International Journal of Information Processing and Communication (IJIPC), Vol 1 No. 1, June, 2024, pp. 18-42

Online ISSN: 2645-2960 Print ISSN: 2141-3959

Detection of Autism Spectrum Disorder using Select Classification Algorithms

Fatima Enehezei Usman-Hamza[1], Mufidah Bibiresanmi Mohammed [1], Amos Orenyi Bajeh[1] Abimbola Ganiyat Akintola[1], Ghaniyyat B. Balogun[1], Peter O. Sadiku[1], Olatinwo Ikeola Suhurat[1], Abdulrafiu Muhammed Jamiu[1], Adebakin Taibat Olaoluwa[1], Taofeekat Tosin Salau-Ibrahim [2]

[1]Department of Computer Science, University of Ilorin, Ilorin, Nigeria
[2]Department of Cybersecurity, Federal University of Lafia.

usman-hamza.fe@unilorin.edu.ng; muffymohd2001@gmail.com; bajehamos@unilorin.edu.ng, akintola.ag@unilorin.edu.ng , balogun.gb@unilorin.edu.ng ;sadiku.po@unilorin.edu.ng ; olatinwo.is@unilorin.edu.ng ; abdulrafiu.mj@unilorin.edu.ng; adebakin.to@unilorin.edu.ng; taofeekat.tosin@cmp.fulafia.edu.ng

Abstract

Autism spectrum disorder (ASD) is a neurological and developmental disorder that often appears in infancy and impacts various developmental processes. This study highlights the use of multiple machine-learning algorithms to accurately diagnose autism spectrum disorder (ASD). It explores the application of classification algorithms, namely Support Vector Machine (SVM), k-Nearest Neighbours (KNN), Logistic Regression (LR), Neural Network (NN), and Random Forest (RF), for ASD detection. The main objective of the project is to assess the precision with which these algorithms can identify persons with ASD and those without. A comparative analysis measures the accuracy of Logistic Regression, KNN, Neural Network, Random Forest, and SVM. The project's overarching objective is to attain high accuracy rates and elevated levels of precision and recall metrics, pivotal for robust and dependable ASD detection. The remarkable performance of the Support Vector Machine (SVM) algorithm is of particular significance, as it achieved an unprecedented accuracy rate of 99.9%. This result underscores SVM's potential as an effective tool for precise ASD identification. The project's findings underline the synergy between advanced computational methods and medical diagnosis, illustrating the capacity of machine learning to aid clinicians and diagnosticians in early ASD detection. In conclusion, this project contributes to advancing ASD diagnosis by strategically selecting and comparing classification algorithms. The exceptional accuracy achieved by SVM signifies a pivotal stride forward in the quest for accurate and dependable ASD detection methods. This study exemplifies the potential of interdisciplinary collaboration between technology and healthcare, focusing on achieving high accuracy and comprehensive metrics for the improved identification and understanding of autism spectrum disorder in patients.

Keywords: *autism spectrum disorder, classification Algorithms, patients, detection.*

Online ISSN: 2645-2960 Print ISSN: 2141-3959

1. Introduction

Autism spectrum disorder (ASD) is a neurological and developmental disorder that often appears in infancy and impacts various developmental processes. It affects how people interact with others, communicate, learn, and behave (Ali et al., Fatimah & Al-Naimat; Yazan, 2019). Improving the quality of life for affected people and their families depends on early diagnosis and management. However, the present ASD diagnosis procedure mainly relies on subjective assessments, which can take time and increase the risk of misdiagnosis. As a result, impartial, precise, and effective techniques are required to diagnose ASD (Hayes et al., Tamsin & McCabe, Rose & Russell; Ginny, 2021). Although the precise etiology of autism is still unknown, it is thought to be brought on by a confluence of hereditary and environmental variables.

According to research, several genes and genetic mutations may have a role in the emergence of ASD. Additionally, potential influences include things like environmental exposures, disturbances in brain development, and prenatal problems. (Ali, Emad & Adwan, Fatimah & Al-Naimat, Yazan, 2019). Machine learning is a branch of artificial intelligence that can revolutionize medical diagnosis by using data to learn patterns and make predictions (Thomas et al., 2019). Several medical sectors, including the diagnosis of diabetes, cancer, and cardiovascular disease, have used machine learning algorithms recently with encouraging results.

Machine learning algorithms can be applied in the context of ASD to analyze and categorize massive datasets of clinical, behavioral, and demographic data to find patterns that distinguish between persons with and without ASD. (Yu-Hsin C, Qiushi C, Lan K, Guodong L, 2022). It is crucial to remember that machine learning algorithms must always be accompanied by a clinical examination conducted by a licensed healthcare practitioner and should never be utilized as the only method of diagnosis. The use of machine learning algorithms to identify autism and the possibility of increasing the speed and precision of autism diagnosis are the main goals of this work. The research is part of an expanding body of work that uses machine learning methods to address significant healthcare concerns and enhance patient outcomes.

In this study, we propose to develop a machine-learning approach to detect ASD using classification algorithms. This approach aims to improve the accuracy, efficiency, and objectivity of the diagnosis process for ASD. Specifically, we aim to collect a large and diverse dataset that contains records with and without ASD, pre-process and clean the data, extract relevant features, train, and evaluate multiple classification algorithms, fine-tune the best-performing algorithm, test the final machine learning system on a separate validation dataset, and analyze the results to identify the essential features for diagnosing ASD and any limitations or areas for improvement.

The exact causes of autism are still unknown, but scientists believe that genetic, environmental, and neurological factors may play a role (Centers for Disease Control and Prevention, 2021). Diagnosis of autism is usually based on behavioral and developmental evaluations by healthcare professionals, including pediatricians, neurologists, and psychologists (National Institute of Mental Health, 2021). The

International Journal of Information Processing and Communication (IJIPC), Vol 1 No. 1, June, 2024, pp. 18-42

Online ISSN: 2645-2960 Print ISSN: 2141-3959

autism spectrum is a range of conditions that includes autism, Asperger's syndrome, and pervasive developmental disorder not otherwise specified (PDD-NOS). The severity of symptoms and level of impairment can vary significantly within this spectrum. (Autism Society, 2021).

Research suggests that the brains of individuals with autism may develop differently than those without the disorder. In particular, there may be differences in the structure and connectivity of various regions of the brain. (Autism Research Institute, 2021). While there is no cure for autism, early intervention and treatment can significantly improve an individual's quality of life. Some of the most effective treatments for autism include behavioral therapy, speech therapy, occupational therapy, and medication. (Autism Speaks, 2021). It is essential to support individuals with autism in a way that respects their unique needs and abilities. This can involve creating a supportive and inclusive environment, providing appropriate accommodations, and working with professionals who specialize in autism (Autism Society, 2021).

