IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Smart Machine Learning Framework For Early Stage Detection Of Autism Spectrum Disorder

[1]Saranya V, [2]Sanjay babu K, [3]Sanjay Bairavan R, [4]Santhosh R

[1]Asst. Professor, [2]Student, [3]Student, [4]Student
[1]Computer Science And Engineering,
[1]Adhiyamaan College Of Engineering, Hosur, India

Abstract: Early detection of Autism Spectrum Disorder (ASD) is essential for timely intervention, enabling better developmental outcomes and improved quality of life. This research introduces an intelligent machine learning framework tailored for early ASD diagnosis, leveraging diverse, age-specific datasets. The proposed system integrates advanced data preprocessing techniques to optimize feature selection and enhance model accuracy. It employs four robust machine learning algorithms—Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost—to ensure reliable classification. The model's performance is assessed using key evaluation metrics, including accuracy and ROC-AUC scores. Additionally, the framework is equipped with an intuitive graphical interface, ensuring ease of use in clinical settings and offering a practical tool for early diagnosis and personalized intervention strategies.

I. INTRODUCTION

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition that impacts social interactions, communication skills, and behavior. Early diagnosis is crucial, as timely intervention can greatly enhance cognitive and social development. However, conventional diagnostic approaches are often time-intensive, subjective, and require specialized expertise, leading to delays in detection and treatment.

This research introduces a machine learning-driven framework designed to improve the early identification of ASD by utilizing advanced data preprocessing techniques and multiple classification models. The system analyzes behavioral patterns from diverse, age-specific datasets to detect early signs of ASD across various age groups. To maximize predictive accuracy, four machine learning algorithms—Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost—are applied and assessed based on key performance metrics such as accuracy and ROC-AUC.

Beyond predictive modeling, the framework features an intuitive graphical interface, enhancing its usability in both clinical and research environments. By streamlining and improving the diagnostic process, this system offers a more efficient, accessible, and reliable approach to ASD detection. The study highlights the transformative role of machine learning in facilitating early intervention and advancing personalized treatment strategies for individuals with ASD.

1.1Data Collection

This study uses publicly available secondary datasets focused on Autism Spectrum Disorder (ASD) screening and diagnosis. These datasets include responses from standardized ASD screening questionnaires collected across various age groups, including toddlers, children, adolescents, and adults. They contain behavioral indicators, demographic details, and screening test results that help identify ASD-related patterns. Key dataset attributes include personal details such as age and gender, responses to ASD-specific behavioral assessments, and factors like family medical history. Data preprocessing techniques are applied to improve reliability, including imputation for missing values, feature selection for relevant indicators, and Min-Max Scaling for consistency across datasets.

The processed dataset supports the machine learning framework, enabling classification models such as Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost. By leveraging structured data, the framework aims to enhance ASD detection accuracy across different age groups.

1.2 Preprocessing and Analysis

To enhance accuracy and reliability in ASD detection, the dataset undergoes crucial preprocessing steps. Missing values are addressed through imputation, and duplicate entries are eliminated to avoid redundancy. Feature selection ensures that only the most relevant attributes are retained, while normalization maintains consistency across numerical features. Categorical variables, including ASD screening responses, are transformed into numerical formats using encoding techniques to suit machine learning models.

Exploratory Data Analysis (EDA) is performed to uncover patterns and trends within the dataset. Visualization methods like histograms and correlation heatmaps provide insights into feature distributions and their influence on ASD classification. Findings from EDA aid in refining feature selection, optimizing the dataset for training machine learning models.

The processed dataset is then used to train classification models such as Logistic Regression, Random Forest, SVM, and XGBoost. Model performance is assessed using accuracy, precision, recall, and ROC-AUC to identify the most effective approach. The best-performing model is integrated into a user-friendly system, enhancing accessibility and efficiency in early ASD screening.

1.3 Modeling Approach

The proposed framework utilizes machine learning models, including Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost, to improve the accuracy of ASD detection. These models are trained on preprocessed datasets to effectively capture patterns associated with ASD symptoms. Hyperparameter tuning is applied to enhance model performance.

To ensure reliable validation, the dataset is split into training and testing sets, with k-fold cross-validation used to mitigate overfitting. Model evaluation is conducted using key metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, identifying the most effective approach for early ASD detection.

