Article

A machine learning approach to predict autism spectrum disorder (ASD) for both children and adults using feature optimization

Khandaker Mohammad Mohi Uddin[1], **Hasibur Rahman**[1], **Mahadi Hasan**[1], **Fatema Akter**[2], **Suman Chandra Das**[3]

[1]Department of Computer Science and Engineering, Dhaka International University, Dhaka- 1205, Bangladesh

[2]Department of Computer Science and Engineering, Jagannath University, Dhaka, Bangladesh

[3]Bangabandhu Sheikh Mujib Medical College, Faridpur, Bangladesh

E-mail: jilanicsejnu@gmail.com, Itzhasib1@gmail.com, Mehedih036@gmail.com, fatema.jnu.cse@gmail.com, sumanfmc21@gmail.com

Received 4 December 2022; Accepted 10 January 2023; Published online 28 January 2023; Published 1 June 2023

Abstract

A central nervous system known as an Autism Spectrum Disorder (ASD) has long-term effects on a person's capacity for engagement and interaction with others. Since its symptoms often manifest in the first two years of life, autism is considered to be a behavioral condition that can be identified at any point in a person's life. This study investigated the potentiality of machine learning techniques such as Logistic Regression, Random Forest, Multinomial Naive Bayes (MNB), Bernoulli Naive Bayes (BNB), Support Vector Machine (SVM), and Gaussian Naive Bayes (GNB) to predict ASD using some health parameters. There are 292 instances and 21 attributes in the first dataset linked to the screening for ASD in children. The adult individuals in the second dataset had a total of 704 occurrences and 21 characteristics related to ASD detection. In order to achieve the highest accuracy possible from the machine learning models, feature optimization is used in this study along with other preprocessing approaches. The findings overwhelmingly support the notion that Random Forest performs better on all of these datasets, with the greatest accuracy (100%) for data on Autistic Spectrum Disorder (ASD) in children and adults, respectively.

Keywords Autism Spectrum Disorder (ASD); Machine Learning; Feature Optimization; Random Forest; Logistic Regression.

Network Biology
ISSN 2220-8879
URL: http://www.iaees.org/publications/journals/nb/online-version.asp
RSS: http://www.iaees.org/publications/journals/nb/rss.xml
E-mail: networkbiology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

ASD, a neurodevelopmental disorder, has a significant influence on many people's day-to-day life especially human interaction (Wing, 1997). The impacts of ASD symptoms degrade social and interaction skills (Vaishali, 2018). As per studies, 1 in 160 children in the world suffers from autism (World Health Organization, 2019). The signs of autism are more noticeable and easier to spot in children between the ages of two and three (Raj, 2020). Healthcare professionals created the testing procedures for ASDs of the Globally Different Type. Using a diagnostic tool like the Autism Diagnostic Interview (ADI), doctors could formally diagnose ASD. For those who already have ASD, early detection and treatments are more vital to reducing symptoms. Since there is no medical test used to diagnose autism, ASD symptoms are usually recognized by observation. In the past, parents and instructors have noticed ASD after a kid starts school. After that, a school's special education program was monitored and evaluated for any signs of ASD. The school staff recommended these kids visit a doctor for essential testing. Since ASD symptoms might overlap with those of other mental illnesses, adults have a lot of difficulties detecting them than children do. Monitoring a child makes it easy to notice behavioral changes in the child because it can be detected earlier than Autism-specific neuroimaging at 6 months of age and because imaging studies can identify the condition after 2 years of age. Machine learning algorithms are attempting to cut computational procedures that automatically find valuable patterns in vast volumes of data. These involve information theory, statistical forecasting, and mathematical learning. This technique can forecast non-experimental data sets with a large number of variables in a reliable and accurate manner (Saxe, 2017). Recent psychiatric research has shown that this approach is effective for diagnosing ASD (Wall, 2012), categorizing attention deficit disorder using altered event-related possibilities (Mueller, 2010), and classifying schizophrenia using free speech analysis. (Bishop et al., 2018) investigated the lifelong health problems of adult ASD patients using machine learning approach, and these methodologies correctly predicted health issues with the respiratory, urinary, and cardiovascular systems.

Several indications of autism: Although the precise etiology of autism spectrum disease is unknown, it is thought that genetic and environmental factors could both play a significant role (Dawson, 1998):

- Odd body expressions or expressions
- Abnormal vocal tone
- Poor eye interaction
- Disabilities of Behavior
- Deficiency of Linguistic Knowledge
- Slow in Talking Development
- Continuous Talk
- Unsuitable Social Interaction
- A Serious Focus on One Subject
- Lack of Sympathy
- Low level of awareness of social cues
- Avoiding playing with peers
- Preoccupation With Particular Subjects
- Would like to live alone
- Echo term usage etc.

Fig. 1 shows the behavior of ASD people. The type of behaviors has seen in ASD affected people normally some behavior has represented in figure (Zayed, 2022).

Additional challenge for those with ASD (Zayed, 2022):

- Repeating certain actions, including using the same words or phrases frequently.
- When a routine is about to change, the person will become angry.
- A slight interest in particular aspects of the subject, such as figures and data, etc.
- Less sensitive in some situations, such as those involving light, noise etc.