2. Machine Learning

Machine learning is the process of using algorithms to automatically detect patterns in data and use these patterns to make predictions or decisions without being explicitly programmed. It is a rapidly growing field that has seen significant advancements in recent years (Mitchell, T. M. 1997). Machine learning has a wide range of applications, including natural language processing, image and speech recognition, recommender systems, fraud detection, medical diagnosis, and autonomous vehicles. It is also widely used in data analysis and predictive modeling. (Jordan, M. I., & Mitchell, T. M. 2015). Machine learning algorithms have been used to analyze various types of data, including behavioral and speech data, to identify patterns associated with ASD. These algorithms can help detect subtle differences in behavior and speech that may not be detectable by human observation alone. (Wall, D. P., Dally, R., Luyster, R., Jung, J. Y., & DeLuca, T. F. 2012).

a. Support Vector Machine (SVM)

Support Vector Machines (SVM) is a robust and widely used machine learning algorithm for both classification and regression tasks. SVM tries to find the best possible decision boundary that can separate the data points into different classes. It works by mapping the data points to a higher-dimensional space and finding the hyperplane that maximizes the margin between the different classes. SVM can handle both linearly separable and non-linearly separable data (Cortes et al.; V., 1995).

b. Random Forest Classifier (RFC)

Random Forest Classifier is an ensemble learning algorithm that combines multiple decision trees to improve the accuracy of the predictions. It works by constructing a multitude of decision trees and combining their outputs. Each tree is constructed using a random subset of the features and a random subset of the data. RFC can handle both binary and multi-class classification problems and is known for its high accuracy and robustness to noisy data (Breiman, L. 2001).

c. Logistic Regression (LR)

Logistic Regression is a popular machine learning algorithm used for binary classification problems. It models the probability of the binary outcome using a logistic function. LR works by finding the best possible decision boundary that can separate the data points into different classes. It can handle both linearly separable and non-linearly separable data (Hosmer et al.; R. X., 2013).

d. K-Nearest Neighbors

K-Nearest Neighbors is a simple and intuitive machine learning algorithm used for both classification and regression tasks. It works by finding the k-nearest neighbors to a given data point and assigning it to the most common class among those neighbors. KNN is a non-parametric algorithm, meaning it does not make any assumptions about the underlying distribution of the data (Cover et al., P. 1967).

e. Neural Network (NN)

Neural networks can be used for autism detection by building a classification model that predicts whether a given toddler has autism or not based on a set of input features. Neural networks are a type of machine learning algorithm that can learn complex nonlinear relationships between the input features and the output, making them a powerful tool for autism detection.
It can be a powerful tool for autism detection when combined with appropriate input features and careful data pre-processing. They can provide a highly accurate and robust model for binary classification, which can be helpful in clinical settings for early diagnosis and intervention (Paul et al., 2014).

3. Related Works

Several studies have made use of machine learning in various ways to improve and speed up the diagnosis of ASD. Machine learning (ML) approaches have shown potential in detecting ASD using classification algorithms. This section provides an overview of some of the relevant studies related to using a machine learning approach to detect ASD in patients.

In their paper, A machine learning autism classification based on logistic regression analysis, Thabtah, Fadi, Abdelhamid, Neda, Peebles, and David employed a machine learning algorithm popularly known as logistic regression to increase the effectiveness and efficiency of screening for autism spectrum disorder (ASD). Based on two datasets gathered via a mobile application called ASDTests, the authors offer a machine-learning framework. These datasets are connected to the adult and adolescent autism quotient (AQ) screening procedure, which is a questionnaire that adults and adolescents can self-administer to evaluate various traits and behaviors connected to autism spectrum disorder (ASD). It was first created by Simon Baron Cohen and his associates to help people spot characteristics associated with Asperger Syndrome.

Thabtah, Fadi, Abdelhamid, Neda, Peebles, and David's paper innovation is the availability of valuable datasets and computational intelligence-based features, which enables researchers to enhance ASD screening and offer knowledge to parents, caregivers, teachers, and healthcare facilities. This research introduces the use of machine learning and mobile technology to streamline the ASD screening process in comparison to the conventional approach. The researchers hope to improve accuracy, lower false positive and false negative rates, and speed up the referral process by adding machine learning techniques. Through the use of reachable intelligent devices, this strategy might yield quicker results and facilitate medical referrals. Although this was a better approach compared to traditional methods, which discussed the use of questionnaires and self-administered screening tools, this still had various limitations.
The datasets used in this investigation are first based on the AQ screening approach, which focuses on distinguishable characteristics associated with Asperger Syndrome in people of average intelligence. Given that the spectrum has a variety of presentations and variations, this may restrict the generalizability of the findings to the overall ASD population. Another difficulty noted is the lack of historical data relevant to behavioral science applications, particularly autism. The robustness and accuracy of the machine learning approach could potentially be impacted by the lack of complete behavior data for ASD screening. To guarantee the suggested approach to other populations and environments, it is crucial to validate it using more extensive and varied datasets.

The majority of these limitations were handled in the works and papers published by Ravindranath, Vaishali, and Ra. Sasikala. The authors suggest a technique that makes use of binary firefly feature selection based on swarm intelligence to find the bare minimal set of features necessary for distinguishing between patients with and without ASD. The innovation in this study is the discovery of a group of 10 variables from the original dataset of 21 qualities that may reliably identify patients with ASD and those without it. The authors identified a feature subset with an average accuracy range of 92.12% to 97.95% utilizing swarm intelligence-based feature selection. This degree of accuracy is comparable to using the complete dataset, showing that the smaller feature set may still capture the crucial data for precise classification.

This approach has several benefits over existing ASD detection techniques, such as logistic regression. First, it uses a machine learning strategy known as swarm intelligence that applies numerous categorization methods, giving it more flexibility in identifying intricate patterns in the data. Second, the algorithm can find the most pertinent features while removing noisy or redundant ones using swarm intelligence-based feature selection, increasing the prediction model's accuracy. In high-dimensional datasets, such as those relating to ASD diagnosis data, where conventional exhaustive search techniques become computationally expensive, this feature selection strategy is beneficial.

By using a smaller feature subset for training machine learning models, the scientists hope to improve prediction performance. They decrease the complexity of the dataset by choosing the most informative features using swarm intelligence-based feature selection, which also improves the model's capacity to generalize and make precise predictions on fresh, untested data. This strategy can considerably increase the performance and accuracy of machine learning models when used to predict ASD, as the article reveals.

Table 1a: Review of Related Works

S/N	AUTHOR	TITLE	OBJECTIVES	METHODOLOGY	GAPS
1.	Thabtah, Fadi & Abdelhamid, Neda & Peebles, David. (2019)	A machine learning autism classification based on logistic regression analysis.	To increase the effectiveness and efficiency of screening for autism spectrum disorder (ASD).	Based on two datasets gathered via a mobile application called ASD-Tests, the authors offer a machine learning framework using logistic regression.	The dataset used focuses on characteristics associated with Asperger Syndrome which restricts the generalizability of the findings to the overall ASD population.
2.	Ravindranath, Vaishali & Ra,.Sasikala. (2018)	A machine learning based approach to classify Autism with optimum behaviour sets.	To make use of binary firefly feature selection based on swarm intelligence to find the bare minimal set of features necessary for distinguishing between patients with and without ASD.	The authors identified a feature subset with an average accuracy range of 92.12% to 97.95% utilizing swarm intelligence-based feature selection.	This research is prone to premature convergence and poor local optimization ability.

