Diagnosing Autism Spectrum Disorder using Machine Learning Techniques

Sanchit Sharma
Dept. of Electronics and Communication Engineering
Nirma University, Ahmedabad, India
20bec108@nirmauni.ac.in

***Abstract*—Autism Spectrum Disorder (ASD) is a serious developmental condition that significantly impairs an individual's communication and social skills, as well as cognitive abilities. The symptoms of ASD may vary from person to person, making it a challenging disorder to diagnose accurately. To address this issue, researchers have turned to the field of machine learning, a subset of artificial intelligence, to develop prediction models that can classify individuals as either affected by ASD or not based on chronological datasets. By leveraging these models, decision-making under ambiguity can be facilitated. In this research paper, we aim to evaluate the performance of different machine learning algorithms like k-Nearest Neighbors (kNN), Naive Bayes, Decision Tree, Random Forest, Logistic Regression, and Support Vector Machines (SVM) in diagnosing ASD using a validated dataset. Our findings will provide valuable insights into the effectiveness of various machine learning techniques in diagnosing ASD, enabling healthcare professionals to make more informed decisions when it comes to diagnosing this challenging disorder.**

***Index Terms*—Introduction, Literature Review, Dataset, Methodology, Implementation, Results, Conclusion, Future Scope, Acknowledgement**

I. Introduction

Autism Spectrum Disorder (ASD) is a multifaceted developmental condition that affects the ability of individuals to communicate, interact socially, and display normal behavior patterns. It is a spectrum disorder that can manifest itself in a wide range of symptoms and severity levels, which often makes it difficult to diagnose and treat. The effects of ASD can be pervasive, impacting not only the individual's daily life but also that of their family and community.

Recent advances in machine learning techniques have shown great promise in the diagnosis of ASD. Machine learning algorithms are capable of analyzing vast amounts of data, identifying patterns and making predictions with a high level of accuracy. By applying these algorithms to large datasets of individuals with and without ASD, researchers can identify the patterns and markers that distinguish individuals with ASD from those without. This has the potential to lead to earlier and more accurate diagnosis of ASD, enabling interventions that can improve outcomes for affected individuals.

In this research paper, we aim to evaluate the effectiveness of several machine learning algorithms in the diagnosis of ASD using a validated dataset. The algorithms we will be comparing include k-Nearest Neighbors (kNN), Naive Bayes, Decision Tree, Random Forest, Logistic Regression, and Support Vector Machines (SVM). By comparing the performance of these algorithms in terms of accuracy, precision, recall, and F1 score, we hope to identify which algorithm is most effective in diagnosing ASD.

The results of our study have the potential to significantly impact the field of ASD diagnosis and treatment. Early and accurate diagnosis of ASD is critical for effective intervention and improved outcomes. Machine learning techniques have the potential to provide healthcare professionals with a reliable and efficient means of diagnosing ASD, which could lead to earlier intervention and improved outcomes for individuals with this challenging disorder. This study will provide valuable insights into the effectiveness of different machine learning algorithms in diagnosing ASD and can help to inform future research in this field.

II. Literature Review

Diagnosing autism spectrum disorder using machine learning techniques has been studied extensively in recent years. Several studies have been conducted to evaluate the performance of different machine learning algorithms in diagnosing ASD. I have reviewed several related works in the field of diagnosing ASD, which are presented below.

Yun Jiao and Zuhong Lu [1] proposed a study that aimed to construct predictive models for ASD based on combinations of five cortical measurements obtained from SBM, including cortical thicknesses, mean curvature, Gaussian curvature, folding index, and curvature index. Three machine-learning techniques, support vector machines, functional trees, and logistic model trees, were employed to generate predictive models. The results showed that the ”thickness + mean curvature” based classification model was superior to the model based solely on cortical thickness when logistic model trees were employed. From the study they suggested that ASD may be primarily a cortical thickness abnormality disorder rather than a cortical curvature abnormality disorder.