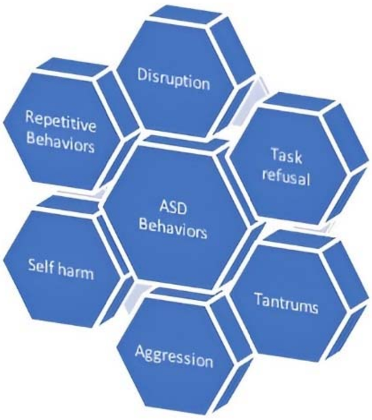
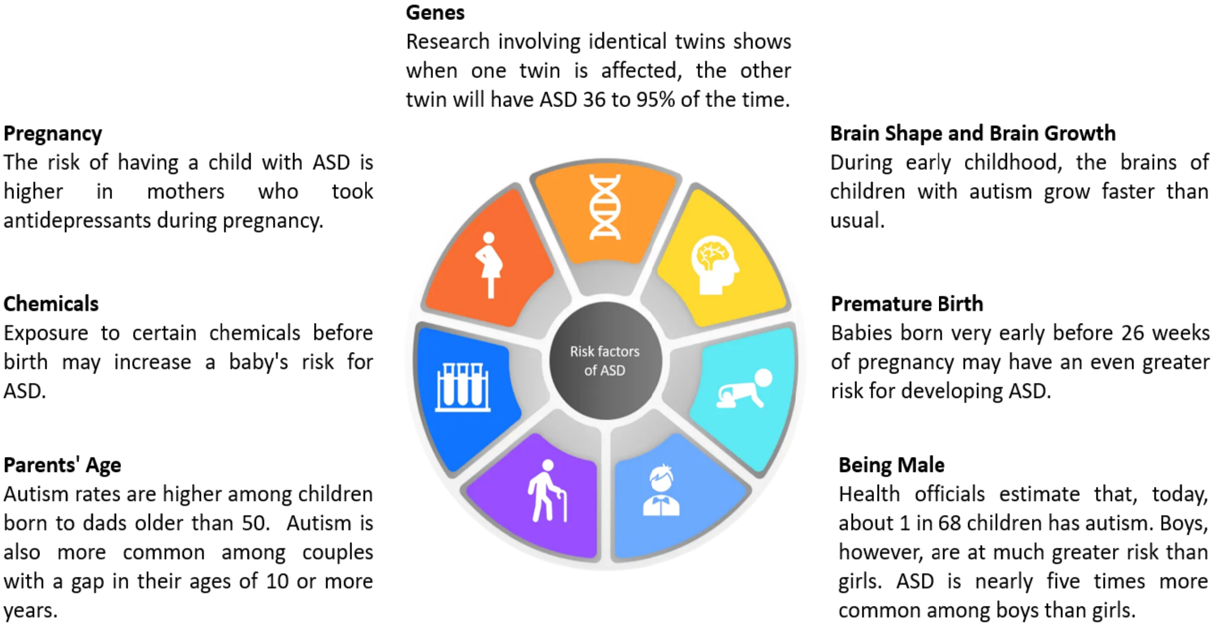


Fig. 1 Behavior of ASD people.

Fig. 2 depicts the risk factor for people with ASD. What causes autism spectrum disorder (ASD) is a mystery to medical professionals. It doesn't seem like there is just one cause. Instead, it's more likely that a number of variables will interact to raise a child's risk for this spectrum of diseases. A child may inherit some risk factors. The surroundings of a child may be important even before birth (Bladen, 2022).



Genes
Research involving identical twins shows when one twin is affected, the other twin will have ASD 36 to 95% of the time.

Pregnancy
The risk of having a child with ASD is higher in mothers who took antidepressants during pregnancy.

Brain Shape and Brain Growth
During early childhood, the brains of children with autism grow faster than usual.

Chemicals
Exposure to certain chemicals before birth may increase a baby's risk for ASD.

Premature Birth
Babies born very early before 26 weeks of pregnancy may have an even greater risk for developing ASD.

Parents' Age
Autism rates are higher among children born to dads older than 50. Autism is also more common among couples with a gap in their ages of 10 or more years.

Being Male
Health officials estimate that, today, about 1 in 68 children has autism. Boys, however, are at much greater risk than girls. ASD is nearly five times more common among boys than girls.

Fig. 2 Risk factor of ASD.

To recognize ASD symptoms, one typically needs some observers as well as an interview with a doctor and patient because ASD prediction does not have a clinical test like other disorders. This procedure takes a great deal of time, and with this kind of procedure, it is challenging to predict ASD at an early stage. The proposed research might reduce the amount of time needed to use artificial intelligence to detect ASD. Prior to this, little research was conducted on this specific aspect, in addition to the numerous ML techniques. In order to obtain more accurate results, the main goal of this suggested work has been to identify ASD utilizing a different sort of machine learning algorithm with feature optimization on data from both children and adults.