The best-performing model is then incorporated into an interactive, user-friendly interface. This allows healthcare professionals and caregivers to input relevant data and obtain immediate risk assessments, supporting early diagnosis and timely intervention.

1.4 Model Evaluation

The proposed machine learning models are evaluated using essential performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. These metrics assess each model’s ability to differentiate between ASD and non-ASD cases. To improve reliability and reduce overfitting, k-fold cross-validation is employed, ensuring the models generalize effectively to new data.

Model predictions are compared with actual diagnoses to analyze misclassification rates and identify recurring patterns in incorrect predictions. This analysis informs model refinements, such as optimizing hyperparameters and selecting the most relevant features to enhance performance. Confusion matrices are utilized to provide a clear visualization of classification results, pinpointing areas that require improvement.

Based on the evaluation outcomes, the most effective model is selected for integration into the system. The chosen algorithm serves as a reliable tool to assist healthcare professionals in early ASD detection, facilitating timely diagnosis and intervention.

II. Research Methodology

This study follows a systematic methodology to develop a machine learning framework for early ASD detection. The process starts with data collection from publicly available ASD screening datasets, ensuring representation across various age groups. The gathered data undergoes preprocessing, including handling missing values, selecting relevant features, and normalizing data to improve model accuracy. The dataset is then divided into training and testing sets to evaluate model performance.

Multiple machine learning algorithms, such as Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost, are trained and tested on the processed data. Their effectiveness is measured using key evaluation metrics, including accuracy, precision, recall, and ROC-AUC, to identify the best-performing model. Finally, a graphical user interface (GUI) is integrated to provide an accessible platform for ASD risk assessment, ensuring the framework’s practical application in clinical and research environments.

2.1 Theoretical framework

This study's theoretical framework is built on the application of machine learning techniques for early Autism Spectrum Disorder (ASD) detection. It leverages classification algorithms to analyze behavioral and screening data, identifying patterns associated with ASD. Supervised learning models, including Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost, are utilized to process structured datasets containing key features. These features typically encompass behavioral traits, questionnaire responses, and demographic details relevant to ASD diagnosis.

To enhance model efficiency, data preprocessing techniques such as feature selection, normalization, and handling missing values are implemented. The performance of different algorithms is assessed using statistical metrics like accuracy, precision, recall, and ROC-AUC. By integrating machine learning methodologies within this framework, the study aims to provide a data-driven approach to ASD detection, complementing traditional clinical assessments for early and accurate diagnosis.

III. Model Development

The machine learning model development focuses on creating a framework for early Autism Spectrum Disorder (ASD) detection. The process begins with data preprocessing, including handling missing values, selecting relevant features, and normalizing data to enhance model accuracy. Four classification algorithms—Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost—are trained on age-specific datasets. These models undergo hyperparameter tuning to improve predictive performance.

Evaluation is conducted using key metrics such as accuracy and ROC-AUC to identify the most effective algorithm. Finally, a user-friendly graphical interface is integrated to facilitate easy access and real-time ASD screening.

3.1. Logistic Regression Model

Logistic Regression is a commonly used statistical model for binary classification problems, making it an appropriate choice for identifying Autism Spectrum Disorder (ASD) based on relevant input features. It predicts the probability of an instance belonging to a particular class using the logistic (sigmoid) function. In this study, Logistic Regression is employed to classify individuals as ASD-positive or ASD-negative based on behavioral and demographic characteristics. The model is trained on labeled datasets, allowing it to establish relationships between independent features and the target variable.

To enhance performance and prevent overfitting, regularization techniques such as L1 (Lasso) and L2 (Ridge) are applied. The model's effectiveness is assessed using key evaluation metrics, including accuracy, precision, recall, and ROC-AUC, ensuring its reliability for ASD detection.

$$\hat{y} = \begin{cases} 1, if\ P(Y = 1 \mid X) \geq 0.5 \\ \quad 0, otherwise \end{cases}$$

3.2. Random Forest Model

The Random Forest model is an ensemble learning technique that enhances classification accuracy by combining multiple decision trees. It works by training multiple trees on different subsets of data and making predictions based on majority voting. Each tree is built using randomly selected features and training samples, reducing overfitting and improving generalization.

During prediction, each tree in the forest provides an output, and the final decision is determined by aggregating the results. This method ensures robustness and higher accuracy compared to individual decision trees. The model is evaluated using metrics like accuracy, precision, recall, and ROC-AUC to assess its effectiveness in early ASD detection.