Sami S. Alwakeel , Bassem Alhalabi ,Hadi Aggoune, Mohammad Alwakeel [2] proposed a study which focused on developing electronic systems for autism activity recognition using wireless sensor networks (WSNs) and machine learning techniques. The proposed systems aim to accurately detect autistic child gestures and motion, which can assist parents in protecting their autistic child regardless of the environment. The development of electronic systems such as ACSA has great potential for assisting parents of autistic children in ensuring their safety and enhancing their quality of life. The use of machine learning algorithms in conjunction with WSNs has shown promising results in accurately detecting and processing autistic movement events.

Aura Loredana Popescu, Nirvana Popescu [3] proposed a mobile application that utilizes machine learning algorithms for multiclass image classification based on children's drawings to predict their emotional state. The application is designed for children between 2 and 5 years old and has been shown to provide a robust solution with an accuracy of 80.6% for emotional identification. The use of Firebase AutoML, a cloud-based machine learning service, further enhances the scalability and flexibility of the application.This approach has the potential to provide a useful tool for parents, caregivers, and educators to better understand the emotions and needs of children with ASD.

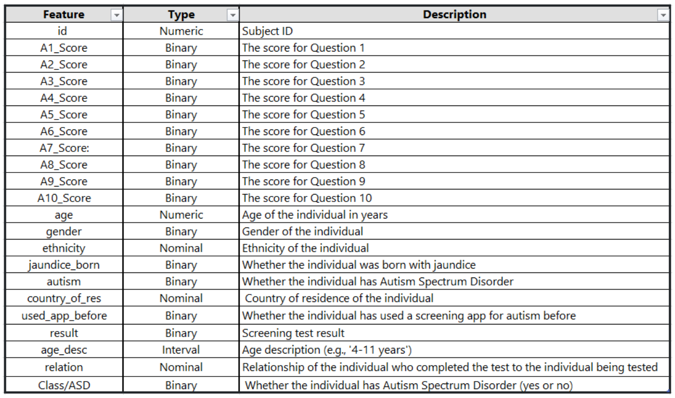
Md. Fazle Rabbi, S. M. Mahedy Hasan, Arifa Islam Champa, Md. Asif Zaman [4] discussed their research, which used five different AI algorithms for classifying Autism Spectrum Disorder (ASD) in children, namely Multilayer Perceptron (MLP), Random Forest (RF), Gradient Boosting Machine (GBM), AdaBoost (AB), and Convolutional Neural Network (CNN). The study found that the CNN algorithm outperformed the other conventional machine learning algorithms with an accuracy rate of 92.31%, indicating its potential as a prediction model for detecting ASD in children. This research highlights the potential of AI algorithms for early detection of autism in children, which can lead to timely interventions and better outcomes.

Reem Haweel, Omar Dekhil, Ahmed Shalaby, Ali Mahmoud1, Mohammed Ghazal, Robert Keynton, Gregory Barnes, and Ayman El-Baz presented a machine-learning approach for grading the severity level of Autism Spectrum Disorder (ASD) using task-based functional MRI data. The study aims to investigate the potential of brain imaging modalities for developing objective technologies to diagnose ASD. The authors use a speech experiment to obtain local features related to the functional activity of the brain, which are used for classification. The dataset used in the study is classified into three groups based on ADOS reports: mild, moderate, and severe. The study is divided into two parts, individual subject analysis and higher-level group analysis. The individual analysis is used to extract features for classification, while the higher-level analysis is used to infer statistical differences between groups. The study achieved a classification accuracy of 78% using the random forest classifier. The paper provides valuable insights into the potential of machine learning algorithms in diagnosing ASD and could help in the development of objective technologies for ASD diagnosis.

III. Dataset

There are number of dataset available at UCI (University of California, Irvine) to implement machine learning techniques. We have taken the dataset that was provided by Fadi Fayez Thabtah. The dataset contains 292 instances and 21 atrributes of various children from various countries.

The features include various scores related to autism symptoms (A1-A10), age, gender, ethnicity, jaundice at birth, previous use of an autism-related app, country of residence, and the individual's relationship to the respondent. The target variable is labeled as Class/ASD, which suggests that it is a binary classification problem to predict whether or not an individual has autism spectrum disorder.