2 Related Works

The detection of autism spectrum disorder (ASD) using machine learning algorithms has been a popular research area in recent years. Many institutions and organizations used different techniques to further improve their ability to detect ASD as accurately as possible. Koemicki et al. (2015) developed a model to detect ASD using machine learning algorithm. The dataset was trained using ADOS modules 2 and 3. 98.90% sensitivity, 98.58% specificity, and 98.76% accuracy were produced by ADOS module 2 and 100% sensitivity, 98.92% specificity, and 99.83% accuracy were produced by ADOS module 3, respectively. This study is constrained by the content of these previous data sets because it focuses on the analysis of archival documents. Raya et al. (2020) studied and invented a technique to detect ASD by observing the body movement parameters of an imitation VR activity and using machine learning. They used Wilcoxon signed-rank to find the body joint significant differences, and the highest classification accuracy was 89.36%. The few participants in each group's sample is the main downside despite the favorable results. The ML approach might be validated in later research with greater sample numbers, allowing the model to be tested. Thabtah et al. (2020) proposed a model to detect ASD using machine learning algorithm. Based on the Covering methodology, they suggested a brand-new classification technique they named Rules-Machine Learning (RML). This approach provides rule sets that serve as automatic classification systems (classifiers). In comparison to PRISM, CART, AdaBoost, Bagging, Nnge, RIDOR, C4.5, and RIPPER algorithms, RML's classifier had lower errors rates of 4.41, 2.7, 0.15, 2.14, 3.7, 3.27, 1.57, and 1.83 percent, higher sensitivity rates of 1.9, 3.3, 2.0, 2.8, 3.2, 1.7, 0.2, and 1.7 percent, and higher specificity rates of 2.52, 3.49, and 1.55. One of the most significant difficulties in ASD screening research is enhancing the screening process to provide patients and their families with a speedier and more accurate service. Hossain et al. (2021) proposed a system to detect ASD using machine learning classification techniques. MLP and Logistic Regression (LR) classifiers exhibit 100% accuracy for the top ten attributes out of the 27 classification strategies they tested. To create a more reliable ASD diagnostic system, they intend to analyze brain signals (such as EEG) and tie them to AQ-based research. Mohanty et al. (2021) suggested a model to detect toddlers' ASD using machine learning classification techniques such as SVM, KNN, DT, DA, and RF. SVM scored best for classifying Toddler ASD among them, outperforming other classification models. It has 99% accuracy, 100% sensitivity and 98% specificity. Akter et al. (2019) created a strategy to identify ASD using an ASD screening dataset. The classifiers model predicted ASD with 98.77% accuracy; 99.98% AUROC; 99.39% sensitivity for SVM and 99.59% specificity for Adaboost classifier. The data sets' contents, which were available, restricted this investigation. The phenotype data used in this study came from repositories of publicly accessible autism research data, which only contain a small number of ADOS tests for non-ASD control subjects. Levy et al. (2017) developed a method to identify stable subsets of predictive variables for the behavioral detection of autism by scarifying machine learning models. They assessed the effectiveness of 17 various machine learning classifiers on these feature sets in order to accurately predict the diagnosis of ASD or non-ASD. They found that, depending on the task, logistic regression or SVM performed best on module 2 (AUC = 0.92 and 0.93), and that logistic regression with L2 regularization performed best on module 3 (AUC= 0.93). More study is required to employ neuroimaging data and apply ML and DL techniques to ASD research. Nogay et al. (2020) proposed a model to recognize autism spectrum disorder (ASD) through brain imaging with the aid of machine learning approaches. The DBN model with a depth of 3 and a combination of triple data (fMRI + GM + WM) has the highest classification accuracy rate (92%). Using machine learning algorithms with brain sMRI and fMRI data, automated diagnosis of ASD has not yet achieved the requisite degree of effectiveness. They are now unable to make a meaningful contribution to the early or quick identification of ASD. Jessey et al. (2019) developed a method to recognize developmental delay and autism using home footage of Bangladeshi toddlers. They experimented with various training-test splits and used Source Classifier to get a maximum AUC of 0.75. Because these original classifiers were trained on clinical score sheets rather than features gleaned from real-time video data, it is most likely that their relatively low accuracy results from this. Raj et al. (2020) demonstrated a method for diagnosing autism spectrum disorder using machine learning techniques such as Naive Bayes, Convolutional Neural Networks, Support Vector Machines, Logistic Regression, KNN, and Neural Networks. Following the use of several machine learning methods and missing value handling, the findings strongly imply that CNN-based on all of these datasets, prediction models perform more accurately 99.53%, 98.30%, and 96.88% of those screened positive for autism spectrum disorders in data for adults, kids, and teenagers, correspondingly. Alcañiz et al. (2022) proposed a model to identify autism spectrum disorder combining virtual reality and machine learning tools like SVM and KNN. Recursive feature selection was used to build a set of multivariate supervised machine learning models based on the obtained eye gaze data. The models identified autistic children with up to 86% accuracy (sensitivity = 91%). The sample size is small, there are a lot of features, and the models weren't used on a separate sample, which has some drawbacks, despite the fact that the results are encouraging. Erkan et al. (2019) proposed an approach to identify autism spectrum disorder using machine learning techniques. We employed the k-Nearest Neighbors approach (KNN), the Support Vector Machine method (SVM), and the Random Forests method to classify the ASD data for children, adolescents, and adults (RF). For all three datasets, the RF technique correctly categorized the data. Vakadkar et al. (2021) presented a model to identify children with autism spectrum disorder using machine learning approaches. Their dataset has been subjected to the use of models including Support Vector Machines (SVM), Random Forest Classifier (RFC), Naive Bayes (NB), Logistic Regression (LR), and KNN. The best accuracy was found to be provided by logistic regression (96%). The primary roadblock to this study's success is the dearth of significant, open-source ASD datasets. A substantial dataset is necessary in order to build a precise model. There weren't enough cases in the dataset used for this investigation. Liao et al. (2022) presented a technique to identify children with autism spectrum disorder by applying machine learning technology. In this work, a machine learning method was put forth to identify children with ASD by combining physiological (electroencephalography, or EEG) and behavioral (eye fixation and facial expression) data. They used hybrid modality fusion based on a weighted naive Bayes algorithm to achieve the greatest accuracy of 87.50%. Nevertheless, there are still limitations despite our encouraging discoveries and their prospective applications. The number of samples used was quite small.

3 Methodology

In the raw data section of this work, there are two types of data sets. One set of data is for adults, while the other is for children. Then both data sets need to be processed to get better results. After preprocessing, there should be feature optimization for both data sets according to this proposed workflow. After preprocessing then should be feature optimization done from both data sets according to this proposed workflow. When features are optimized, various types of machine learning algorithms are used to find the best possible classifications and various types of machine learning algorithms were used to find the best possible classifications. At last, this proposed workflow has seen the performance of the best classifications or machine learning algorithms. Fig. 3 shows the proposed workflow of the study. This figure shows all of the workflows from data preprocessing to performance analysis.

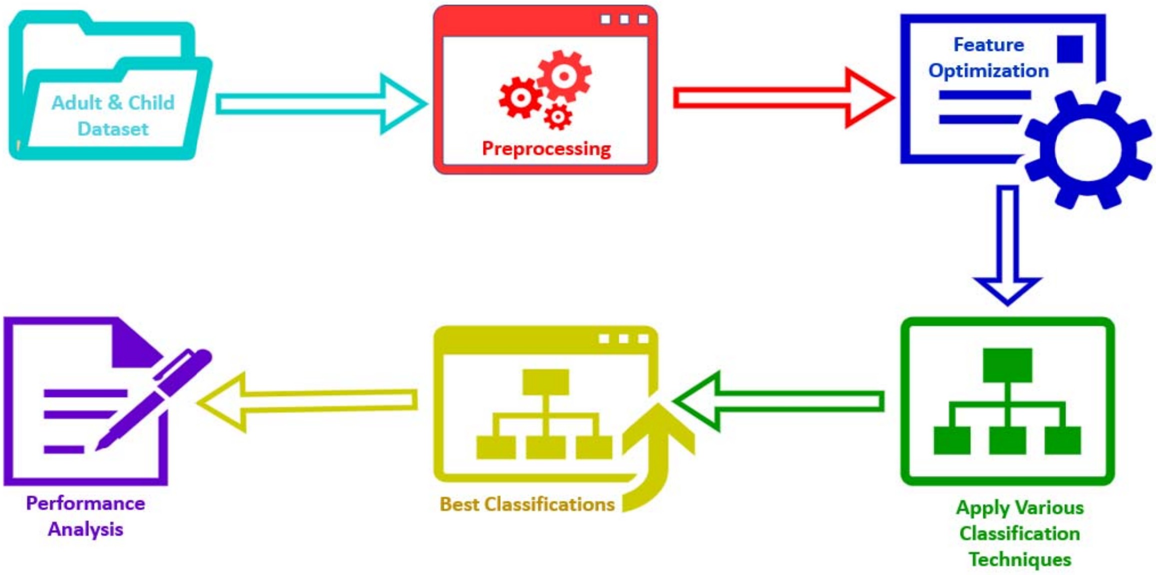


Fig. 3 Work flow of the proposed system.

3.1 Data set

The publicly accessible UCI Repository was used to collect the dataset for this study (Thabtah, 2017). Two different types of datasets have mostly been employed in this study. Table 1 shows the summary of ASD datasets for both children and adults.

Table 1 List of ASD datasets.

