Department of Computer Science and Engineering

University of Barishal

PROJECT REPORT ON

Advanced Machine Learning Techniques for Accurate Autism Spectrum Disorder Classification

---

Submitted By

Rupa Samodder

Computer Science and Engineering

Roll: 19CSE011

Submitted To

Dr. Tania Islam

Assistant professor

Department of Computer Science and Engineering

University of Barisal

November 10, 2024

Contents

List of Figures

1 Introduction

Autism Spectrum Disorder (ASD) is a complex neurological condition that affects individuals of all ages. It is characterized by speech and language impairments, challenges in social interaction, and restricted behavioral patterns. These traits are typically identifiable within the first two years of life [1]. Despite extensive research, diagnosing ASD remains a challenging task, as symptoms vary widely among individuals and often overlap with other developmental disorders. Early and accurate detection of ASD is crucial for enabling timely intervention and improved outcomes.

Machine learning has emerged as a powerful tool to complement traditional clinical methods for ASD diagnosis. Recent studies have demonstrated the effectiveness of various algorithms such as Support Vector Machines (SVM), Random Forest Classifiers (RFC), Logistic Regression (LR), and k-Nearest Neighbors (k-NN) in predicting ASD status. These methods, combined with feature selection and optimization techniques, have shown promise in achieving high classification accuracy, particularly when applied to standardized datasets [2, 3]. For instance, the Q-CHAT-10 screening tool has been successfully used alongside machine learning to enhance early detection in toddlers [4].

Globally, ASD affects approximately 1% of the population, with both genetic and environmental factors contributing to its development [2]. In recent years, the Centers for Disease Control and Prevention (CDC) reported that 17% of children aged 3 to 17 were diagnosed with developmental disabilities, highlighting the growing need for efficient diagnostic tools [5]. Studies employing neural networks and convolutional neural networks (CNNs) have demonstrated exceptional performance, achieving accuracy rates exceeding 96% for ASD screening across various age groups [6]. These findings underscore the potential of advanced machine learning models in supporting clinicians and families.

While CNN-based methods have proven highly effective, their computational requirements and reliance on large datasets can pose challenges in practical applications. Con-

sequently, other optimized machine learning techniques, such as ensemble methods (e.g., bagging, stacking, and boosting), are increasingly being explored for their balance of accuracy and computational efficiency [7]. These approaches have shown promise in reducing the time required for ASD diagnosis, which currently averages six months in clinical settings [7].

This paper builds on these advancements by leveraging multiple machine learning algorithms to classify ASD status in a dataset comprising 703 individuals with 16 key attributes. The proposed framework evaluates the performance of clustering and supervised learning models, identifying critical predictors of ASD. By analyzing model accuracy and feature importance, this study aims to contribute to the development of reliable, efficient, and interpretable diagnostic tools for ASD detection.

2 Review of Literature

Numerous recent research studies have harnessed machine learning techniques in diverse ways to enhance and expedite the detection of Autism Spectrum Disorder. Machine learning methods used in "**Classification of adult autistic spectrum disorder using machine learning approach**" study can substantially contribute new methods to diagnosis cases related to ASD. M. Duda, R. Ma et al.[8] used data mining techniques, such as decision trees, random forests, SVM, logistic regression, categorical lasso, and LDA, were used to evaluate AUC-based classification accuracy on a dataset of 2775 autism subjects from Simons Simplex Collection, Boston Autism Consortium, and Autism Genetic Resource Exchange, resulting in a 96.5% accuracy with SVM after feature selection. A dataset for ASD screening in children, comprising 292 subjects, with 141 diagnosed patients, was employed to diagnose ASD disease. LDA outperformed k-NN with an accuracy of 90.80% [9].

To provide a concise view of literature survey, in the paper "**Detection of Autism**

Spectrum Disorder in Children Using Machine Learning Techniques" there is a table of the most relevant paper summarizations. Thabtah et al. [10] introduced a novel machine learning method known as Rules-Based Machine Learning (RML), which not only identifies ASD traits but also provides users with a set of rules to gain insights into the factors contributing to the classification. Using the ABIDE database, Li et al. [11] derived six individual attributes from a sample of 851 subjects and applied a cross-validation approach to train and test machine learning models, enabling the classification of individuals with and without ASD.