Table 1b: Review of Related Works

International Journal of Information Processing and Communication (IJIPC), Vol 1 No. 1, June, 2024, pp. 18-42

Online ISSN: 2645-2960 Print ISSN: 2141-3959

S/N	AUTHOR	TITLE	OBJECTIVES	METHODOLOGY	GAPS
3.	Wall, Dennis & Dally, Rebecca & Luyster, Rhiannon & Jung, Jae-Yoon & Deluca, Todd. (2012).	Use of Artificial Intelligence to Shorten the Behavioral Diagnosis of Autism.	To develop a condensed version of the ADIR that could offer an accurate and speedy diagnosis of autism by examining the item-level responses and clinical diagnoses.	Data from more than 800 autistic people who had previously taken the through 93-item Autism Diagnostic Interview-Revised (ADI-R) were analyzed. The researchers looked for a subset of ADI-R questions that would produce results that were as accurate as the complete test.	The test, however, had age restrictions of 5 to 17 and was unable to accurately detect ASD at an early stage. since making an early discovery can greatly enhance their chances of developing normally.
4.	Bone D, Bishop SL, Black MP, Goodwin MS, Lord C, Narayanan SS. (2016).	Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion.	To create better screening and diagnostic tools for autism spectrum disorder (ASD) by integrating information from the Autism Diagnostic Interview-Revised (ADI-R) and Social Responsiveness Scale (SRS).	The study used a dataset with scores from the ADI-R and SRS for 1,264 people with ASD and 462 people with developmental or psychiatric illnesses that are not ASD, divided by age 10. The best-estimate clinical diagnosis of ASD versus non-ASD was the focus of the support vector machine classifier, which was used to build the machine learning methods.	Despite their great strides and use of a relatively larger dataset than previous works, their work faced similar limitations that Wall's experienced. Although it allowed for vast age range (4-55 years), their research was not approved as a screening method for all age groups as it didn't accommodate early discovery in young children.

Table 1c: Review of Related Works

S/N	AUTHOR	TITLE	OBJECTIVES	METHODOLOGY	GAPS
5.	Duda, M & Ma, R & Haber, N & Wall, Dennis. (2016).	Use of machine learning for behavioral distinction of autism and ADHD.	To make use of ML approaches to create classifiers based on a small subset of behavioral variables obtained from the Social Responsiveness Scale (SRS) in order to accurately diagnose autism spectrum disorder (ASD).	They developed models that successfully distinguished people with ASD from those with other mental illness using mutual information-based feature ranking, forward feature selection, and six ML algorithms.	The analysis was based mostly on collections of archival data related to autism, which led to an imbalance in favor of the ASD class. The researchers used repeated random under sampling and stratified 10-fold cross-validation to balance the classes.

Overall, these studies demonstrate the potential of ML approaches and classification algorithms in detecting ASD using various types of data, including imaging data, behavioral and developmental assessments, and biological markers. However, more research is needed to validate the reliability and effectiveness of these approaches, particularly with more extensive and standardized datasets. Additionally, ethical considerations such as data privacy and bias must be carefully considered and addressed. This is the focus of this study.

International Journal of Information Processing and Communication (IJIPC), Vol 1 No. 1, June, 2024, pp. 18-42

Online ISSN: 2645-2960 Print ISSN: 2141-3959

4. Method

4.1 Datasets

For comparing various data mining classification techniques, the Autistic Spectrum

Disorder Screening Data for Toddlers, 2019, was used. The dataset has 18 attributes and 1054 instances of records. However, only 14 attributes are used for this study. The dataset can be found on Kaggle.

Table 2: Autism Spectrum Disorder Dataset

Feature	Type	Description
A1-A10	Binary (0, 1)	The answer code of the question based on the screening method used
Speech delay	String	Based on how the child speaks
Voice pattern	Number	sluggish, difficulty speaking, etc.
Age	Number	Toddlers (months)
Score by Q-chat-10	Number	1-10 (Less than or equal to 3 no ASD traits; > 3 ASD traits
Sex	Character	Male or Female
Epilepsy	String	Whether the case was born with epilepsy
Born with jaundice	Boolean (yes or no)	Whether the case was born with jaundice
Family member with ASD history	Boolean (yes or no)	Whether any immediate family member has ASD
Aggressive Level	String	Aggressiveness in toddler
Echolisis	String	Use input textbox
Class variable	String	ASD traits or No ASD traits (automatically assigned by the ASD Tests app). (Yes / No)

4.2 Data Preprocessing

The first step in classifying algorithms is to prepare the data for analysis. This involves cleaning the data, applying feature selection, transforming it into a usable format, and selecting relevant features.

Feature Selection: Feature selection is the process of identifying and selecting a subset of input features that are most relevant to the target variable. To achieve this, we will use several techniques listed below:

Online ISSN: 2645-2960 Print ISSN: 2141-3959

I. Numeric Variables

a. Variables with near zero variance

It is advisable to remove variables with a single level (zero variance). As there are no variables with 1 level, we don't need to eliminate any variable.

Constant and almost constant predictors across samples (called zero and near-zero variance predictors, respectively) happen pretty often. One reason is that we usually break a categorical variable with many categories into several dummy variables. Hence, when one of the categories has zero observations, it becomes a dummy variable full of zeroes. From the code below, we took a sample of the dataset to detect if there is zero variance which is part of the data pre-processing technique; we can see from the result that we do not have variables with zero variance. We do this because we need to be sure of the nature of the dataset to give us an insight into what and how to handle our predictor class or variable.

Variables with near zero variance*

1. Count Unique Values using sapply()
2. For each numeric column in autism_train
3. Use unique(x) to extract distinct values.
4. Use length() to count unique values.
5. Sort the results using sort().
6. Identify Near-Zero Variance (NZV) Features with nearZeroVar()
7. Apply nearZeroVar() on numeric columns of autism_train.
8. Set saveMetrics = TRUE to store detailed feature metrics.
9. Save Results using saveRDS()
10.Store the NZV analysis result in "autism_nzv.rds".
11. Display the stored NZV metrics.

```
autism_nzv <- nearZeroVar(autism_train[, numeric_vars],
                          saveMetrics = TRUE)
saveRDS(autism_nzv, "autism_nzv.rds")
autism_nzv
```

Figure 1. Variables with near zero variance

Online ISSN: 2645-2960 Print ISSN: 2141-3959

	freqRatio	percent unique	zeroVar	nzv
A1	1.368590	0.270636	FALSE	FALSE
A2	1.205970	0.270636	FALSE	FALSE
A3	1.530822	0.270636	FALSE	FALSE
A4	1.041436	0.270636	FALSE	FALSE
A5	1.099432	0.270636	FALSE	FALSE
A6	1.346032	0.270636	FALSE	FALSE
A7	1.898039	0.270636	FALSE	FALSE
A7.1	1.898039	0.270636	FALSE	FALSE
A8	1.173529	0.270636	FALSE	FALSE
A9	1.081690	0.270636	FALSE	FALSE
A10	1.463333	0.270636	FALSE	FALSE

b. Correlations

Correlation expresses the extent to which two variables are related to each other. Features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable. Therefore, when two features have a high correlation, we can drop one of them. In this case, explanatory variables that can be said to be more correlated than others are Fam_mem_with_ASD and Q10 score chat. However, the correlation equals -0.34. Hence, there is no need to omit any of them.

Correlations

1. Compute Correlation Matrix using cor()
2. Select numeric variables from autism_train.
3. Calculate the correlation matrix using "complete.obs" to handle missing values.
4. Visualize Correlation Matrix with corrplot.mixed()
5. Display correlations as numbers in the upper triangle.
6. Represent correlations as circles in the lower triangle.
7. Set text labels in black (tl.col = "black") and position them at the top left (tl.pos = "lt").