Sr. No.	Dataset Name	Sources	Attribute Type	Number of Attributes	Number of Instances
1.	ASD Screening Data for Adult	UCI Repository	Categorical, continuous and binary	21	704
2.	ASD Screening Data for children	UCI Repository	Categorical, continuous and binary	21	292

There are 20 common attributes in these datasets that are used for prediction. These attributes are listed below in Table 2.

Table 2 List of Attributes in the dataset.

Attribute Id	Attributes Description
1	Age
2	Sex
3	Nationality
4	The patient was born with a jaundice condition.
5	Every member of the family had pervasive developmental problems.

6	Who is doing the study?
7	Country
8	Screening Inquiry
9	Type of screening test
10-19	responses to the ten screening questions
20	Screening Score

Heat Map: A heat map is a type of data visualization that uses color to represent a phenomenon's size in two dimensions. The reader is helped by the color variation to visualize how the phenomena clusters or fluctuates in space. In the table, the first dimension's values are shown as rows, and the second dimension's values are shown as columns. The percentage of measurements that match the dimensional value determines the cell's color. The pattern highlights distinctions and variations within the same set of data. Warmer colors in Figs 4(a) and 4(b) represent greater values, whereas colder colors indicate lower values. Fig. 4(a) shows the heat map for the child data set, whereas Fig. 4(b) shows the heat map for the adult data set.

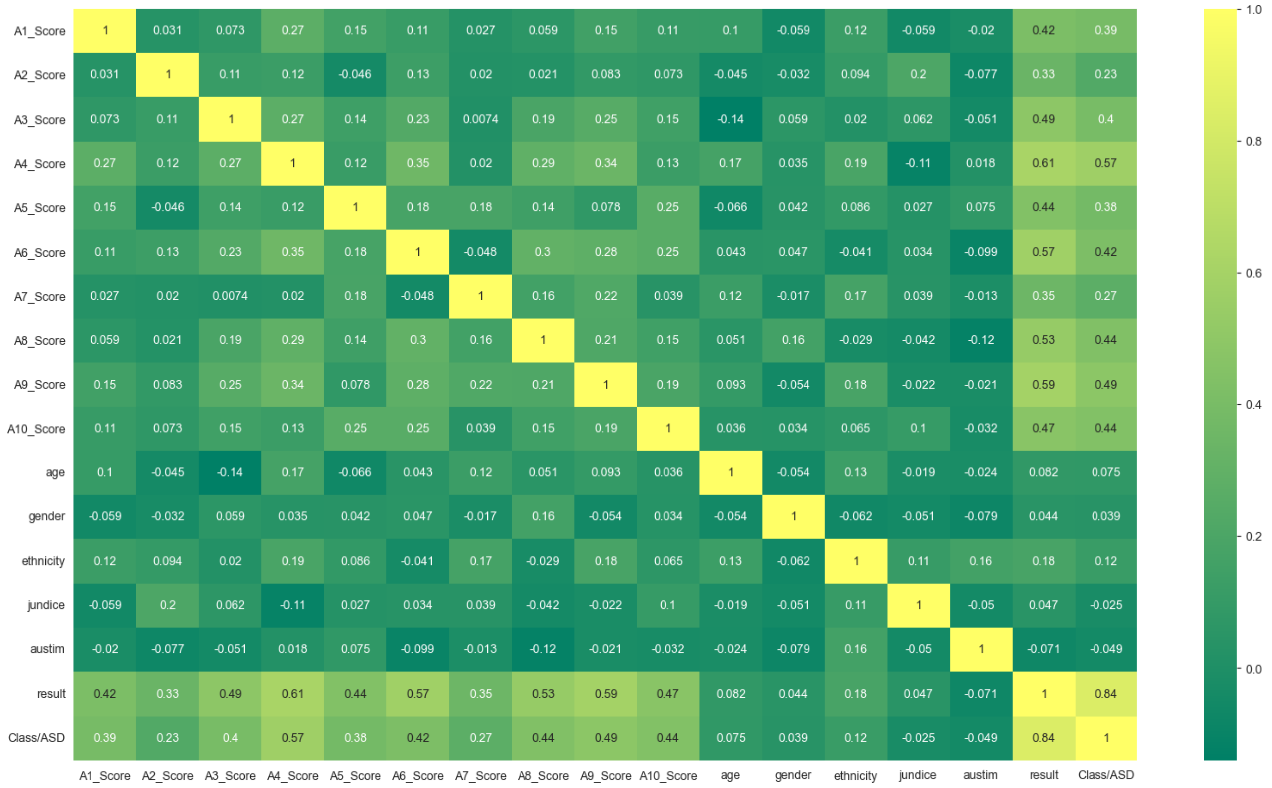


Fig. 4(a) Heat Map for children.