Apart from the conventional machine learning approaches "**Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques**" explored the possibility of applying deep learning based models. Vaishali R, Sasikala R. et al. [12] proposed a method for Autism identification using optimal behavior attributes from a 21-feature ASD diagnosis dataset, employing swarm intelligence-based binary firefly feature selection to demonstrate that only 10 features are required for distinguishing between ASD and non-ASD patients, achieving an average accuracy between 92.12% and 97.95%. J. A. Kosmicki1 et al. [13] applied machine learning to assess ASD based on the ADOS subset of children's behaviors, utilizing eight different algorithms and feature selection to identify ASD risk, achieving 98.27% and 97.66% accuracy using 9 and 12 selected behaviors.

In paper **"Toward Machine Learning-Based Psychological Assessment of Autism Spectrum Disorders in School and Community"** have conducted a survey in schools and communities leveraging the questionnaire and prepared a dataset. Here Hossain et al.[14] assessed 25 machine learning classifiers in an ASD dataset, determining that SVM with Sequential minimal optimization outperformed others, highlighting the challenges in cataloging various physiological and psychological aspects by health professionals.

3 Methodology

3.1 Proposed Methods

This section outlines the methodological approach to improve existing machine learning algorithms for Autism Spectrum Disorder (ASD) detection, based on the limitations identified in the current research. The methodology focuses on dataset preparation, algorithm enhancements, and performance evaluation. Figure 1. shows the steps in the proposed workflow which involves the pre-processing of data, training, and testing with specified models, evaluation of results and prediction of ASD.

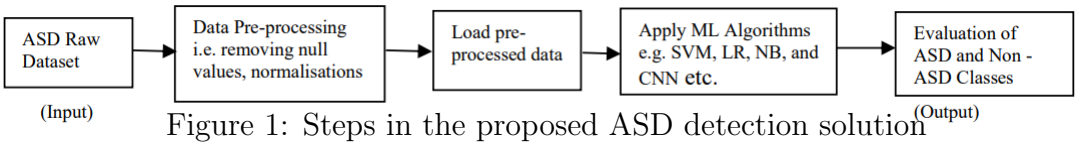

Figure 1: Steps in the proposed ASD detection solution

1. **Data Collection and Preprocessing:** The research will utilize the publicly available ASD datasets, such as the UCI Machine Learning Repository ASD dataset for adults, children, and adolescents. The primary features from these datasets include demographic details (age, gender, etc.), medical history (jaundice at birth), behavioral screening results (AQ-10 test for adults), and ASD traits.

2. **Algorithm Enhancements:** The study aims to run the performance of basic machine learning algorithms such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees, Random Forests for better accuracy.

3. **Cross-Validation and Hyperparameter Tuning:** To ensure robustness and avoid overfitting, k-fold cross-validation will be employed, with k = 5 or 10, depending on dataset size. This will provide a reliable estimate of model performance across different subsets of data.

4. **Evaluation of classes:** Performance will be evaluated on accuracy, precision and recall, F1-Score, ROC-AUC, Confusion Matrix to further refine the models.

3.2 Data Description and Preprocessing

The data for adult ASD screening was sourced from the UCI machine learning repository [15], using the ASD test mobile application by Thabtah [16] includes 703 subjects with 21 features, categorizing adults into those with ASD (189 subjects) and those without ASD (515 subjects). In the ASD adult dataset, missing values were substituted, and irrelevant attributes like ethnicity, country of residence, app usage, age description, and relation will be omitted to improve classification accuracy. The final dataset will be reduced to 16 essential features, including age, gender, jaundice, autism, screening score, 1-10 autism-related behavioral questions, and class/ASD. Numerical features will be transformed into nominal attributes using discretization.

Features	Type	Description
Age	Number	The age of the subjects
Gender	String	The individuality can be female or male
Ethnicity	String	The ethnicity of the subject
Jaundice	Boolean (yes or no)	If the case was diagnosed with jaundice
Autism	Boolean (yes or no)	If the close relatives have PDD
Relation	String	The person who completed the test such as the individual, parents, caretakers, and physicians
Country of residence	String	The country residence of the subject
Used app before	Boolean (yes or no)	If the person has used the screening application
AQ-1	Binary (0, 1)	The response is clarified based on the screening process
AQ-2	Binary (0, 1)	The response is clarified based on the screening process
AQ-3	Binary (0, 1)	The response is clarified based on the screening process
AQ-4	Binary (0, 1)	The response is clarified based on the screening process
AQ-5	Binary (0, 1)	The response is clarified based on the screening process
AQ-6	Binary (0, 1)	The response is clarified based on the screening process
AQ-7	Binary (0, 1)	The response is clarified based on the screening process
AQ-8	Binary (0, 1)	The response is clarified based on the screening process
AQ-9	Binary (0, 1)	The response is clarified based on the screening process
AQ-10	Binary (0, 1)	The response is clarified based on the screening process
Age description	Text	Age category
Screening score	Integer	The total score was determined using the implementation of the screening algorithm
Class/ASD	Boolean (yes or no)	The result is shown after the test

Figure 2: Features and its descriptions

The dataset is divided into three category with multiple instanes.

