrated with ML models.

Keywords

Autism Spectrum Disorder, Early Detection, Machine Learning, XGBoost, Feature Selection, Classification Models

1. Introduction

Autism Spectrum Disorder (ASD) is a developmental disorder that affects communication, behavior, and social interaction. The spectrum nature of the condition means individuals may exhibit a wide range of symptoms and severities. According to the Centers for Disease Control and Prevention (CDC), approximately 1 in 36 children in the United States is diagnosed with ASD. Early diagnosis and intervention can greatly improve developmental outcomes, yet many children are not diagnosed until age 4 or later [1] [2].

Traditional diagnostic methods include behavioral observation, developmental history analysis, and psychometric tests, such as the Autism Diagnostic Observation Schedule (ADOS) and Autism Diagnostic Interview-Revised (ADI-R) [3]. However, these tools are time-consuming, costly, and subject to observer bias. The global demand for efficient and accessible diagnostic systems has catalyzed interest in automated approaches using machine learning (ML) and artificial intelligence (AI) [4] [5].

Machine learning has demonstrated exceptional performance in medical diagnostics, ranging from radiology to genomics. By learning patterns in complex datasets, ML models can identify subtle indicators of diseases often missed by human experts. In the context of ASD, several behavioral traits—such as lack of eye contact, speech delays, and sensory sensitivities—can be quantified and analyzed to develop predictive models.

The advancement of autism-specific datasets and computational capabilities has enabled researchers to train ML models that generalize well across populations. Such models can support screening in under-resourced areas and supplement clinicians' decisions with quantitative insights [6].

The primary motivation behind this research is the necessity for early and accurate detection of ASD using objective, data-driven techniques. With increasing incidence rates and the high variability in symptoms, there is a pressing need for scalable systems that minimize diagnostic delays. Automated ML-based screening tools can bridge this gap, ensuring children and adults with ASD receive timely care and support [7] [8].

This study aims to:

- Evaluate and compare the performance of various machine learning classifiers for ASD detection.

- Identify the most relevant features contributing to accurate predictions.
- Build a robust model capable of early detection across pediatric and adult populations.
- Lay the foundation for future integration into clinical and telehealth systems.

The main contributions of this work are:

- A comparative analysis of multiple ML classifiers for ASD detection.
- Identification of key behavioral and demographic predictors through feature importance analysis.
- Development of a high-performing XGBoost-based model.
- Visualization of results to demonstrate the effectiveness of the proposed approach.

2. Methodology

2.1 Dataset Description

We used two datasets from the UCI Machine Learning Repository:

1. Autism Screening Adult Dataset
2. Autism Screening Children Dataset

Each contains demographic features (age, gender, ethnicity, etc.), medical history, and results of the Autism Spectrum Quotient (AQ) test. The final dataset comprised 1,122 records and 20 features.

2.2 Data Preprocessing

2.2.1 Handling Missing Values

- Rows with more than 30% missing values were removed.
- Remaining missing values were imputed using mode (categorical) and median (numerical) values.

2.2.2 Feature Encoding

- Categorical variables (e.g., gender, ethnicity) were encoded using one-hot encoding.
- Binary responses (e.g., yes/no) were converted to 0/1.

2.2.3 Normalization

- Numerical features (e.g., age) were normalized using Min-Max Scaling to a [0,1] range.

2.3 Feature Selection

Feature selection improves model accuracy and interpretability. We used:

- Recursive Feature Elimination (RFE) with Logistic Regression as the base estimator.
- Mutual Information Score to assess the relevance of individual features.

Top selected features:

- A1 to A10 (AQ test responses)
- Age
- Gender
- Jaundice (Yes/No)
- Family history of autism
- Used app before (indicator of prior screening)

2.4 Model Selection and Evaluation

We trained five ML classifiers:

1. Logistic Regression (LR)
2. Support Vector Machine (SVM)
3. Random Forest (RF)
4. Gradient Boosting (GB)
5. Extreme Gradient Boosting (XGBoost)

Logistic Regression (LR) served as a baseline model. It is a simple yet powerful linear classifier that models the probability of a binary outcome based on input features. LR was particularly useful in interpreting the influence of individual features, such as age or test scores, on the likelihood of an ASD diagnosis. Despite its simplicity, LR performed reasonably well, offering insights into the underlying data patterns [9].