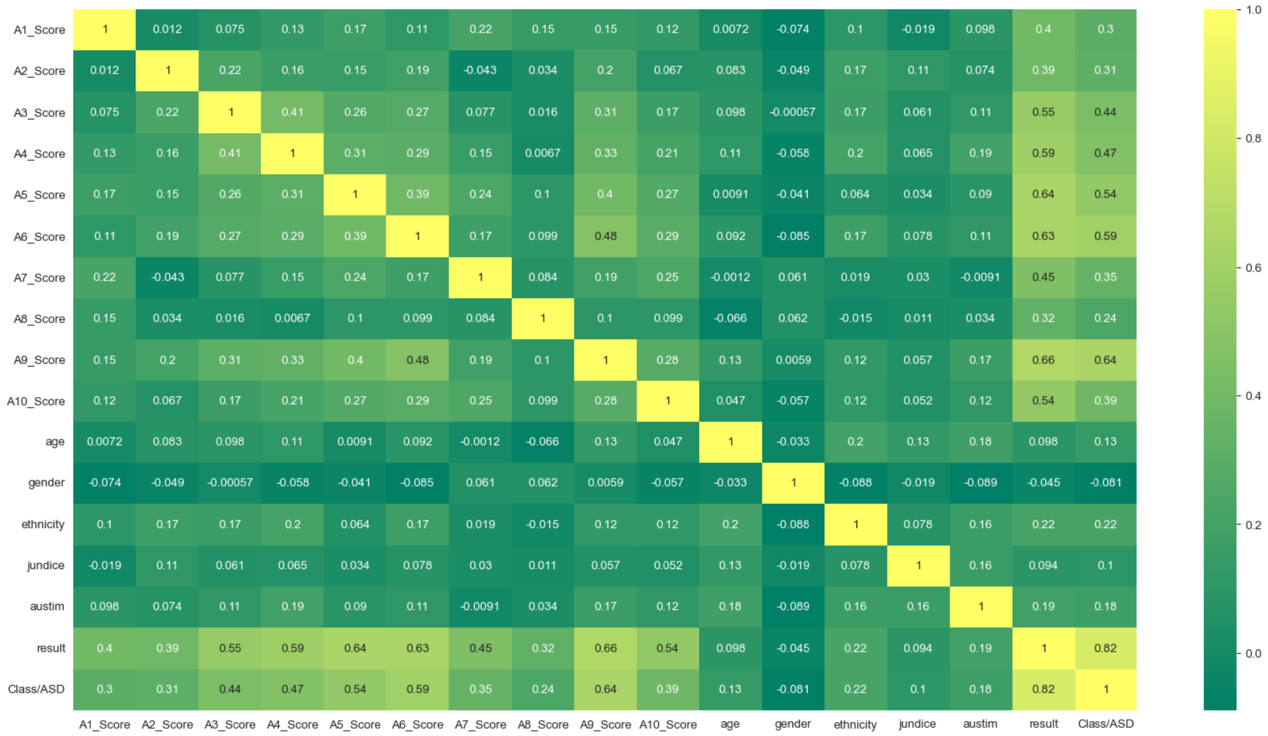


Fig. 4(b) Heat Map for Adults dataset.

3.2 Data preprocessing

Data preparation is a method that turns the raw data into a format that is useful and clear. Real-world data usually has errors and null values, which causes it to be inconsistent and incomplete. A great result is always produced by well-preprocessed data. A range of data preprocessing methods, such as dealing with missing values, outlier identification, data discretization, data reduction (dimension and numerosity reduction), etc., can be used to deal with incomplete and inconsistent data. This work has applied the mean method for replacing the age column and also used a label encoder to transform categorical data into numerical data. Then feature engineering is used to remove unnecessary data.

3.2.1 Label encoder

Label encoding is a common encoding technique that can handle category information (Sailasya, 2021). This approach assigns a unique number based on alphabetical order to every label. Equation 1 helps to perform level encoding.

$$S = \frac{1}{1 + \exp\left(-\frac{n - mdl}{a}\right)} \tag{1}$$

3.2.2 Missing values

Missing data or a missing value refers to data that is not stored for some variables or participant information (Mohammed, 2020). Equation 2 helps to sort out the problem of missing values.

$$(\boldsymbol{X}) = \boldsymbol{A} + \frac{\sum \boldsymbol{fd}}{\boldsymbol{N}} \times \boldsymbol{C} \tag{2}$$

3.2.3 Feature engineering

Feature engineering is the process of selecting, manipulating, and producing features from raw data that can be used in supervised learning. After using feature selection, some columns in this section were removed based on the feature selection score (Li, 2017). Country_of_Res, Relation, Age_desc, and Used_app_before are the columns.

$$z = \frac{x_{i-\mu}}{\sigma} \tag{3}$$

4 Classification Techniques

To complete this task, 6 machine learning algorithms are used. In this section, the used algorithms are fully described.

4.1 Gaussian Naive Bayes (GNB)

A simple prediction model is Naive Bayes. Naive Bayes (Gaussian) assumes that each class has a Gaussian distribution. Naive Bayes (Gaussian) implies feature independence, hence the covariance matrices are diagonal matrices. This is how QDA differs from Naive Bayes (Ontivero-Ortega, 2017).

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma^2{}_y}} \exp\left(-\frac{\left(x_{i-\mu_y}\right)^2}{2\sigma^2 y}\right) \tag{4}$$

4.2 Multinomial Naïve Bayes (MNB)

The most widely used probabilistic learning method in Natural Language Processing (NLP) is the Multinomial Naive Bayes algorithm. The Bayes theorem serves as the foundation for the system that anticipates the tag of a text, such as an email or newspaper article. For a given sample, it determines the probabilities of each tag, and then outputs the tag with the highest probability (Kibriya, 2004).

$$P(A|B) = P(A) * P(B|A)/P(B) \tag{5}$$

where, P(B) = prior probability of B, P(A) = prior probability of class A, P(B|A) =the chance of predictor B occurring given class A

4.3 Bernoulli Naive Bayes (BNB)

This uses the discrete data set and the Bernoulli distribution. Only binary values—such as true or false, yes or no, success or failure, 0 or 1 are acceptable for features in Bernoulli Naive Bayes, which is their main distinguishing feature (Ponte, 2017; Rahman, 2022).

$$\mathrm{P}(x_i|y) = P(i|y)x_i + \left(1 - p(i|y)\right)(1|x_i) \tag{6}$$

4.4 Support Vector Machine (SVM)

Closer to the hyperplane data points, called support vectors, influence the position and direction of the hyperplane. These support vectors are used to boost the classifier's margin. If the support vectors are eliminated, the position of the hyperplane will change (Noble, 2006). These are the concepts that support the growth of our SVM.

$$Q(a) = \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j d_i d_j \emptyset(x_i)\,\emptyset(x_j) \qquad (7)$$

4.5 Quadratic discriminant analysis (QDA)

Linear discriminant analysis (LDA) is quite popular since it can be used as both a classifier and a dimensionality reduction method. Quadratic discriminant analysis is a version of LDA that permits non-linear separation of data (QDA) (Wu, 1996).

$$In\,P(t|\theta) = In \prod_{n-1}^{N}\prod_{k-1}^{K} t_n k\left(In\pi_k + In\left(\frac{1}{\sqrt{(2\pi)^D det \sum c}} exp\left(-\frac{1}{2}(x_n - p_c)^t \sum_{c}^{-1}(x_n - \mu_c)\right)\right)\right) \qquad (8)$$

4.6 Random Forest (RF)

As a result of avoiding the issue of overfitting, a random forest model outperforms a decision tree. Decision trees of numerous types, each slightly different from the others, make up models based on random forests. Based on each decision tree model, the ensemble uses the majority voting procedure to create predictions (bagging). As a result, less overfitting occurs while yet keeping each tree's capacity for prediction. In Random Forest, a forest or forests are created by using several trees, and each tree is then continuously examined (Belgiu, 2016). Use the following equation to calculate the Gini Index for the classification:

$$Gini = 1 - \sum_{i-1}^{n}(p_i)^2 \qquad (9)$$

The probability that the object will be categorized into a certain class or feature is represented by the value of p_i in equation 9.

5 Result and Discussion

Through the use of a confusion matrix and classification report, the outcome is evaluated in terms of accuracy, precision, recall, and f-measure. The outcome is determined by how precisely the model was trained.

5.1 Confusion matrix

To explain how effectively a categorization system performs, a confusion matrix is utilized. In a confusion matrix, the results of a classification algorithm are displayed and condensed (Dey, 2022; Biswas, 2022).