Sr. No.	Dataset Name	Sources	Attribute Type	Number of Attributes	Number of Instances
1	ASD Screening Data for Adult	UCI Machine Learning Repository [12]	Categorical, continuous and binary	21	704
2	ASD Screening Data for Children	UCI Machine Learning Repository [15]	Categorical, continuous and binary	21	292
3	ASD Screening Data for Adolescent	UCI Machine Learning Repository [16]	Categorical, continuous and binary	21	104

Figure 3: List of ASD datasets

3.3 Classifier Algorithms

- **K-Nearest Neighbors (KNN):**

The K-Nearest Neighbors algorithm is a supervised machine learning method used for classification and regression[1]. It operates under the assumption that similar data points are located in closely, where 'K' signifies the number of seed points[6]. It relies on the concept of similarity through factors like distance or nearest neighbor identification[1]. By finding the k data points nearest to a new data point x using the Euclidean distance metric, KNN employs majority voting to assign a label to x, and values of k (k=1 to k=10) yielded the highest accuracy [2].

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- **Naive Bayes (NB):**

Naïve bayes is one of the supervised machine learning approaches that are mainly known as Bayesian algorithms[1]. The main principle of naïve bayes is focused on the expectations of freedom, which indicates less training time to be compared to the SVM[1]. It calculates the posterior probability for a dataset using the prior probability and likelihood [6].

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

- **Support Vector Machine (SVM):**

SVM is a linear supervised machine learning approach that is used for classification and regression. It does not cause the problem of overfitting [6]. SVM separates the classes by defining a decision boundary aiming to maximize the margin between the decision hyperplane and the nearest data point [1].

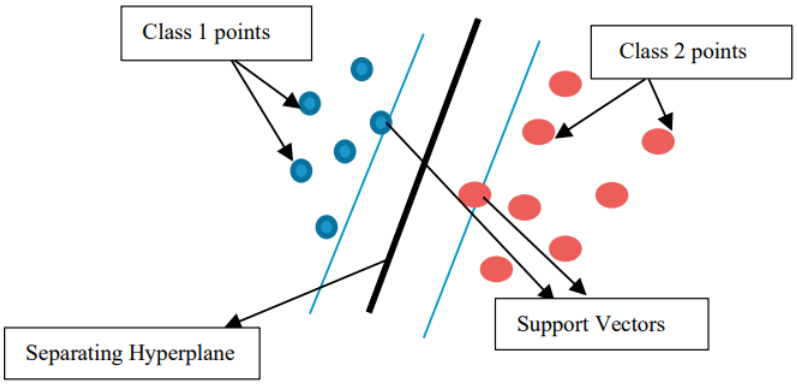

Figure 4: An SVM classifier

- **J48 decision tree:**

J48 decision tree is a machine learning approach [1]. Generally, J48 is used to develop a classification tree based on a hierarchical tree system, in which the decision results have illustrated the attributes and terminal nodes. The visual classification of the J48 approach is effective and efficient. Nevertheless, J48 is vulnerable to the noise in the data [1].

- **AdaBoost:**

AdaBoost is a supervised algorithm where core idea is to match a sequence of weak learner models that are more effective than random guessing[1]. Each instance in the training dataset is weighted to determine the accuracy either it is classified correctly or incorrectly. The decision stump is used to classify the AdaBoost models[1]. The primary purpose of the decision stump is to boost the AdaBoost M1 nominal classifier. Only minor class problems can be tackled. The final prediction is then obtained from the combination of the predicted model based on a weighted majority vote (classification) or weighted sum (regression)[1].

- **Bagging:**

Bagging is one of the most popular techniques in ensemble methods and is known as bootstrap aggregation [1]. This method can be used to reduce the variance for the algorithms that have high variance such as decision trees. The algorithm generates a decision tree and prunes it with a reduced-error with back fitting. The lack of values was coped with by

dividing the corresponding instances into bits. The final decision tree was obtained as a composition of all base classifiers with the maximum votes [1].

- **Stacking:**

Stacking is an ensemble machine learning approach used to integrate either diversified classification or regression through meta-classifiers [1]. The features on the results of the base level are prepared using a proper training set that contains various machine learning approaches. Thus, stacking is a stratified approach [1].