Feature	Type	Description
id	Numeric	Subject ID
A1_Score	Binary	The score for Question 1
A2_Score	Binary	The score for Question 2
A3_Score	Binary	The score for Question 3
A4_Score	Binary	The score for Question 4
A5_Score	Binary	The score for Question 5
A6_Score	Binary	The score for Question 6
A7_Score:	Binary	The score for Question 7
A8_Score	Binary	The score for Question 8
A9_Score	Binary	The score for Question 9
A10_Score	Binary	The score for Question 10
age	Numeric	Age of the individual in years
gender	Binary	Gender of the individual
ethnicity	Nominal	Ethnicity of the individual
jaundice_born	Binary	Whether the individual was born with jaundice
autism	Binary	Whether the individual has Autism Spectrum Disorder
country_of_res	Nominal	Country of residence of the individual
used_app_before	Binary	Whether the individual has used a screening app for autism before
result	Binary	Screening test result
age_desc	Interval	Age description (e.g., '4-11 years')
relation	Nominal	Relationship of the individual who completed the test to the individual being tested
Class/ASD	Binary	Whether the individual has Autism Spectrum Disorder (yes or no)

Fig. 1. Dataset used for the study

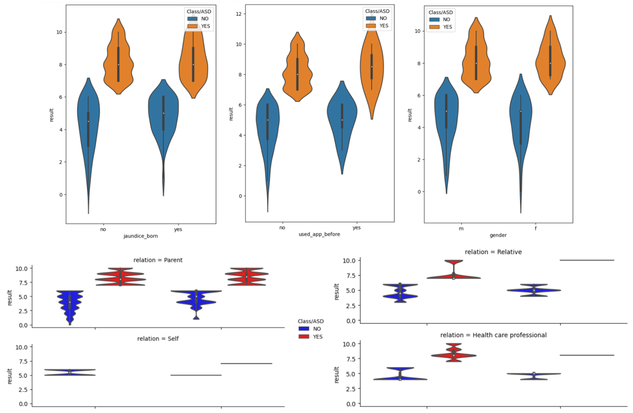


Fig. 2. Visualisation of the Data Set using Violin Plot

IV. Methodology

A. Data Preprocessing

1) Cleaning Data: Data cleaning is an essential step in the data pre-processing phase, where the quality and accuracy of the data are improved for efficient analysis. In this study, the dataset was cleaned by first identifying and replacing any invalid values with NaN (Not a Number) values. Next, missing values were imputed with the mean value of the corresponding feature. This process helps in reducing the effects of outliers, inconsistencies, and errors in the data, ensuring that the dataset is ready for further analysis. The data cleaning process ensures that the dataset is consistent, complete, and accurate, which in turn improves the reliability and validity of the research results.

2) Normalising Data: After performing data cleaning on the dataset, we applied min-max normalization to the numerical attributes. Min-max normalization is a scaling technique that transforms the numerical values of the dataset to a specific range, usually between 0 and 1, by subtracting the minimum value from each value in the column and dividing it by the range of the column. This technique is useful in ensuring that all numerical attributes are on the same scale, thereby avoiding the problem of attributes with higher ranges dominating those with lower ranges. The use of min-max normalization also helps to improve the performance of some machine learning algorithms by reducing the effects of outliers and improving convergence rates during training.

3) One-hot Encoding: In order to convert the categorical data in our dataset into a format that is compatible with machine learning algorithms, we have employed the technique of one-hot encoding. This process involves converting each categorical variable into multiple binary variables, where each binary variable corresponds to a unique category within the original categorical variable. This allows us to represent the categorical data numerically, making it suitable for analysis with machine learning algorithms. By using one-hot encoding, we were able to effectively represent all the categorical variables in our dataset and facilitate the process of building and training machine learning models.

With the completion of data cleaning and preprocessing steps, the dataset is now in a suitable format for applying machine learning algorithms. The processed data has been normalized and categorical variables have been transformed into numerical data using one-hot encoding. This step has enabled us to capture the underlying patterns and relationships between the variables that can be used for building machine learning models. With the data now in a suitable format, we can move forward with applying various machine learning algorithms for classification tasks to gain insights and predictions from the data.