Table 3 Elements of confusion matrix.

		Actual Values	
		Positive (1)	Negative (0)
Predicted values	Positive (1)	TP	FP
	Negative (0)	FN	TN

$$\text{Accuracy} = \frac{TP+TN}{(TN+TP+FP+FN)} \quad (10)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (11)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (12)$$

$$\text{F1} = 2\frac{Precision.Recall}{Precision+Recall} \quad (13)$$

For data from ASD screening for adults and children’s, the experimental findings of several machine learning approaches with all features selected have been presented. To evaluate the precision, recall, f1 measure, and accuracy of the predicted model, all 21 features were chosen. Table 4 showed the result for adults screening data for ASD after optimization.

Table 4 After optimization, the overall findings for the adult autistic spectrum disorder screening data.

Classifier	Accuracy	Precision	Recall	F1 score
GNB	.96	.94	.91	.93
BNB	.95	.89	.91	.90
MNB	.85	.76	.56	.64
SVM	.84	1	.35	.52
Random F	**1**	**1**	**1**	**1**
QDA	.95	.94	.85	.89

After optimization, the ASD adult diagnosis dataset's machine learning models' accuracy ranged from .84 to 1.

Table 5 showed the result for children screening data for ASD after optimization.

Table 5 After optimization, the overall findings for the children autistic spectrum disorder screening data.

Classifier	Accuracy	Precision	Recall	F1 score
GNB	.93	1	.88	.98
BNB	.93	.94	.94	.94
MNB	.76	.82	.72	.77
SVM	.64	1	.34	.51
Random F	1	1	1	1
QDA	.54	.54	.5	.70

After optimization, the ASD child diagnosis dataset's machine learning models' accuracy ranged from .54 to 1.

5.2 Roc curves

The relationship or trade-off between clinical sensitivity and specificity for each conceivable cut-off for a test or set of tests is usually shown graphically using ROC curves. Additionally, the ROC curve's area under the curve shows how useful the test or tests in question are. Fig. 5 (a) represented the roc curves for adults' data set after optimization and Fig. 5(b) represented the roc curves for children data set after optimization. The genuine positive rate is shown on the y-axis in all four ROC curves, while the false positive rate is shown on the X-axis.

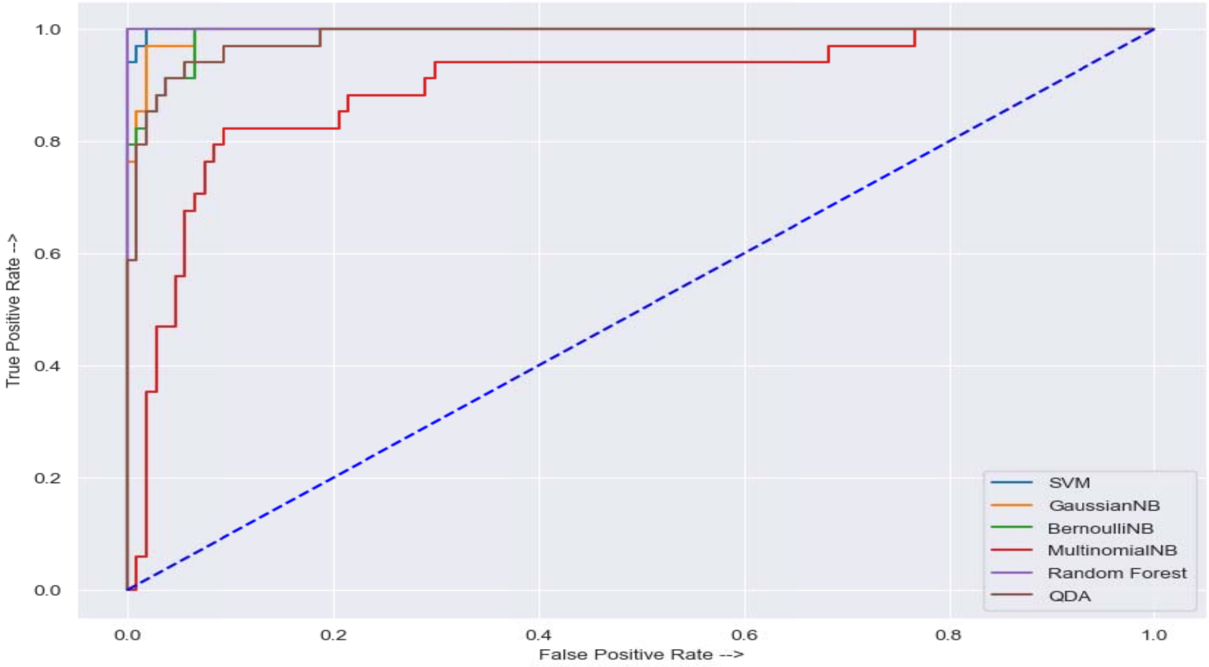


Fig. 5 (a) Roc Curve for adult’s data set after optimization.