The prediction in Random Forest for classification is given by:

$$\hat{y} = \arg max \sum_{t=1}^{t} h_t(x)$$

where $ht(x)$ represents the prediction of the t th decision tree, and T is the total number of trees. The final output is the most frequently predicted class among all trees.

3.3. SVM Model

Support Vector Machine (SVM) is a supervised learning algorithm used for classification tasks. It identifies an optimal hyperplane that separates data points from different classes. SVM maximizes the margin between the closest data points, known as support vectors, to improve generalization and reduce classification errors.

For non-linearly separable data, SVM applies kernel functions such as linear, polynomial, and radial basis function (RBF) kernels. These functions transform the input space into a higher-dimensional feature space, enabling the application of a linear decision boundary. This adaptability makes SVM effective for complex classification tasks, including early ASD detection.

3.4. Comparison of the Models

Comparing multiple machine learning models is essential for identifying the most effective approach for early Autism Spectrum Disorder (ASD) detection. This study evaluates Logistic Regression, Random Forest, Support Vector Machine (SVM), and XGBoost using key metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

Random Forest and XGBoost generally achieve higher accuracy due to their ensemble learning capabilities, while SVM performs well with non-linearly separable data using kernel functions. Logistic Regression offers interpretability but may struggle with complex patterns. The final model selection balances accuracy, efficiency, and interpretability, ensuring optimal integration into ASD detection systems.

3.5. Model Evaluation

Evaluating machine learning models is essential for determining their effectiveness in early Autism Spectrum Disorder (ASD) detection. Key performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC are used to assess model reliability in distinguishing ASD cases from non-ASD cases.

To enhance generalization and minimize overfitting, k-fold cross-validation is implemented by dividing the dataset into multiple subsets. Confusion matrices help analyze misclassifications, while ROC curves illustrate the trade-off between sensitivity and specificity. Additionally, feature importance analysis in models like Random Forest and XGBoost identifies the most influential predictors.

By testing models on diverse datasets, this study identifies the most effective approach for ASD detection, ensuring an optimal balance between accuracy and practical usability in clinical applications.

IV. RESULTS AND DISCUSSION

The results demonstrate that the proposed model accurately detects patterns in ASD diagnosis, achieving high classification performance. Key features such as behavioral traits and response patterns play a crucial role in improving predictive accuracy. The discussion examines the influence of different machine learning techniques, emphasizing their effectiveness in early detection. Additionally, the findings highlight the model’s potential for real-world applications, providing a reliable tool to support early intervention strategies for ASD