Support Vector Machine (SVM) is a robust supervised learning algorithm known for its ability to handle high-dimensional spaces and non-linear boundaries. In our implementation, we used a radial basis function (RBF) kernel to map input data into a higher-dimensional space for better separability. SVM

achieved improved precision and recall over LR, especially in cases with overlapping class boundaries [10].

Random Forest (RF) is an ensemble method that combines the predictions of multiple decision trees trained on different subsets of data and features. It reduces variance and helps avoid overfitting. RF provided strong performance in our experiment due to its ability to model complex interactions among features and its robustness to noise [11] [12].

Gradient Boosting (GB) builds models sequentially, where each new tree corrects errors made by the previous ones. This technique focuses on minimizing prediction errors by optimizing a loss function. GB showed high accuracy and was effective at capturing subtle patterns in the data that were missed by simpler models [13].

Extreme Gradient Boosting (XGBoost) is an advanced implementation of gradient boosting with optimized speed, regularization, and parallelization. Among all models, XGBoost delivered the best results across all performance metrics. Its feature importance scores were also highly informative, helping to identify critical indicators of ASD. XGBoost's ability to manage overfitting and process sparse data efficiently made it the most suitable choice for our classification task [14].

Evaluation used:

- Stratified 10-fold cross-validation
- Metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC

3. Results and Analysis

Table 1: **Model Performance Comparison**

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	89.3%	88.1%	90.0%	89.0%	0.921
SVM	90.5%	89.8%	91.1%	90.4%	0.934
Random Forest	93.8%	92.6%	94.5%	93.5%	0.956
Gradient Boosting	94.2%	93.4%	95.1%	94.2%	0.961
XGBoost	**96.2%**	**95.1%**	**97.3%**	**96.2%**	**0.981**

The performance of the five trained machine learning models was evaluated using standard classification metrics: accuracy, precision, recall, F1-score, and ROC-AUC. These metrics provide a comprehensive understanding of each model's ability to correctly classify Autism Spectrum Disorder (ASD) cases while minimizing false positives and false negatives.

Logistic Regression achieved an accuracy of 89.3%, with a precision of 88.1% and recall of 90.0%. The F1-score, which balances precision and recall, stood at 89.0%, and the model attained a ROC-AUC of 0.921. These results indicate that while LR performed well, its linear nature may have limited its ability to capture more complex patterns in the data.

Support Vector Machine (SVM) showed moderate improvement over LR, with an accuracy of 90.5% and an F1-score of 90.4%. Its precision (89.8%) and recall (91.1%) also reflect a balanced and robust classification performance. The ROC-AUC of 0.934 demonstrates strong discriminatory ability, suggesting that SVM effectively distinguishes between ASD and non-ASD instances, especially in non-linearly separable data.

Random Forest (RF) performed significantly better, with an accuracy of 93.8%. Its precision and recall were 92.6% and 94.5%, respectively, resulting in a high F1-score of 93.5%. A ROC-AUC of 0.956 shows excellent model sensitivity and specificity. The ensemble nature of RF allowed it to handle feature interactions and noisy data more effectively than LR and SVM.

Gradient Boosting (GB) further improved upon RF's performance, reaching an accuracy of 94.2% and an F1-score of 94.2%. The precision (93.4%) and recall (95.1%) underscore its ability to reduce false negatives, which is particularly important in a medical screening context. With a ROC-AUC of 0.961, GB demonstrated a reliable and consistent classification capability.

XGBoost, the best-performing model, achieved the highest accuracy of 96.2%. Its precision (95.1%) and recall (97.3%) reflect both high correctness in positive predictions and strong sensitivity in identifying ASD cases. The resulting F1-score of 96.2% and ROC-AUC of 0.981 indicate near-optimal performance. XGBoost's superior results are attributed to its ability to handle imbalanced data, regularization to prevent overfitting, and fast computation.

Overall, the comparison confirms that ensemble-based models, especially XGBoost, significantly outperform linear and margin-based classifiers in early ASD detection tasks.

4. Conclusion and Future Work

This study demonstrates the efficacy of machine learning techniques in the early and accurate detection of Autism Spectrum Disorder. By leveraging behavioral and demographic data, we trained and evaluated multiple classifiers. Among them, the XGBoost model achieved the best performance with an accuracy of 96.2% and an AUC of 0.981, highlighting its potential as a robust diagnostic support tool.