Figure 2. Correlations

International Journal of Information Processing and Communication (IJIPC), Vol 1 No. 1, June, 2024, pp. 18-42

Online ISSN: 2645-2960 Print ISSN: 2141-3959

II. CATEGORICAL VARIABLES

a. Stepwise Regression

Bayesian Information Criterion (BIC) is an estimate of a function of the posterior probability of a model being true under a certain Bayesian setup so that a lower BIC means that the model is considered to be more likely to be the true model.

```
Null_model <- lm(class_with_ASD ~ 1 , data = autism_train_numeric)

full_model <- lm(autism_EVENT ~ . , data = autism_train_numeric)

BIC_model <- step( null_model, scope = list( lower = null_model, upper = full_model), k = log(nrow(autism_train_numeric)), direction = “forward”)
```

Figure 3. Stepwise Regression

b. Selected Classification Algorithms

The dataset is first pre-processed to remove noise, missing values, and outliers and encode category properties. Additionally, we use dimensionality Reduction to improve the model performance and to select the features from the data collection that will be most useful. Increasing training efficiency and speed, this decreases the data’s dimensionality. After the data set has been pre-processed, our selected classification algorithms—Support Vector Machines (SVM), Random Forest Classifier (RFC), Logistic Regression (LB), K-Nearest Neighbours (KNN), and Naive Bayes—are utilized to predict the output label (ASD or no ASD). Each classifier’s accuracy is evaluated and contrasted. To assess each classifier more effectively, metrics like the F1 score and precision-recall values have also been generated. The training accuracy will be higher than the test accuracy if the classifier works appropriately.

5. Findings and Discussion

A new technique for predicting and detecting autism using a patient data set is presented. This new technique uses SVM, KNN, Logistic Regression, Neural Network, and Random Forest models that will produce the results based on the resulting input dataset. This new system produces results from mathematical calculations based on probabilistic theory. The results produced using mathematical calculations tend to be more accurate, thus improving system efficiency and accurate prediction of autism system development.

Online ISSN: 2645-2960 Print ISSN: 2141-3959

5.1 Prediction of results using the Logistic Regression Algorithm

Figure 4b shows the number of instances of toddlers with autism and without autism with respect to the dataset; over 300 toddlers show signs of autism, and 700 toddlers show the absence of autism.

```
counts <- table(data_autism$Class.ASD.Traits.)
barplot(counts,
   xlab="Autism Events",
      ylab="Number of instances")
```

Figure 4a. Autism Events

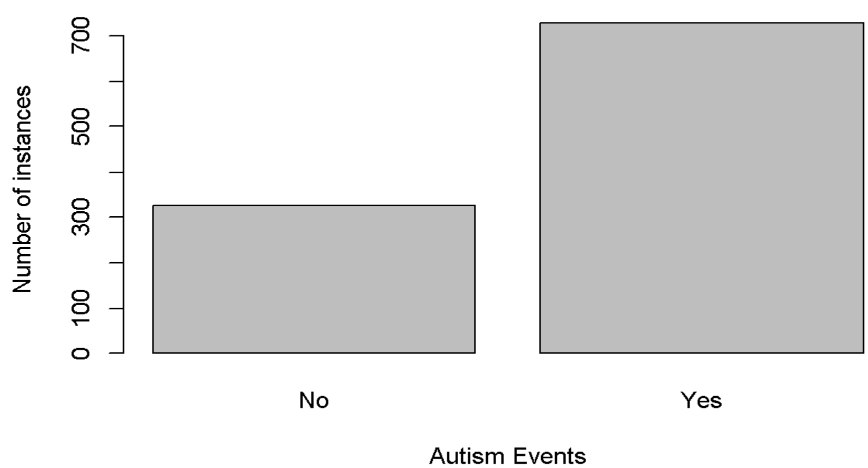

Figure 4b. Autism Events

Likely, the graphical image below depicts the relationship among A1-A10, age, case number, and qchat score, this helps give a vivid understanding of the dataset and how to wriggle around it.

International Journal of Information Processing and Communication (IJIPC), Vol 1 No. 1, June, 2024, pp. 18-42

Online ISSN: 2645-2960 Print ISSN: 2141-3959

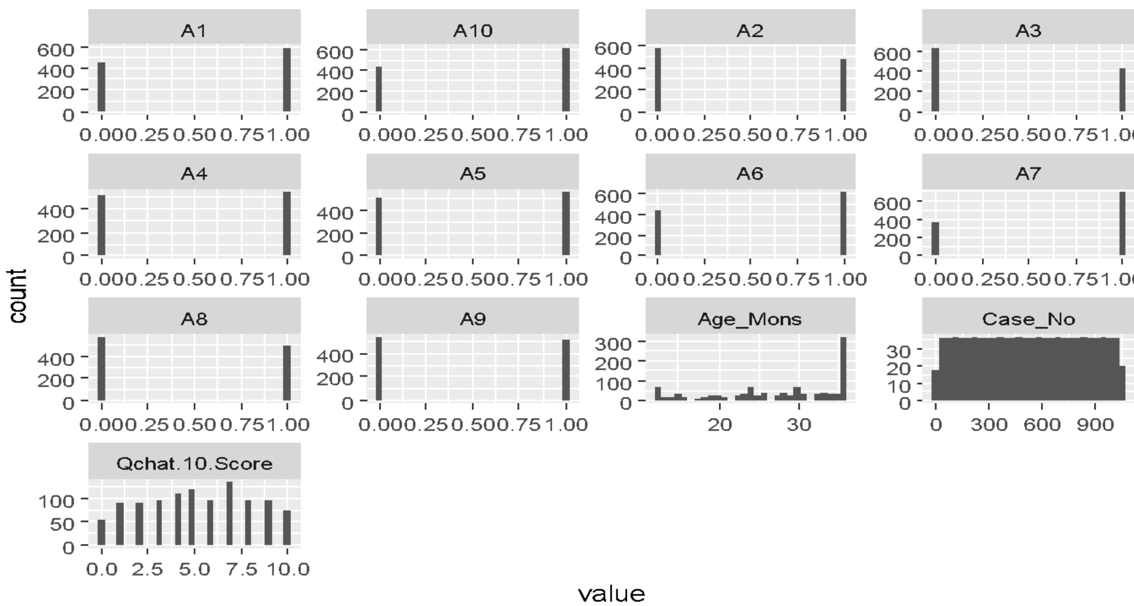

Figure 5. Graphical Relationship

The majority of the toddlers are in their 10th and 25th month; the graph gives us a visual insight that toddlers below 10th month and 25th month are likely not to show any sign of autism, but as they grow older, chances of being diagnosed is possible.

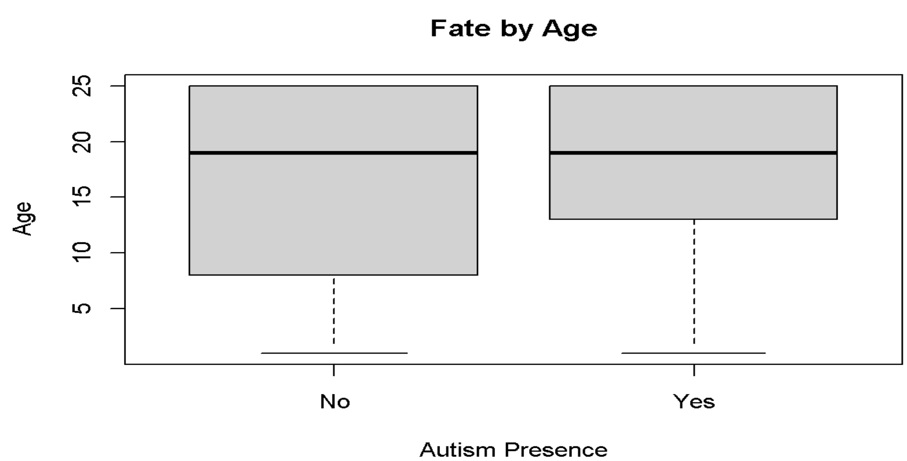

Figure 6. Most Frequent Age Difference

Data columns having an excessive number of missing values would not be very useful. As a result, we must determine whether any values are missing from our dataset. As can be seen in Figure 7, there are no missing values.