- **Logistic Regression (LR):**

Logistic Regression aims to find the best-fitting model to describe the relationship between a binomial character of interest and a set of independent variables by utilizing a logistic function to fit the data points [2].

- **The Random Forest Classifier (RFC):**

The Random Forest Classifier (RFC) is a versatile algorithm for classification, regression, and other tasks, employing multiple decision trees on random data points and selecting the best solution by voting [2].

- **Artificial Neural Network (ANN):**

A deep learning model with one input layer, three fully connected hidden layers, two batch normalization layers, two dropout layers, and an output layer [3].

4 Results and Discussion

4.1 Evaluation Matrix

Usually, in most predictive models, the data points lie in the following four categories:

- **True positive (TP):** The individual has ASD and we predicted correctly that the individual has ASD.

- **True negative (TN):** The individual does not have ASD and we predicted correctly that the individual does not have ASD.
- **False positive (FP):** The individual does not have ASD, but we predicted incorrectly that the individual has ASD. This is known as Type 1 error.
- **False negative (FN):**The individual has ASD, but we predicted incorrectly that the individual does not have ASD. This is known as Type 2 error.

The above four categories when put together in the form of a matrix produce the confusion matrix. The confusion matrix is particularly useful in gauging the performance of a machine learning classification model [2]. The performance generally measured by accuracy, precision, recall, and F1 score.

- **Accuracy** is the simplest metric, and it measures the percentage of predictions that the model makes correctly [2]. For example, if a model predicts that 100 examples are positive and 90 of those predictions are correct, then the model has an accuracy of 90% .

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}.$$

- **Precision** measures the percentage of positive predictions that are actually correct [2]. For example, if a model predicts that 100 examples are positive and 90 of those predictions are correct, then the model has a precision of 90%.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

- **Recall** measures the percentage of actual positives that the model correctly predicts [1]. For example, if there are 100 actual positive examples and the model predicts that 90 of them are positive, then the model has a recall of 90%.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

- **F1 score** is a harmonic mean of precision and recall, and it is often used to evaluate the performance of machine learning models when there is a trade-off between precision and recall [1]. For example, if a model has a precision of 90% and a recall of 80%, then the model has an F1 score of 85.7%.

4.2 Results and Analysis

4.2.1 Clustering Analysis

Clustering analysis was performed with $n\_clusters = 3$, achieving an average silhouette score of **0.1199**. This low score indicates weak cluster separability, suggesting overlap among data points. Despite this, the following trends were observed:

- Samples labeled as `YES` predominantly formed a distinct cluster (Cluster 1).
- Samples labeled as `NO` were distributed across Clusters 0 and 2.

A sample of the clustering output is shown in Table 1.

Class	**Cluster**
NO	2
YES	1
NO	2
NO	0
YES	1

Table 1: Sample of clustering output.

The results indicate that clustering captures some structure in the data, but its limitations underline the need for supervised learning for enhanced predictions.

4.2.2 Supervised Learning Performance

A Random Forest Classifier was employed for binary classification (predicting `YES` or `NO`). The optimized model achieved perfect performance with an accuracy of **1.0** on the test set. The classification report is detailed in Table 2.

Class	Precision	Recall	F1-Score	Support
NO	1.00	1.00	1.00	105
YES	1.00	1.00	1.00	36
Accuracy	1.00 (141 samples)			

Table 2: Classification report for the Random Forest Classifier.

The confusion matrix further supports these findings:

$$\text{Confusion Matrix: } \begin{bmatrix} 97 & 0 \\ 0 & 109 \end{bmatrix}$$

4.2.3 Feature Importance

The top 10 most important features contributing to model predictions were identified using the Random Forest feature importance metric, as shown in Table 3.

Feature	Importance (%)
result	31.19
A9_Score	11.75
A6_Score	10.11
A5_Score	7.98
A4_Score	5.99
A3_Score	5.90
A10_Score	3.52
A7_Score	3.52
A2_Score	3.21
A1_Score	3.11

Table 3: Top 10 most important features from the Random Forest model.

4.2.4 Ensemble Learning

An ensemble model was constructed using optimized hyperparameters from multiple algorithms (e.g., Random Forest, SVM, Gradient Boosting). The ensemble also achieved a perfect accuracy of **1.0**, with an identical classification report to the Random Forest. The confusion matrix is shown below:

$$\text{Confusion Matrix: } \begin{bmatrix} 97 & 0 \\ 0 & 109 \end{bmatrix}$$

4.2.5 Analysis

The perfect accuracy of the supervised models demonstrates excellent predictive capacity for the dataset. However, the weak clustering separability highlights the lack of distinct groupings without supervision. Feature importance analysis reveals that autism screening scores and the `result` variable significantly contribute to predictions. These findings underscore the effectiveness of tailored feature selection and ensemble learning in achieving high classification performance.