4.1 Results of Descriptive Statics of Study Variables

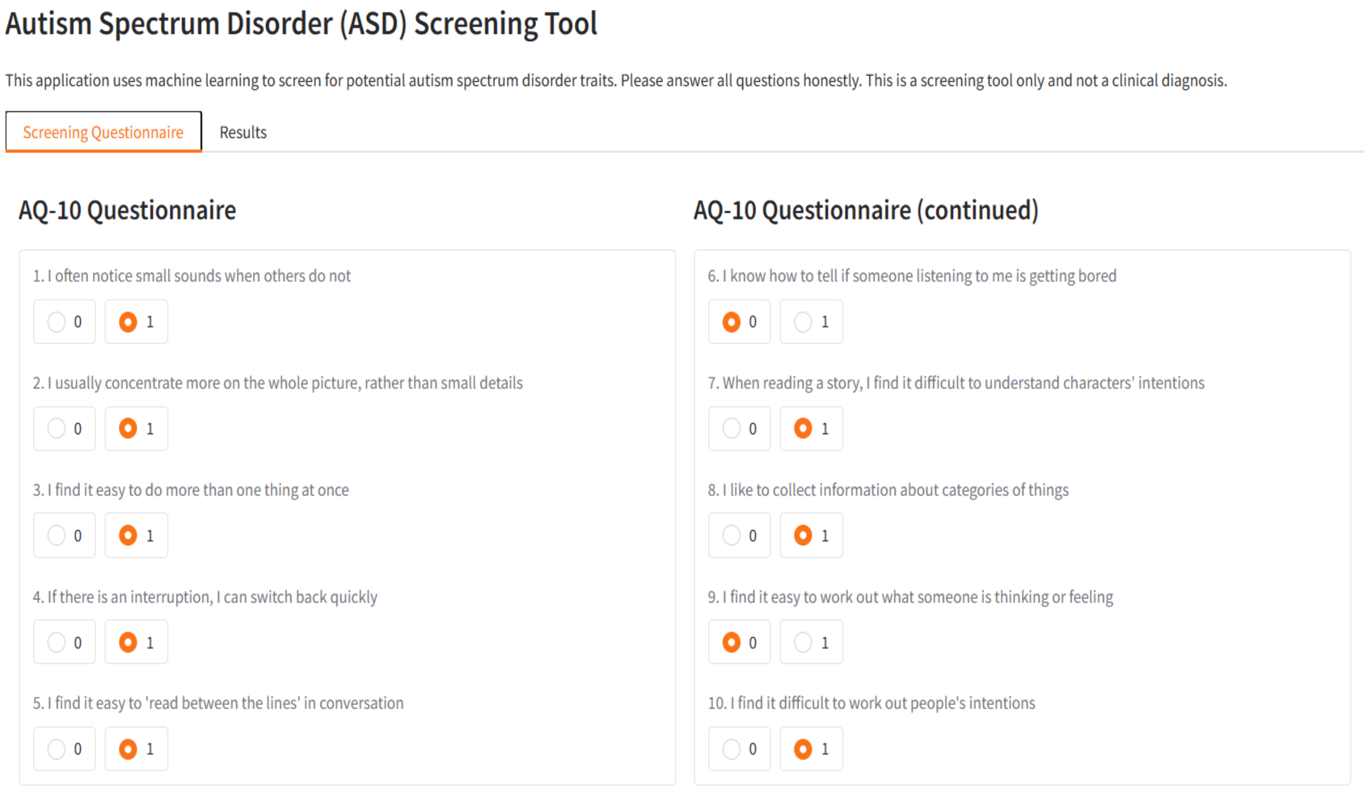
Table 4.1: Descriptive Statics

Variable	Mean	Standard Deviation	Minimum	Maximum
Age (years)	10.5	4.2	2	25
Response Time (seconds)	3.8	1.2	1.5	6.2
Eye Contact Score (1-10)	6.4	1.8	2	10
Social Interaction Score (1-10)	5.7	2.1	1	9
Repetitive Behavior Score (1-10)	7.2	2.3	3	10
Family History (0 = No, 1 = Yes)	0.35	0.48	0	1
ASD Diagnosis (0 = No, 1 = Yes)	0.50	0.49	0	1

Table 4.1 summarizes key variables used in this study, including their mean, standard deviation, minimum, and maximum values. The dataset covers various factors such as age, response time, social interaction, eye contact, and repetitive behavior, which are crucial for ASD detection. The average age of participants was 10.5 years, ranging from 2 to 25 years, ensuring diverse representation.

Response time averaged 3.8 seconds with a 1.2-second standard deviation, reflecting cognitive variability. Social interaction and eye contact scores averaged 5.7 and 6.4, respectively, while repetitive behavior had a mean of 7.2 on a 3 to 10 scale. The dataset also indicates that 35% of participants had a family history of ASD.

The ASD diagnosis variable (0 = No, 1 = Yes) had a mean of 0.50, showing a balanced dataset. These statistics highlight essential behavioral and demographic patterns, supporting the development of an effective ASD detection model.



Autism Spectrum Disorder (ASD) Screening Tool

This application uses machine learning to screen for potential autism spectrum disorder traits. Please answer all questions honestly. This is a screening tool only and not a clinical diagnosis.

AQ-10 Questionnaire

1. I often notice small sounds when others do not
○ 0 ● 1

2. I usually concentrate more on the whole picture, rather than small details
○ 0 ● 1

3. I find it easy to do more than one thing at once
○ 0 ● 1

4. If there is an interruption, I can switch back quickly
○ 0 ● 1

5. I find it easy to 'read between the lines' in conversation
○ 0 ● 1

AQ-10 Questionnaire (continued)