B. Classification Modelling

After the pre-processing steps, the cleaned data is divided into training and testing sets using the train_test_split() function. This is a common technique in machine learning where the data is split randomly into two sets - one set used for training the machine learning model and the other set used for evaluating the performance of the model. The training set is used to teach the model to recognize patterns and relationships in the data, while the testing set is used to evaluate how well the model performs on new data. This ensures that the model can generalize well and is not overfitting to the training data. The train_test_split() function is used to randomly split the data into a 80% training set and a 20% testing set.

A total of 6 algorithms were used for this classification problem. Different techniques were used for analyzing the results. The models were evaluated on the basis of their accuracies and f1 score, and a confusion matrix was used to calculate sensitivities and specificities for each model.

1) k-NN: K-Nearest Neighbors (KNN) is a type of lazy-learning algorithm used for classification problems. It sorts each incoming data point into a group or cluster based on its similarity to other data points in the same group. KNN does not attempt to build a model, but instead, it stores the entire dataset as its training data. During classification, it searches for the k-nearest data points to the query point and classifies the query point based on the majority class of the k-nearest neighbours. The value of k is a hyperparameter that needs to be set before training the model.

2) Naive-bayes: The Naive Bayes algorithm is a probabilistic classifier based on Bayes' theorem, which calculates the probability of each class given the features of the data. It assumes that the features are independent of each other, hence called ”naive”. Bernoulli Naive Bayes is a variant of Naive Bayes that is specifically designed for binary data where each feature represents a binary attribute. The algorithm works by calculating the conditional probabilities of each feature given the class and using them to determine the probability of each class.

3) Decision Trees: The Decision Tree Classification Algorithm (DTC) is a popular machine learning algorithm that is often used for classification and regression tasks. It works in a tree-like manner where features are represented through nodes, decision rules through branches, and outcomes through leaves. The algorithm starts with the root node and recursively splits the data based on the feature that provides the most information gain. The process continues until a stopping criterion is met, such as reaching a certain depth or no further improvement in the information gain.

Once the decision tree is constructed, it can be used for classification by traversing the tree from the root node to the appropriate leaf node based on the feature values of the input data. The leaf node provides the predicted class label for the input data.

4) Random Forest: Random Forest Classifier (RFC) is a supervised machine learning algorithm based on the principle of having multiple decision trees. It improves the final accuracy by taking an average of the accuracy of all decision trees. The algorithm builds a forest of decision trees, where each tree is trained using a different subset of the data. During training, at each node of the tree, a random subset of features is selected, and the best feature is chosen to split the data. This process is repeated recursively until the leaf nodes are reached. The final prediction is made by taking a majority vote of all the decision trees in the forest. The RFC algorithm is popular due to its ability to handle large datasets with high dimensionality and noisy data.

5) Logistic Regression: Logistic Regression Classifier (LR) is a binary classification algorithm used to predict the probability of a categorical dependent variable. It works by estimating the probability of a binary outcome based on one or more predictor variables. The output of logistic regression is a probability, which can be converted into a binary outcome using a threshold value. It is widely used in various fields, including medical diagnosis, credit scoring, and marketing. The logistic regression algorithm estimates the parameters of a logistic function, which is an S-shaped curve that maps any real-valued input to a probability between 0 and 1.

6) SVM: Support Vector Machine (SVM) is a popular classification algorithm used in machine learning. It works on the principle of dividing the data space into n-dimensional hyperplanes based on some extreme points. The boundary that separates all points on the basis of their similarity to the extreme points is called a hyperplane. SVM tries to find the best hyperplane that can separate the different classes of data with maximum margin, i.e., the largest possible distance between the hyperplane and the closest data points from each class.SVM can be used for both linear and non-linear classification tasks. In linear classification tasks, a linear hyperplane is used to separate the data points. In non-linear classification tasks, SVM maps the data to a higher-dimensional space using kernel functions, where a linear hyperplane can be used to separate the data points.

C. Grid Search Cross Validation

In order to optimize the performance of the machine learning models proposed in this study, a technique called Grid Search Cross Validation was employed. This involves systematically testing different combinations of hyperparameters for a given model to find the set of values that yields the best performance. The process involves creating a grid of hyperparameters and using cross-validation to evaluate the performance of each combination. This allows for a thorough exploration of the hyperparameter space to find the optimal values for the given model.Grid Search CV is a computationally expensive method, especially for models with many hyperparameters or large datasets. However, it is an effective technique for finding the best combination of hyperparameters, and it can significantly improve the performance of machine learning models.