Online ISSN: 2645-2960 Print ISSN: 2141-3959

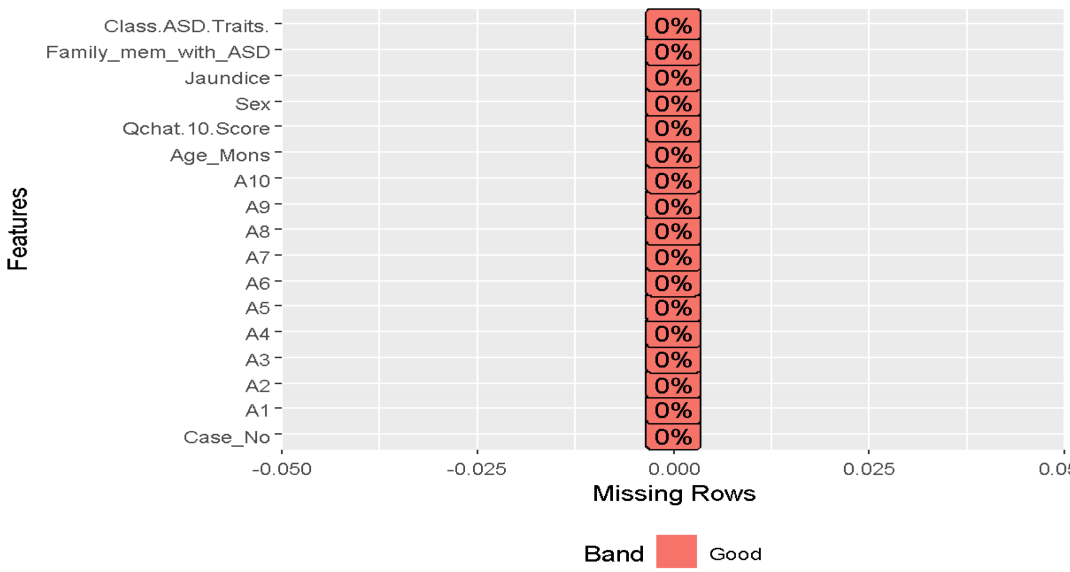

Figure 7. Missing Values

From our analysis below, we can clearly see that logistic regression is not suitable for this study with an accuracy of 27%; therefore, it is not recommendable.

```
>x <- data.frame(Case_No="10")
p<- predict(logit,x)
p
1
26.56607
```

Figure 8. Autism Prediction using Logistic Regression

5.2 Prediction of results using K-nearest neighbour Algorithm

The k-nearest neighbor was applied to the dataset; accuracy was used to select the optimal model using the most significant value. The final value used for the model was k = 1, giving an accuracy of 76%, which is far better and more productive than the logistic regression algorithm.

Online ISSN: 2645-2960 Print ISSN: 2141-3959

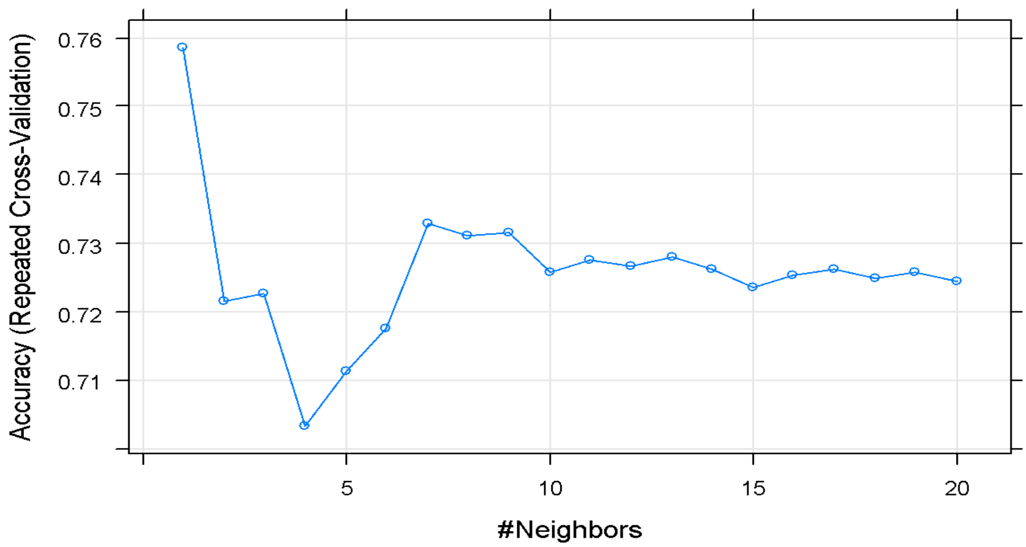

Figure 9. K-Nearest Neighbour

5.3 Prediction of results using Neural Network Algorithm

Once an input layer is determined, weights are assigned. These weights help determine the importance of any given variable, with larger ones contributing more significantly to the output compared to other inputs. All inputs are then multiplied by their respective weights and then summed. Afterward, the output is passed through an activation function, which determines the output. If that output exceeds a given threshold, it "fires" (or activates) the node, passing data to the next layer in the network. This results in the output of one node becoming the input of the next node. This process of passing data from one layer to the next layer defines this neural network as a feed-forward network.

As we can see from our diagram below in Figure 10b, the Neural network algorithm produced an accuracy of 99%, approximately 100%, proving that it is suitable for this study for further research works.

International Journal of Information Processing and Communication (IJIPC), Vol 1 No. 1, June, 2024, pp. 18-42

Online ISSN: 2645-2960 Print ISSN: 2141-3959

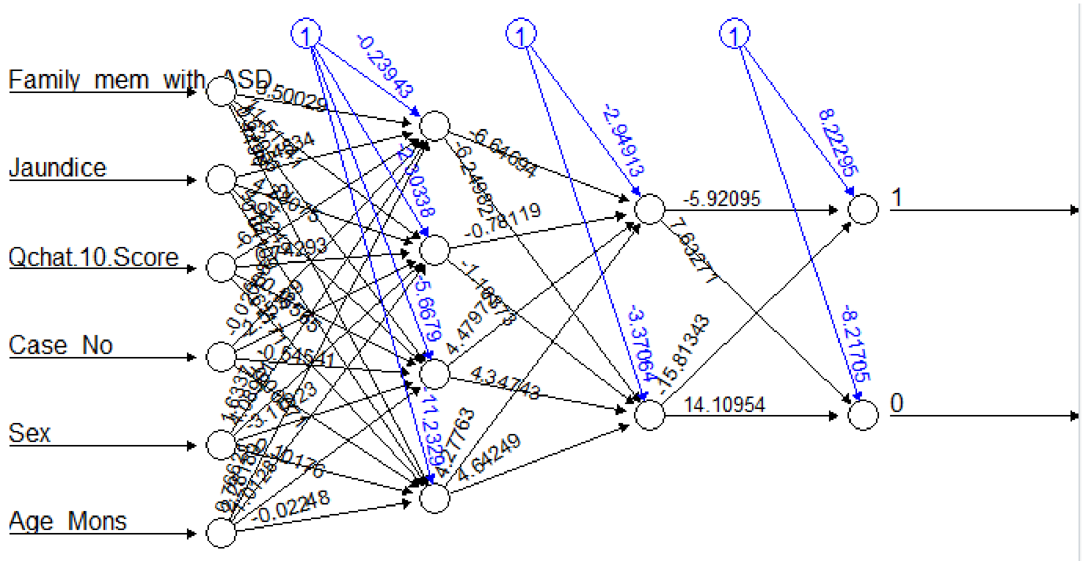

Figure 10a. Neural Network

```
table(test_data$Class.ASD.Traits, prediction_label)
prediction_label
Age_Mons Case_No
0    1    63
1    147    0
check = as.numeric(test_data$Class.ASD.Traits) == max.col(pred)
accuracy = (sum(check)/nrow(test_data))*100
print(accuracy)
[1]    99.52607
```