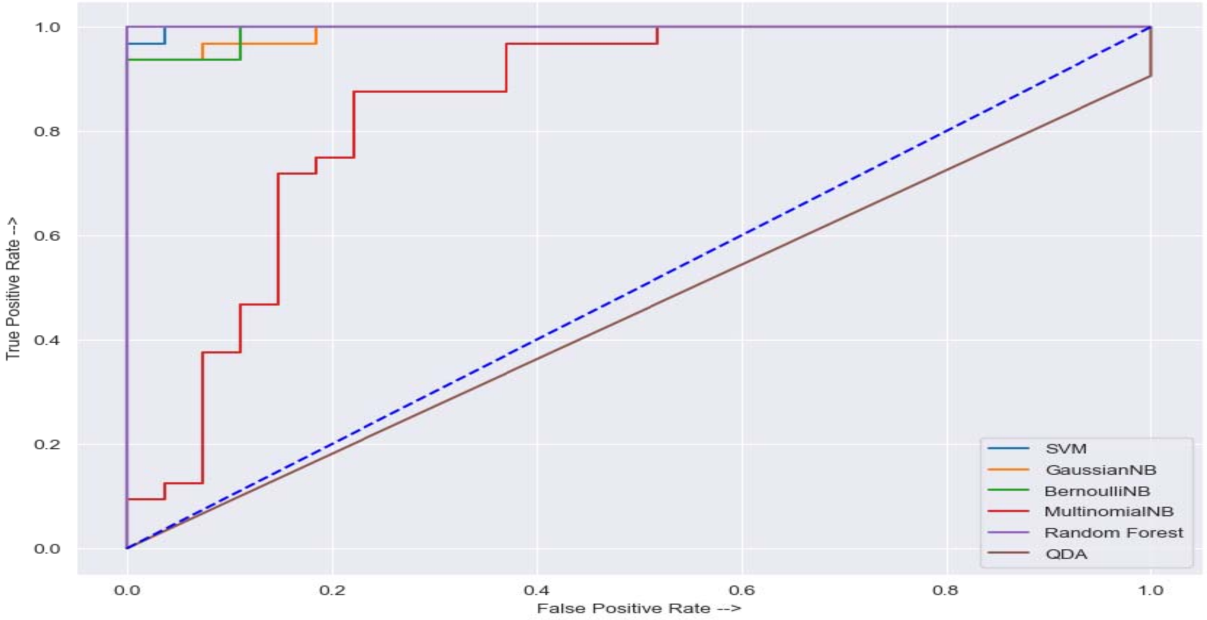


Fig. 5(b) Roc Curves for children data set after optimization.

5.3 Accuracy comparison according to machine learning algorithm

The accuracy comparison chart represents the accuracy according to the machine learning algorithm that is used in this study. The blue color columns represent the accuracy before optimization and red color columns represent the accuracy after optimization for the child data set. Fig. 6 represented the accuracy comparison according to machine learning algorithm for children data set and adult's data set. Table 6 presents the comparison of the proposed work with other existing works.

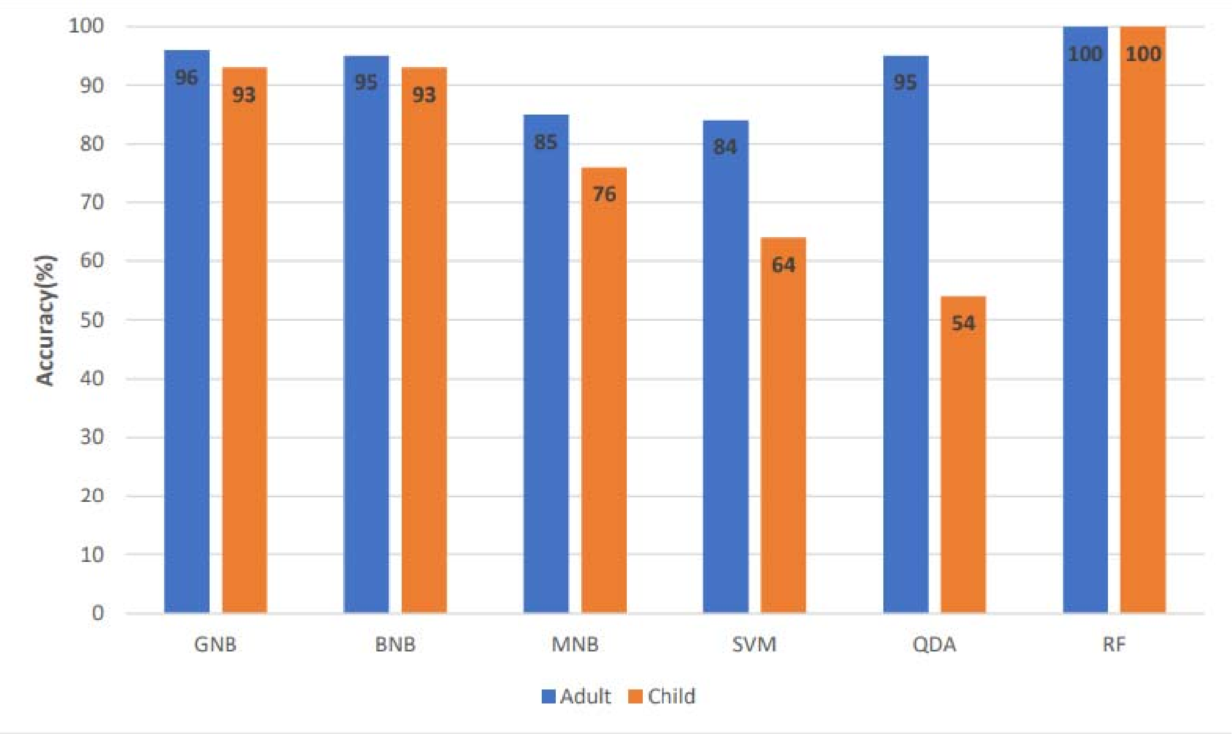


Fig. 6 Accuracy comparison for children data set and adult's data set after optimization.

Table 6 Comparison with the existing work for ASD prediction using ML.

Reference	Algorithm	Proposed Method	Accuracy
Suman Raj et al. [3]	SVM	Machine learning	99.53%
J. A. Koemicki et al [11]	LibSVM*	Machine learning	99.83%
Mariano Alcaniz Raya et al [12]	Not mentioned Algorithm name on this paper	Machine learning	89.36%
Ashima Sindhu Mohanty et al [15]	SMV	Machine learning	99%
Tania Akter et al [16]	SVM	Machine learning	98.77%
HidirSelcukNogay et al [18]	SVM	Machine learning	92%
Mariano Alcañiz et al [21]	SVM, KNN	Machine learning	86%
Kaushik Vakadkar et al [23]	Logistic Regression (LR)	Machine learning	96%
Mengyi Liao et al [24]	Naive Bayes	Machine learning	87.50%
Our Proposed Model	Random Forest	Machine learning	100%

6 Conclusion

In this experiment, a variety of machine learning approaches were used to try and detect autism spectrum disorder. Different performance assessment metrics were used to analyze the performance of the models built for ASD identification after optimizing records from two sets of age groups. One data set is for children's and another is for adult. The Random Forest, classifiers performed better than SVM, MNB, GNB, BNB and QDA with all of its features' characteristics included after managing missing data when compared to the results of another recent works. For both the children's and adult's data sets in this work, Random Forest, attained the greatest accuracy (100%). As well as GNB and BNB achieved up to 92% accuracy for both children's and adults. These findings strongly suggest that, rather than the other typical machine learning classifiers described in previous studies, a model based on random forest can be used to detect autism spectrum disorder..

The main contribution of this study is that recently published papers have shown the best model according to the SVM to predict ASD. However, this paper demonstrated that the RF model outperforms the SVM model in terms of maximum accuracy (100%). So, in this study strongly suggested to build a RF based model to predict ASD as early-stage detection.