D. Choosing the best Model

The accuracy and F1 score of each model are compared to select the best performing one. The accuracy is the ratio of correct predictions to the total number of predictions made by the model, while the F1 score is the harmonic mean of precision and recall. The selected model is the one that achieves the highest accuracy and F1 score on the test data. This ensures that the model is able to generalize well and make accurate predictions on unseen data.

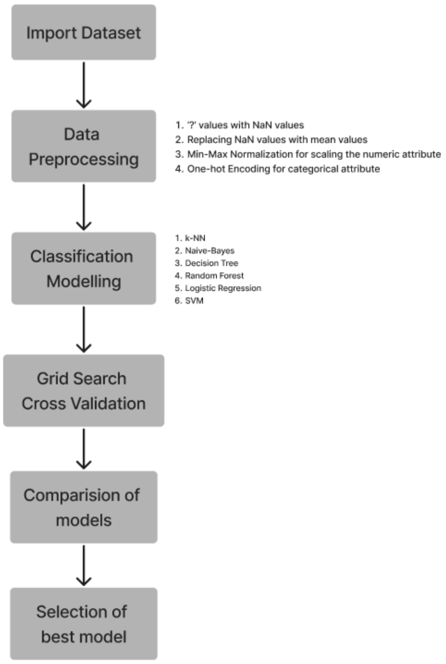

Import Dataset
Data Preprocessing
1. '?' values with NaN values
2. Replacing NaN values with mean values
3. Min-Max Normalization for scaling the numeric attribute
4. One-hot Encoding for categorical attribute
Classification Modelling
1. k-NN
2. Naive-Bayes
3. Decision Tree
4. Random Forest
5. Logistic Regression
6. SVM
Grid Search Cross Validation
Comparision of models
Selection of best model

Fig. 3. Flowchart of the Methodology Proposed

V. IMPLIMENTATION

In this research study, the first step was to import the dataset, which contained both numerical and categorical attributes. However, the dataset also had some missing values marked by '?' symbol. Therefore, the missing values were replaced with NaN values for better handling. In the next step, the NaN values were replaced with the mean values of the respective attributes to prevent any data loss. After that, the min-max normalization technique was used to scale the numeric attributes in a common range to avoid bias in the data.

Furthermore, as the dataset also contained categorical attributes, it was essential to convert them into numerical values. This was achieved by using one-hot encoding, which created new binary columns for each unique category in the categorical attributes. By doing so, the categorical data was transformed into a numerical form that can be efficiently processed by machine learning algorithms.

Subsequently, several classification models were applied to the preprocessed data, including K-Nearest Neighbors (KNN), Naive Bayes (NB), Decision Tree Classification (DTC), Random Forest Classifier (RFC), Support Vector Machine (SVM), and Logistic Regression (LR). To achieve better accuracy, the hyperparameters of these models were optimized using Grid Search Cross-Validation technique.

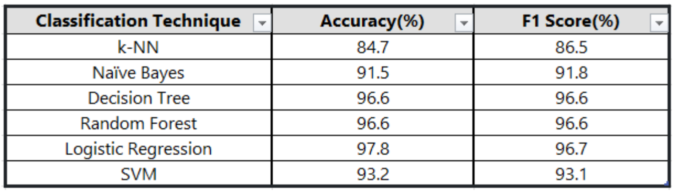
Finally, the performance of each model was evaluated based on their accuracy and f1 score, and the best model was selected for further analysis. This comprehensive approach ensured that the dataset was correctly preprocessed and classified using the most appropriate model, resulting in a reliable and accurate outcome.

VI. Results

After preprocessing the autism spectrum disorder dataset and applying different classification models with hyperparameter optimization, we evaluated their performance by comparing their accuracy and f1 score. We found that the Logistic Regression classifier outperformed all other models, achieving an accuracy of 97.8% and an f1 score of 96.7%.