Metric against Value

Accuracy: 99.53%
Sensitivity: 30.00%
Specificity: 100.00%
Precision (PPV:100.00%
Negative Predictive Value (NPV): 0.68%
Balanced Accuracy: 65.00%

International Journal of Information Processing and Communication (IJIPC), Vol 1 No. 1, June, 2024, pp. 18-42

Online ISSN: 2645-2960 Print ISSN: 2141-3959

This is the Confusion Matrix:

1. True Positives (TP): 63
2. False Positives (FP): 0
3. False Negatives (FN): 147
4. True Negatives (TN): 1

Figure 10b. Autism Prediction using Neural Network

5.4 Prediction of results using Random Forest Algorithm

The graphical representation below shows the relationship between sets of selected attributes from the dataset; it shows that A1 to A4 information is a better predictor compared to the case_no information. The last column and the last row demonstrate the point. As we can see, Random Forest gave an accuracy of 31%, which is not suitable for this study but can be applied to another health dataset to prove its efficiency.

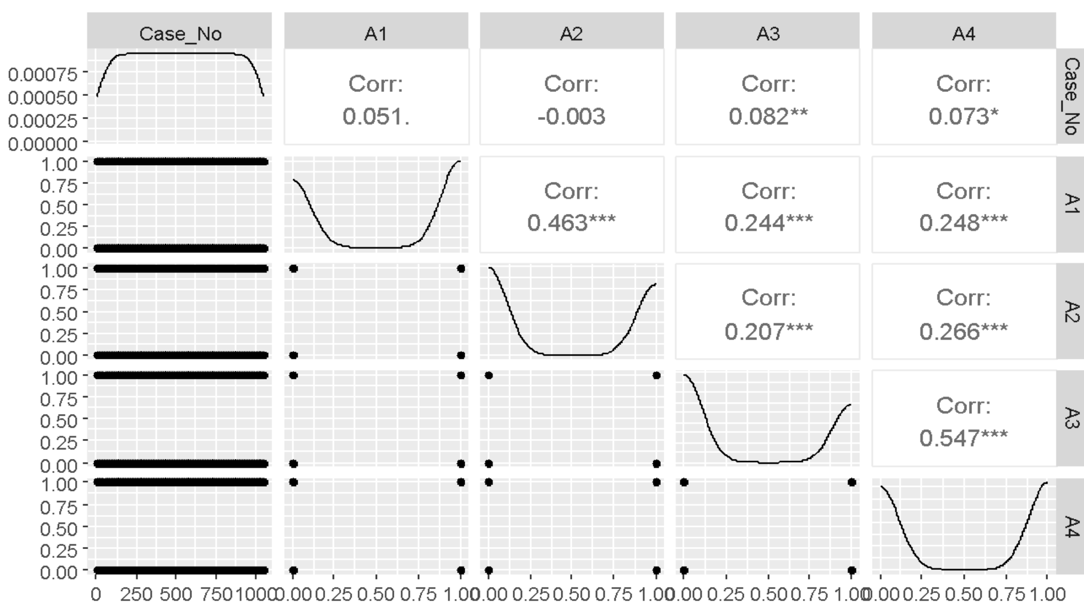

Figure 11a. Relationship between variables

International Journal of Information Processing and Communication (IJIPC), Vol 1 No. 1, June, 2024, pp. 18-42

Online ISSN: 2645-2960 Print ISSN: 2141-3959

```
>print(cm)
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 32  0
         1  0 72

               Accuracy : 30.769
                 95% CI : (0.9652, 1)
    No Information Rate : 0.6923
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.3077
         Detection Rate : 0.3077
   Detection Prevalence : 0.3077
      Balanced Accuracy : 1.0000

       'Positive' Class : 0
```

Figure 11b. Autism Prediction using Random Forest

5.5 Prediction of results using Support Vector Machine Algorithm

Support vector machine (SVM) has proven to be the best algorithm suitable for this research work with an accuracy of 99.9% and can be implemented in real life situations for the prediction of autism.

International Journal of Information Processing and Communication (IJIPC), Vol 1 No. 1, June, 2024, pp. 18-42

Online ISSN: 2645-2960 Print ISSN: 2141-3959

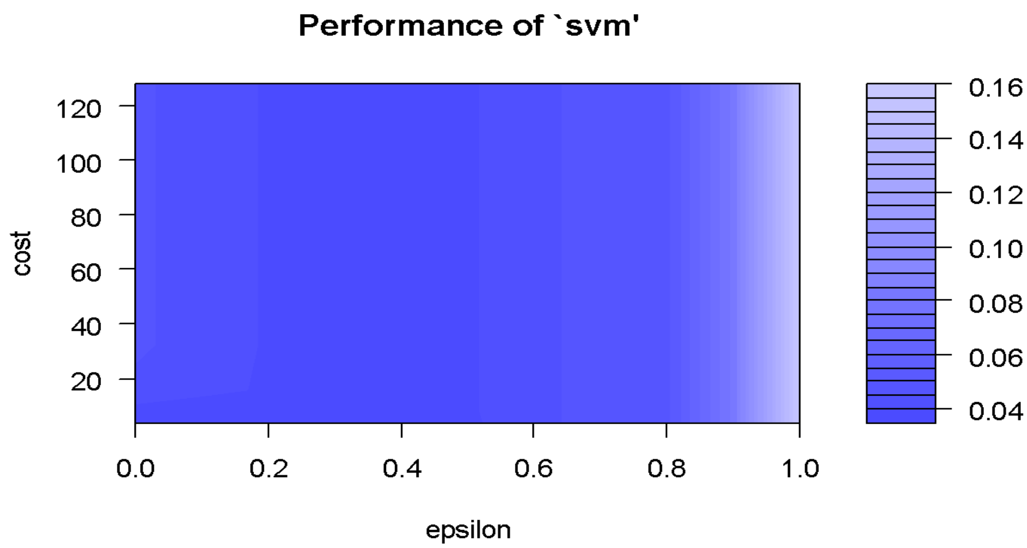

Figure 12a. Performance of Support Vector Machine

The cost vs. epsilon parameter decides how much an SVM should be allowed to "bend" with the data. For a low cost, you strive for a smooth decision surface; for a higher cost, you want to correctly classify more points. It is also known as the epsilon cost of misclassification.

```
[ reached getOption("max.print") -- omitted 554 rows ]
> 1-sum(diag(tab1)/sum(tab1))
[1] 0.9990512
> pred = predict(svm_model,data_autism)
> tab = table(Predicted=pred, Actual = data_autism$Class.ASD.Traits)
> tab
          Actual
Predicted
```

Figure 12b. Autism Prediction using Support Vector Machine

5.6 Discussion

This study compares the accuracy of the Logistic regression, KNN, Neural Network, Random Forest, and SVM algorithm. The goal is to have high accuracy, besides high precision and recall metrics. Although these metrics are used more often in the field of information retrieval, here we have considered them as they are related to the other existing metrics, such as specificity and sensitivity. These metrics can be derived from the confusion matrix and can be easily converted to true-positive (TP) and false-positive (FP) metrics; Table 2 below shows the performance accuracy in percentage % with the time taken to build each model.