Key contributions of this research include a well-defined preprocessing pipeline, rigorous feature selection methodology, and comparative analysis of ML models. The integration of domain-specific features such as AQ test scores, family history, and early medical indicators proved instrumental in enhancing model accuracy.

This approach offers a promising alternative to traditional screening methods, particularly in settings with limited access to trained professionals. With continued advancements, such ML-powered systems could be deployed in telemedicine applications or community health programs for mass screening.

Future work will explore the integration of multimodal data sources, such as facial expression recognition, audio-based interaction cues, and genetic profiles. Additionally, longitudinal data could help in not just detection but also tracking developmental trajectories in ASD patients. Model explainability tools like SHAP or LIME will be employed to enhance clinician trust and understanding of predictions.

Deployment of this framework as a mobile or web application, integrated with healthcare systems, can further broaden its utility and impact. Overall, this research lays a solid foundation for AI-assisted autism diagnosis, with promising implications for global healthcare accessibility and equity.

References

1. Eslami, T.; Mirjalili, V.; Fong, A.; Laird, A.R.; Saeed, F. ASD-DiagNet: A hybrid learning approach for detection of autism spectrum disorder using fMRI data. *Front. Neuroinform.* **2019**, *13*, 70.
2. Prelock, P.A. Autism Spectrum Disorders. *Handb. Lang. Speech Disord.* **2021**, 129–151.
3. Klin, A.; Mercadante, M.T. Autism and the pervasive developmental disorders. *Rev. Bras. de Psiquiatr.* **2006**, *28*, S1–S2.
4. Russell, A.J.; Murphy, C.M.; Wilson, E.; Gillan, N.; Brown, C.; Robertson, D.M.; Murphy, D.G. The mental health of individuals referred for assessment of autism spectrum disorder in adulthood: A clinic report. *Autism* **2016**, *20*, 623–627.
5. Dawson, G. Early behavioral intervention, brain plasticity, and the prevention of autism spectrum disorder. *Dev. Psychopathol.* **2008**, *20*, 775–803.

6. Loth, E.; Charman, T.; Mason, L.; Tillmann, J.; Jones, E.J.; Wooldridge, C.; Buitelaar, J.K. The EU-AIMS Longitudinal European Autism Project (LEAP): Design and methodologies to identify and validate stratification biomarkers for autism spectrum disorders. *Mol. Autism* **2017**, *8*, 1–19.
7. Kwon, M.K.; Moore, A.; Barnes, C.C.; Cha, D.; Pierce, K. Typical levels of eye-region fixation in toddlers with autism spectrum disorder across multiple contexts. *J. Am. Acad. Child Adolesc. Psychiatry* **2019**, *58*, 1004–1015.
8. Constantino, J.N.; Kennon-McGill, S.; Weichselbaum, C.; Marrus, N.; Haider, A.; Glowinski, A.L.; Jones, W. Infant viewing of social scenes is under genetic control and is atypical in autism. *Nature* **2017**, *547*, 340–344.
9. Gredebäck, G.; Johnson, S.; von Hofsten, C. Eye tracking in infancy research. *Dev. Neuropsychol.* **2010**, *35*, 340–344.
10. Falck-Ytter, T.; Nystrom, P.; Gredeback, G.; Gliga, T.; Bolte, S. Reduced orienting to audiovisual synchrony in infancy predicts autism diagnosis at 3 years of age. *J. Child Psychol. Psychiatry* **2018**, *59*, 872–880.
11. Guillon, Q.; Hadjikhani, N.; Baduel, S.; Roge, B. Visual social attention in autism spectrum disorder: Insights from eye tracking studies. *Neurosci. Biobehav. Rev.* **2014**, *42*, 279–297.
12. Lord, C.; Risi, S.; DiLavore, P.S.; Shulman, C.; Thurm, A.; Pickles, A. Autism from 2 to 9 years of age. *Arch. Gen. Psychiatry* **2006**, *63*, 694–701.
13. Chlebowski, C.; Green, J.A.; Barton, M.L.; Fein, D. Using the childhood autism rating scale to diagnose autism spectrum disorders. *J. Autism Dev. Disord.* **2010**, *40*, 787–799.
14. Thorup, E.; Nystrom, P.; Gredeback, G.; Bolte, S.; Falck-Ytter, T. Altered gaze following during live interaction in infants at risk for autism: An eye tracking study. *Mol. Autism* **2016**, *7*, 1–10.