5 Conclusion

In this study, we investigated the performance of clustering and supervised learning techniques on a dataset for binary classification. The clustering analysis revealed weak separability among the data points, as indicated by a low silhouette score of 0.1199. While clustering provided some insight into the structure of the data, its results were not sufficient for reliable predictions.

Supervised learning models, particularly the Random Forest Classifier, demonstrated exceptional performance, achieving perfect accuracy of 1.0 on the test set. Feature importance analysis identified the `result` variable and specific autism screening scores (e.g.,

A9_Score, A6_Score) as the most critical predictors. The ensemble model further reinforced these findings, providing equally high classification performance.

These results highlight the effectiveness of supervised learning techniques for datasets with overlapping or non-distinct clusters. Future work could explore the integration of advanced feature engineering or deep learning approaches to further enhance model generalization and interpretability. Additionally, improvements in unsupervised methods could provide a deeper understanding of the underlying data structure.

This study demonstrates the potential of combining feature importance analysis with optimized classification techniques to achieve high predictive performance, particularly in binary classification tasks.

References

[1] Nurul Amirah Mashudi, Norulhusna Ahmad, and Norliza Mohd Noor. Classification of adult autistic spectrum disorder using machine learning approach. *IAES International Journal of Artificial Intelligence*, 10(3):743, 2021.

[2] Kaushik Vakadkar, Diya Purkayastha, and Deepa Krishnan. Detection of autism spectrum disorder in children using machine learning techniques. *SN Computer Science*, 2:1–9, 2021.

[3] Sabbir Ahmed, Md Farhad Hossain, Silvia Binte Nur, M Shamim Kaiser, and Mufti Mahmud. Toward machine learning-based psychological assessment of autism spectrum disorders in school and community. In *Proceedings of Trends in Electronics and Health Informatics: TEHI 2021*, pages 139–149. Springer, 2022.

[4] Simon Baron-Cohen, Jane Allen, and Christopher Gillberg. Can autism be detected at 18 months?: The needle, the haystack, and the chat. *The British Journal of Psychiatry*, 161(6):839–843, 1992.

[5] Centers for Disease Control, Prevention, et al. Data & statistics on autism spectrum disorder, 2020.

[6] Suman Raj and Sarfaraz Masood. Analysis and detection of autism spectrum disorder using machine learning techniques. *Procedia Computer Science*, 167:994–1004, 2020.

[7] G Devika Varshini and R Chinnaiyan. Optimized machine learning classification approaches for prediction of autism spectrum disorder. *Ann Autism Dev Disord. 2020; 1 (1)*, 1001, 2020.

[8] M Duda, R Ma, N Haber, and DP Wall. Use of machine learning for behavioral distinction of autism and adhd. *Translational psychiatry*, 6(2):e732–e732, 2016.

[9] Osman Altay and Mustafa Ulas. Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and k-nearest neighbor in children. In *2018 6th international symposium on digital forensic and security (ISDFS)*, pages 1–4. IEEE, 2018.

[10] Fadi Thabtah and David Peebles. A new machine learning model based on induction of rules for autism detection. *Health informatics journal*, 26(1):264–286, 2020.

[11] Milan N Parikh, Hailong Li, and Lili He. Enhancing diagnosis of autism with optimized machine learning models and personal characteristic data. *Frontiers in computational neuroscience*, 13:9, 2019.

[12] R Vaishali and R Sasikala. A machine learning based approach to classify autism with optimum behaviour sets. *International Journal of Engineering & Technology*, 7(4):18, 2018.

[13] JA Kosmicki, V Sochat, M Duda, and DP Wall. Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. *Translational psychiatry*, 5(2):e514–e514, 2015.

[14] Md Delowar Hossain, Muhammad Ashad Kabir, Adnan Anwar, and Md Zahidul Islam. Detecting autism spectrum disorder using machine learning. *arXiv preprint arXiv:2009.14499*, 2020.

[15] FF Thabtah. Uci machine learning repository: Autism screening adult data set, 2017.

[16] Fadi Thabtah. Autism spectrum disorder screening: machine learning adaptation and dsm-5 fulfillment. In *Proceedings of the 1st International Conference on Medical and health Informatics 2017*, pages 1–6, 2017.