References

Akter T, Satu MS, Khan MI, Ali MH, Uddin S, Lio P, Quinn JM, Moni MA. 2019. Machine learning-based models for early-stage detection of autism spectrum disorders. IEEE Access, 7: 166509-166527

Alcaniz Raya M, Marín-Morales J, Minissi ME, Teruel Garcia G, Abad L Chicchi Giglioli IA. 2020. Machine learning and virtual reality on body movements' behaviors to classify childrenwith autism spectrum disorder. Journal of Clinical Medicine, 5: 1260

Alcañiz M, Chicchi-Giglioli IA, Carrasco-Ribelles LA, Marín-Morales J, Minissi ME, Teruel-García G, Sirera M, Abad L. 2022. Eye gaze as a biomarker in the recognition of autism spectrum disorder using virtual reality and machine learning: A proof of concept for diagnosis. Autism Research, 15(1):131-145

Bladen M, Thorpe N, Ridout D, Barrie A, McGibbon E, Mance A, Watson L, Main E. 2022. Autism Spectrum Disorders in boys at a major UK hemophilia center: prevalence and risk factors. Research and Practice in Thrombosis and Haemostasis, 00013

Belgiu M, Drăguţ L. 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114: 24-31

Bishop-Fitzpatrick L, Movaghar A, Greenberg JS, Page D, DaWalt LS, Brilliant MH, Mailick MR. 2018. Using machine learning to identify patterns of lifetime health problems in decedents with autism spectrum disorder. Autism Research, 11(8): 1120-1128

Biswas N, Uddin KMM, Rikta ST, Dey SK. 2022. A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach. Healthcare Analytics, 2: 100116

Dawson G, Meltzoff AN, Osterling J, Rinaldi J. 1998. Neuropsychological correlates of early symptoms of autism. Child Development, 69(5): 1276-1285

Dey SK, Uddin KMM, Babu HMH, Rahman MM., Howlader A, Uddin KA. 2022. Chi2-MI: A hybrid feature selection based machine learning approach in diagnosis of chronic kidney disease. Intelligent Systems with Applications, 16: 200144

Erkan U, Thanh DN. 2019. Autism spectrum disorder detection with machine learning methods. Current Psychiatry Research and Reviews Formerly: Current Psychiatry Reviews, 15(4): 297-308

Hossain MD, Kabir MA, Anwar A, Islam MZ. 2021. Detecting autism spectrum disorder using machine learning techniques. Health Information Science and Systems, 9(1): 1-13

Kibriya AM, Frank E, Pfahringer B, Holmes G. 2004. December. Multinomial naive bayes for text categorization revisited. In: Australasian Joint Conference on Artificial Intelligence. 488-499, Springer, Berlin, Heidelberg

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

Levy S, Duda M, Haber N,Wall DP. 2017. Sparsifying machine learning models identify stable subsets of predictive features for behavioral detection of autism. Molecular Autism, 8(1): 1-17

Li Z, Ma X, Xin H. 2017. Feature engineering of machine-learning chemisorption models for catalyst design. Catalysis Today, 280: 232-238

Liao M, Duan H, Wang G. 2022. Application of machine learning techniques to detect the children with autism spectrum disorder. Journal of Healthcare Engineering, 2022.

Mueller A, Candrian G, Kropotov JD, Ponomarev VA, Baschera GM. 2010. Classification of ADHD patients on the basis of independent ERP components using a machine learning system. In Nonlinear biomedical physics, 4(1): 1-12

Mohanty AS, Patra KC, Parida P. 2021. Toddler ASD classification using machine learning techniques. International Journal of Online and Biomedical Engineering, 17(7)

Mohammed R, Rawashdeh J, Abdullah M. 2020, April. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In: 2020 11th international conference on information and communication systems (ICICS). 243-248, IEEE

Noble WS. 2006. What is a support vector machine? Nature Biotechnology, 24(12): 1565-1567

Nogay HS, Adeli H. 2020. Machine learning (ML) for the diagnosis of autism spectrum disorder (ASD) using brain imaging. Reviews in the Neurosciences, 31(8): 825-841

Ontivero-Ortega M, Lage-Castellanos A, Valente G, Goebel R, Valdes-Sosa M. 2017. Fast Gaussian Naïve Bayes for searchlight classification analysis. Neuroimage, 163: 471-479

Ponte JM, Croft WB. 2017, A language modeling approach to information retrieval. In: ACM SIGIR Forum. 51(2): 202-208, ACM, New York, USA

Rahman MM, Rana MR, Alam MNA, Khan MSI, Uddin KMM. 2022. A web-based heart disease prediction system using machine learning algorithms. Network Biology, 12(2): 64-80

Raj S, Masood S. 2020. Analysis and detection of autism spectrum disorder using machine learning techniques. Procedia Computer Science, 167: 994-1004

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

Saxe GN, Ma S, Ren J, Aliferis C. 2017. Machine learning methods to predict child posttraumatic stress: a proof of concept study. BMC Psychiatry, 17(1): 1-13

Sailasya G, Kumari GLA. 2021. Analyzing the performance of stroke prediction using ML classification algorithms. International Journal of Advanced Computer Science and Applications, 12(6)

Tariq Q, Fleming SL, Schwartz JN, Dunlap K, Corbin C, Washington P, Kalantarian H, Khan NZ, Darmstadt GL, Wall DP. 2019. Detecting developmental delay and autism through machine learning models using home videos of Bangladeshi children: Development and validation study. Journal of Medical Internet Research, 21(4): e13822

Thabtah, Fadi Fayez. 2017. Autistic Spectrum Disorder Screening Data for Adult.

Thabtah F, Peebles D. 2020. A new machine learning model based on induction of rules for autism detection. Health informatics journal, 26(1), pp.264-286.

Vakadkar K, Purkayastha D, Krishnan D. 2021. Detection of autism spectrum disorder in children using machine learning techniques. SN Computer Science, 2(5): 1-9

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

Wing L. 1997. The autistic spectrum. The Lancet, 350(9093): 1761-1766

Wall DP, Kosmicki J, Deluca TF, Harstad E, Fusaro VA. 2012. Use of machine learning to shorten observation-based screening and diagnosis of autism. Translational psychiatry, 2(4): e100-e100

Wu W, Mallet Y, Walczak B, Penninckx W, Massart DL, Heuerding S, Erni F. 1996. Comparison of regularized discriminant analysis linear discriminant analysis and quadratic discriminant analysis applied to NIR data. Analytica Chimica Acta, 329(3): 257-265

Zayed KMES, Ibrahim MA, Farid MN, El-Shourbagy OESO, Tarkan RS. 2022. Zinc Levels Assay in Children with Autism Spectrum Disorder by Quantum Magnetic Resonance Analyzer and Direct Colorimetry. Europe PMC, preprint