Online ISSN: 2645-2960 Print ISSN: 2141-3959

Table 2. Performance accuracy of different classification techniques

Classification	Accuracy
Logistic Regression	26.6 %
KNN	76 %
Neural Network	99.5 %
Random Forest	31 %
SVM	99.9%

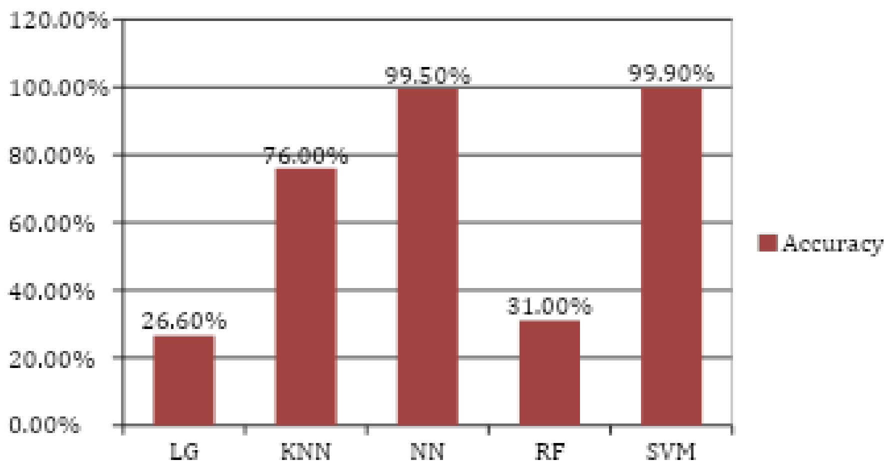

Figure 13. Performance analysis of Autism prediction

6. Conclusion

The accurate detection of ASD in patients can result in early interventions, specialized assistance, and improved long-term outcomes. By using advanced algorithms, this approach not only highlights the feasibility of accurate ASD detection but also emphasizes the importance of precision and recall for a comprehensive and effective diagnosis.

The primary objective of the project was to evaluate the performance of these classification algorithms in terms of accuracy. The study involved a comparison of the accuracy achieved by each algorithm, with a specific focus on Logistic Regression, KNN, Neural Networks, Random Forest, and SVM. Among the evaluated algorithms, Support Vector Machine (SVM) emerged as the most effective one, achieving an impressive accuracy rate of 99.9%.

This outcome indicates that the SVM algorithm exhibited the highest capability in correctly classifying persons with or without an autism spectrum disorder. The project's findings suggest that SVM could potentially serve as a reliable tool for accurately detecting ASD, with its excellent performance on accuracy metrics. In conclusion, the study presents a promising step towards enhancing ASD diagnosis and intervention strategies through innovative machine learning methodologies.

The significance of the project lies in its potential to contribute to the field of healthcare, specifically in early ASD detection. With the high accuracy achieved through SVM, which recorded an impressive 99.9% accuracy rate, the project showcases the potential of machine learning techniques in assisting clinicians and diagnosticians. The implications of this project extend to the healthcare domain, where early detection of ASD holds immense value. Timely identification enables timely interventions, personalized therapeutic strategies, and improved outcomes for patients on the autism spectrum. The integration of machine learning techniques into the diagnostic process exemplifies the synergy between technology and healthcare, showcasing how innovation can lead to tangible benefits for individuals, families, and society.

References

Ali, Emad & Adwan, Fatimah & Al-Naimat, Yazan. (2019). Autism Spectrum Disorder (ASD); Symptoms, Causes, Diagnosis, Intervention, and Counseling Needs of the Families in Jordan. Modern Applied Science. 13. 48. 10.5539/mas.v13n5p48. https://www.researchgate.net/publication/332778242_Autism_Spectrum_Disorder_ASD_Symptoms_Causes_Diagnosis_Intervention_and_Counseling_Needs_of_the_Families_in_Jordan/citation/download

Alpaydin, E. (2010). Introduction to machine learning (2nd ed.). MIT Press. (Supervised Learning)

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3), 175-185. https://doi.org/10.2307/2685208

American Occupational Therapy Association. (2020). Autism spectrum disorder. https://www.aota.org/-/media/Corporate/Files/AboutOT/Advocacy/State/Resources/Autism%20Spectrum%20Disorder%20Fact%20Sheet.pdf

American Speech-Language-Hearing Association. (2020). Autism spectrum disorder. https://www.asha.org/public/speech/disorders/Autism-Spectrum-Disorder/

Autism Research Institute. (2021). Brain Development in Autism. Retrieved from https://www.autism.com/science_brain

Autism Society. (2021). About Autism. Retrieved from https://www.autism-society.org/what-is/

Autism Society. (2021). Support & Services. Retrieved from https://www.autism-society.org/living-with-autism/how-we-can-help/

Online ISSN: 2645-2960 Print ISSN: 2141-3959

Autism Speaks. (2021). Treatment. Retrieved from https://www.autismspeaks.org/treatment

Autism Speaks. (2021). What is Autism? Retrieved from https://www.autismspeaks.org/what-autism

Baio, J., Wiggins, L., Christensen, D. L., Maenner, M. J., Daniels, J., Warren, Z., ... & Dowling, N. F. (2018). Prevalence of autism spectrum disorder among children aged 8 years—Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014. MMWR Surveillance Summaries, 67(6), 1-23.

Bone D, Bishop SL, Black MP, Goodwin MS, Lord C, Narayanan SS. Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion. J Child Psychol Psychiatry. 2016 Aug;57(8):927-37. doi: 10.1111/jcpp.12559. Epub 2016 Apr 19. PMID: 27090613; PMCID: PMC4958551.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Centers for Disease Control and Prevention. (2021). Autism Spectrum Disorder (ASD). Retrieved from https://www.cdc.gov/ncbddd/autism/index.html

Chapelle, O., Schölkopf, B., & Zien, A. (Eds.). (2006). Semi-supervised learning (1st ed.). MIT Press. (Semi-Supervised Learning)

Chien, L. N., Lin, H. C., & Chung, C. H. (2014). Increased risk of autism in children born to mothers with systemic lupus erythematosus: A retrospective cohort study. Lupus, 23(1), 43-48.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297. https://doi.org/10.1007/BF00994018

Donnellan AM, Hill DA, Leary MR. Rethinking autism: implications of sensory and movement differences for understanding and support. Front Integr Neurosci. 2013 Jan 28; 6:124. doi: 10.3389/fnint.2012.00124. PMID: 23372546; PMCID: PMC3556589.

Duda, M & Ma, R & Haber, N & Wall, Dennis. (2016). Use of machine learning for behavioural distinction of autism and ADHD. Translational psychiatry. 6. e732. 10.1038/tp.2015.221.