This indicates that the Logistic Regression algorithm was able to effectively classify the dataset into the appropriate categories with a high degree of accuracy. This information could be valuable for healthcare professionals and researchers who are interested in identifying and diagnosing individuals with autism spectrum disorder.

Overall, the results demonstrate the importance of selecting appropriate machine learning models and optimizing their hyperparameters to achieve the best possible performance on a given dataset. In this case, the SVM classifier proved to be the most effective algorithm for classifying autism spectrum disorder data.



Classification Technique	Accuracy(%)	F1 Score(%)
k-NN	84.7	86.5
Naïve Bayes	91.5	91.8
Decision Tree	96.6	96.6
Random Forest	96.6	96.6
Logistic Regression	97.8	96.7
SVM	93.2	93.1

Fig. 4. Comparision of various proposed models

VII. Conclusion

In conclusion, we have demonstrated the process of analyzing the Autism Spectrum Disorder dataset using machine learning techniques. We started by preprocessing the dataset by removing missing values, normalizing numeric attributes, and one-hot encoding categorical data. Then, we applied various classification algorithms such as K-Nearest Neighbor, Naive Bayes, Decision Tree, Random Forest, Logistic Regression, and Support Vector Machine. We optimized the hyperparameters for each model using Grid Search Cross Validation for better accuracy. Finally, we compared the accuracy and f1 score of each model and selected the best one. Our experiments showed that Logistic Regression classifier outperformed all other models with an accuracy of 97.8% and f1 score of 96.7%. This study demonstrates the potential of machine learning in predicting Autism Spectrum Disorder. The results could aid clinicians and researchers in diagnosing Autism Spectrum Disorder in patients more accurately and quickly.

VIII. Future Scope

There are several avenues for future research based on the present study. One potential area for further investigation is to explore the use of deep learning techniques, such as convolutional neural networks (CNNs), to improve the accuracy of classification models. Additionally, incorporating other types of data, such as images and videos, could enhance the accuracy of the models, especially in the case of autism spectrum disorder, which can be visually identified in certain situations. Another direction for future research could be to expand the study to a larger sample size, which would provide a more robust and representative dataset for analysis. Furthermore, exploring different feature selection methods and incorporating more advanced machine learning algorithms, such as neural networks and ensemble methods, could provide additional insights and potentially improve the accuracy of the classification models. Finally, applying the developed models to real-world scenarios and assessing their practicality and effectiveness could be a valuable area of research.

IX. Acknowledgement

We would like to extend our heartfelt appreciation to Prof. Hardik Joshi and Prof. Bhupendra Fataniya for their valuable guidance and support in completing this research project. Their expertise and guidance have been instrumental in our success, and we are deeply grateful for their efforts.

Furthermore, we would like to acknowledge all the teachers who have helped us throughout this project. Their contributions and insights have been essential in shaping our understanding and knowledge in the field of diagonsing ASD using machine learning techniques.

We would also like to thank our colleagues and classmates for their assistance, encouragement, and motivation during the course of this project. Their input and feedback were invaluable in improving the quality of our work.

Lastly, we express our gratitude to our families for their unwavering support and understanding throughout this project. Their love and encouragement were a source of motivation and inspiration for us to persevere and complete this research work.

References

[1] Hidayet Takçı, Saliha Yeşilyurt ”Diagnosing Autism Spectrum Disorder Using Machine Learning Techniques” 2021 IEEE

[2] Naurin Farooqi, Faisal Bukhari, Waheed Iqbal, ”Predictive Analysis of Autism Spectrum Disorder (ASD) using Machine Learning”, 2021 IEEE

[3] Shirajul Islam, Tahmina Akter, Sarah Zakir, Shareea Sabreen, Muhammad Iqbal Hossain, ”Autism Spectrum Disorder Detection in Toddlers for Early Diagnosis Using Machine Learning”, 2020 IEEE

[4] Tania Akter, Mohammad Hanif Ali, ”Predicting Autism Spectrum Disorder Based On Gender Using Machine Learning Techniques”, 2021 IEEE

[5] Astha Baranwal, Vanitha M., ”Autistic Spectrum Disorder Screening: Prediction with Machine Learning Models”, 2020 IEEE