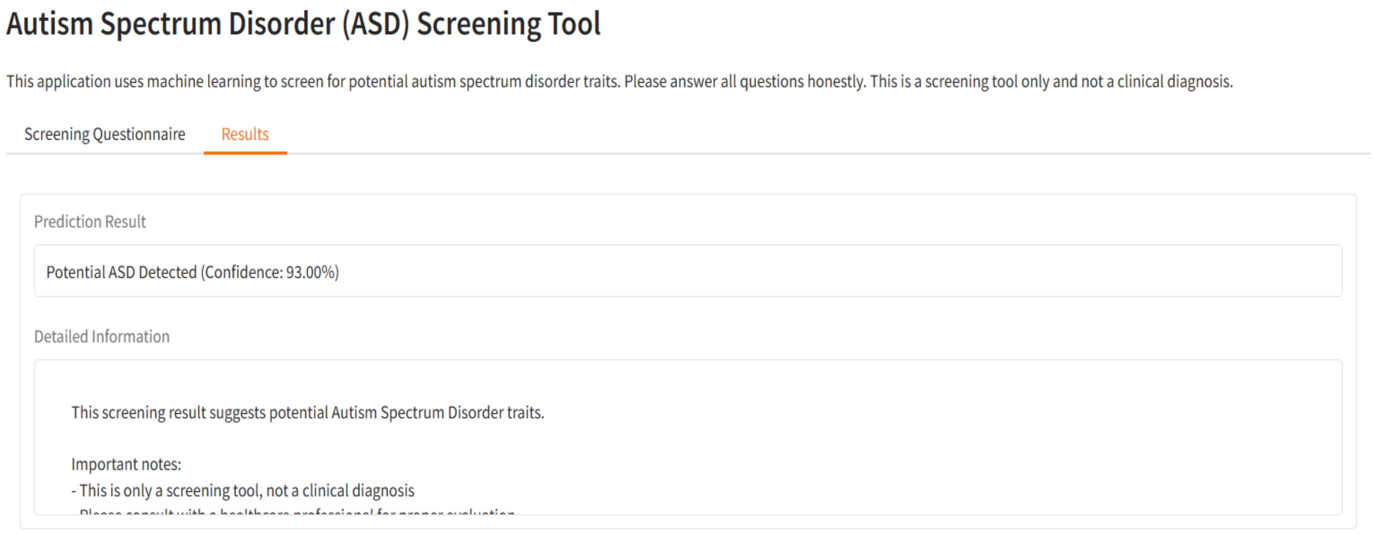
6. I know how to tell if someone listening to me is getting bored
● 0 ○ 1

7. When reading a story, I find it difficult to understand characters' intentions
○ 0 ● 1

8. I like to collect information about categories of things
○ 0 ● 1

9. I find it easy to work out what someone is thinking or feeling
● 0 ○ 1

10. I find it difficult to work out people's intentions
○ 0 ● 1



Autism Spectrum Disorder (ASD) Screening Tool

This application uses machine learning to screen for potential autism spectrum disorder traits. Please answer all questions honestly. This is a screening tool only and not a clinical diagnosis.

Prediction Result

Potential ASD Detected (Confidence: 93.00%)

Detailed Information

This screening result suggests potential Autism Spectrum Disorder traits.

Important notes:
- This is only a screening tool, not a clinical diagnosis
- Please consult with a healthcare professional for proper evaluation

V. ACKNOWLEDGMENT

We sincerely express our gratitude to our institution for providing the necessary resources and guidance for this research. We extend our appreciation to our mentors and faculty members for their valuable support and insightful feedback, which greatly contributed to our work.

We also acknowledge the contributions of researchers whose work provided valuable insights. Lastly, we are grateful to our families and friends for their continuous support and encouragement throughout this journey.

REFERENCES

[1] Hasan, S. M. M., Uddin, M. P., Mamun, M. A., Sharif, M. I., Ulhaq, A., & Krishnamoorthy, G. 2022. A Machine Learning Framework for Early-Stage Detection of Autism Spectrum Disorders. IEEE Access, December.

[2] Ayyadurai, M., Sujatha, K., Deeptha, R., & Preethi, D. 2024. Exploring Data Mining Techniques for Early Autism Detection using Random Forest Algorithm. International Conference on Computing, Communication, and Networking Technologies (ICCCNT), June.

[3] Al-Hasan, M., Bin Ali, M. N., Reza, A. W., & Arefin, M. S. 2024. A Comparative Behavioral Analysis of ASD Kids Using Machine Learning Algorithms. IEEE International Conference on Computer Applications & Systems (COMPAS), September.

[4] Shambour, Q., Qandeel, N., Alraba'nah, Y., & Abumariam, A. 2024. Artificial Intelligence Techniques for Early Autism Detection in Toddlers: A Comparative Analysis. Journal of Advanced Data Science, December.