Ecker, C., Bookheimer, S. Y., & Murphy, D. G. (2017). Neuroimaging in autism spectrum disorder: Brain structure and function across the lifespan. The Lancet Neurology, 16(6), 545-555. doi: 10.1016/S1474-4422(17)30100-3

Gardener, H., Spiegelman, D., & Buka, S. L. (2011). Prenatal risk factors for autism: Comprehensive meta-analysis. The British Journal of Psychiatry, 198(4), 266-272. doi: 10.1192/bjp.bp.110.077091

Online ISSN: 2645-2960 Print ISSN: 2141-3959

Gardener, H., Spiegelman, D., & Buka, S. L. (2011). Prenatal risk factors for autism: comprehensive meta-analysis. The British Journal of Psychiatry, 199(6), 417-427.

Hallmayer, J., Cleveland, S., Torres, A., Phillips, J., Cohen, B., Torigoe, T., ... & Risch, N. (2011). Genetic heritability and shared environmental factors among twin pairs with autism. Archives of General Psychiatry, 68(11), 1095-1102.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. (Unsupervised Learning)

Hayes, Jennie & Ford, Tamsin & McCabe, Rose & Russell, Ginny. (2021). Autism diagnosis as a social process. Autism. 26. 136236132110303. 10.1177/13623613211030392.

Heoden I. Netoff (2019), Chapter 14 - The Ability to Predict Seizure Onset,Pages 365-378, ISBN 9780128130681, https://doi.org/10.1016/B978-0-12-813068-1.00014-2. (https://www.sciencedirect.com/science/article/pii/B9780128130681000142)

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons. https://doi.org/10.1002/9781118548387

Hossain, Md & Kabir, Ashad & Anwar, Adnan & Islam, Md. (2021). Detecting autism spectrum disorder using machine learning techniques. Health Information Science and Systems. 9. 10.1007/s13755-021-00145-9. https://www.researchgate.net/publication/350682394_Detecting_autism_spectrum_disorder_using_machine_learning_techniques/

Hus, Yvette & Segal, Osnat. (2021). Challenges Surrounding the Diagnosis of Autism in Children. Neuropsychiatric Disease and Treatment. 17. 3509–3529. 10.2147/NDT.S282569. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8654688/

Ibrahim Rihan,(2023) The Feasibility Study https://www.academia.edu/29462582/The_Feasibility_Study

Kalkbrenner, A. E., Daniels, J. L., Chen, J. C., Poole, C., Emch, M., Morrissey, J., ... & Sandler, D. P. (2010). Perinatal exposure to hazardous air pollutants and autism spectrum disorders at age 8. Epidemiology, 21(5), 631-641.

Kaushik V, Diya P, Deepa K, 2021: Vakadkar, K., Purkayastha, D. & Krishnan, D. Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques. *SN COMPUT. SCI.* 2, 386 (2021). https://doi.org/10.1007/s42979-021-00776-5

Online ISSN: 2645-2960 Print ISSN: 2141-3959

Kennedy Krieger Institute, August 21, 2014, Twins Study Finds Large Genetic Influence in Autism- Interactive Autism Network (IAN) Research Project.https://www.kennedykrieger.org/stories/interactive-autism-network-ian/autism-twins-study

Koegel, Lynn & Koegel, Robert & Ashbaugh, Kristen & Bradshaw, Jessica. (2013). The importance of early identification and intervention for children with or at risk for autism spectrum disorders. International journal of speech-language pathology. 16. 10.3109/17549507.2013.861511. https://www.researchgate.net/publication/259313969_The_importance_of_early_identification_and_intervention_for_children_with_or_at_risk_for_autism_spectrum_disorders/

Krakowiak, P., Walker, C. K., Bremer, A. A., Baker, A. S., Ozonoff, S., Hansen, R. L., ... & Hertz-Picciotto, I. (2012). Maternal metabolic conditions and risk for autism and other neurodevelopmental disorders. Pediatrics, 129(5), e1121-e1128.

Lyall, K., Croen, L., Daniels, J., Fallin, M. D., Ladd-Acosta, C., Lee, B. K.,... & Newschaffer, C. (2017). The changing epidemiology of autism spectrum disorders. Annual Review of Public Health, 38, 81-102. doi: 10.1146/annurev-publhealth-031816-044318

Mahesh, Batta. (2019). Machine Learning Algorithms -A Review. 10.21275/ART20203995.

Nadeem, Muhammad & Murtaza, Bibi & Al-Ghamdi, Mariam & Ali, Akbar & Zamzami, Mazin & Khan, Jalaluddin & Ahmad, Aftab & Rehman, Mujaddad & Kazmi, Imran. (2021). Autism - A Comprehensive Array of Prominent Signs and Symptoms. Current Pharmaceutical Design. 27. 10.2174/1381612827666210120095829. https://www.researchgate.net/publication/348793156_Autism_-_A_Comprehensive_Array_of_Prominent_Signs_and_Symptoms#:~:text=Speech%20or%20hearing%20impairments%2C%20poor,pain%20include%20common%20physiological%20symptoms.

National Institute of Mental Health. (2021). Autism Spectrum Disorder. Retrieved from https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd/index.shtml

National Institute of Mental Health. (n.d.). Autism spectrum disorder. https://www.nimh.nih.gov/health/topics/autism-spectrum-disorders-asd/index.shtml

Ravindranath, Vaishali & Ra,.Sasikala. (2018). A machine learning-based approach to classify Autism with optimum behaviour sets. International Journal of Engineering and Technology. 7. 10.14419/ijet.v7i3.18.14907Published.

Reaven, Judy & Hepburn, Susan & Ross, Randy. (2008). Use of the ADOS and ADI-R in Children with Psychosis: Importance of Clinical Judgment. Clinical child psychology and psychiatry. 13. 81-94. 10.1177/1359104507086343. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4426195/

Rish, I. (2001). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, pp. 41-46).

Sandin, S., Lichtenstein, P., Kuja-Halkola, R., Larsson, H., Hultman, C. M., & Reichenberg, A. (2014). The familial risk of autism. JAMA, 311(17), 1770-1777.

Schipul, S. E., Wallace, A. S., Williams, D. L., Robins, D. L., & McPartland, J. C. (2012). Elevated basal serum glutamate and unchanged glutamine and GABA in individuals with autism spectrum disorder. JAMA Psychiatry, 69(12), 1318-1325. doi: 10.1001/archgenpsychiatry.2012.104

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press. (Reinforcement Learning)

Thabtah, Fadi & Abdelhamid, Neda & Peebles, David. (2019). A machine learning autism classification based on logistic regression analysis. Health Information Science and Systems. 7. 10.1007/s13755-019-0073-5.

Thomas Davenport and Ravi Kalakota 2019: Thomas Davenport and Ravi Kalakota 2019) Future Healthc J. 2019 Jun; 6(2): 94–98. doi: 10.7861/futurehosp.6-2-94

U.S. Department of Education. (n.d.). A guide to the individualized education program. https://www2.ed.gov/parents/needs/speced/iepguide/index.html

Wall, Dennis & Dally, Rebecca & Luyster, Rhiannon & Jung, Jae-Yoon & Deluca, Todd. (2012). Use of Artificial Intelligence to Shorten the Behavioral Diagnosis of Autism. PloS one. 7. e43855. 10.1371/journal.pone.0043855.

Werling, D. M., & Geschwind, D. H. (2013). Sex differences in autism spectrum disorders. Current Opinion in Neurology, 26(2), 146-153.

Yu-Hsin C, Qiushi C, Lan K, Guodong L, 2022: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9462117/.