Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques

Srilakshmi Bai Rasaputhra
Department of Computer Science and Engineering
Narasaraopeta Engineering college
Narasaraopet, Anshra Pradesh, India
rsrilakshmi53@gmail.com

Navya Sri Thogati
Department of Computer Science and Engineering
Narasaraopeta Engineering college
Narasaraopet, Anshra Pradesh, India
navyasrithogati@gmail.com

Sakhitha Penugonda
Department of Computer Science and Engineering
Narasaraopeta Engineering college
Narasaraopet, Anshra Pradesh, India
sakhithapenugonda2206@gmail.com

Abstract

A person with autism spectrum disorder (ASD) may experience lifetime difficulties with language acquisition, communication, cognition, and social skills. It affects around 1% of the world's population and often manifests as developmental problems in the first two years following birth. Although environmental or genetic factors are the primary causes of ASD, early detection and treatment can ameliorate the condition. Currently, the only techniques available for diagnosing ASD are clinical standardized testing. This results in longer diagnosis times as well as a sharp rise in medical expenses. In order to increase the accuracy and time needed for diagnosis. The traditional approaches are being supplemented by machine learning techniques. Based on the results, we built predictive models using models like Support Vector Machines (SVM) with 97.2%, Random Forest Classifier (RFC) with 100%, Naïve Bayes (NB) with 97.32%, Logistic Regression (LR) with 100%, and KNN with 96.86%. In order to expedite the diagnostic procedure, our paper's primary goal is to ascertain whether the kid is prone to ASD at an early age. For the dataset we chose, logistic regression has the best accuracy, according to our findings.

Keywords—Autisms spectrum disorder, Dataset, Preprocessing, Encoding, SVM, KNN, Random Forest, Logistic Regression, Precision, Recall, F1 score and Accuracy.

I. Introduction

One serious condition that might make it difficult for a person to interact or communicate with others is autism spectrum disorder. It shows up in a person when they are still developing

A neuro-developmental disease, autism spectrum disorder (ASD) often manifests in the first three years of life in humans [1]. Basically, it is characterized by a number of symptoms, including communication and social interaction difficulties, interest restrictions, and repetitive behaviour [2]. People with ASD have trouble comprehending the thoughts and feelings of others. They have a lot of difficulty interacting with people. The World Health Organization (WHO) estimates that 1 in 270 people worldwide suffer with ASD [3]. Everyperson with ASD is different, and some have remarkable visual, intellectual, and musical ability. In this instance, the most crucial actions are needed to identify ASD and make sure that the right care is provided as soon as feasible. By doing these actions, you may lessen the consequences of the sickness and enhance their condition. Various kinds of observations are used to identify the symptoms associated with ASD. When it comes to effective therapy for ASD patients, early diagnosis does, however, need a substantial investment of time and energy. These days, a lot of people utilize machine learning techniques to analyse the symptoms of a variety of serious illnesses, such as diabetes, cancer, heart disease, and so on. In order to more accurately reduce the effects of ASD and identify ASD patients, several researchers have investigated a variety of techniques [4, 5]. This paper aimed to develop a machine learning model that investigates ASD in young children and toddlers.

II. *LITERATURE SURVEY*

Al Banna et al. [6] in order to help ASD patients deal with the COVID-19 epidemic, used a customized AI-based system for monitoring and assistance. Muhammad Hanif Ali [9] in addition to identifying ASD symptoms, presented a novel machine learning approach called Rules Machine Learning (RML), which provides users with a knowledge base of rules for comprehending the underlying causes of the categorization. Uddin MJ et al. [7] to gather information on ASD in babies, kids, teens, and adults created the smartphone application ASD Tests

.

Akter et al. [8] to gather information about ASD in toddlers, kids, teens, and adults, created the smartphone application ASD Tests. This software was developed to determine whether or not a user has ASD using the Q-CHAT and AQ-10 tools. Using this app, they gathered ASD data, which they then uploaded to the University of California-Irvine's (UCI) Machine Learning (ML) portal. Hossain et al. [9] similar dataset types were examined by, who also created subsets using the CFS, CHI, IG, One-R, and Relief-F techniques.

Raj et al. [10], SVM, LR, NB, and Convolutional Neural Network (CNN) were used examine these datasets (i.e., omitting toddlers). CNN demonstrated the best accuracy, with 98.30% for children, 96.88% for adolescents, and 99.53% for adults, respectively. Thabtah et al. [11] for the purpose of extracting ASD features, created a Rules-based Machine Learning (RML) algorithm, which outperformed previous machine learning techniques in terms of predicting accuracy.

Akter et al. [12] the ASD dataset was collected by at various age ranges, and several modified subsets were produced. They used a number of classifiers to examine them, and LR retrieved important attributes and outperformed the others. Chowdhury et al. [13] an association classification strategy with seven algorithms was offered by, and this method demonstrated 97% accuracy in detecting ASD.Akter

et al. [14] also used k-means algorithms to uncover many categories of autism. Next, they determined which of them was discriminating. Omar et al. [15] in order to develop an efficient machine learning model, used Random Forest (RF), Classification and Regression Trees (CART), and Random Forest-Iterative Dichotomies 3 (ID3) to evaluate 250 actual datasets and the AQ-10.

Sharma et al. [16] the CFS-greedy stepwise feature selector was used by to analyse these datasets. Naïve Bayes (NB), Stochastic Gradient Descent (SGD), K-Nearest Neighbours (KNN), Random Tree (RT), and K-Star (KS) were also used to these datasets. Satu et al. [17] using a variety of tree-based classifiers, gathered samples of children aged 16 to 30 and identified various rules for both normal and autistic development.

III. *DATASET DESCRIPTION*

The dataset contains [20] with 1054 instances and 18 attributes based on the Quantitative Checklist for Autism in children. The features in the dataset tells us things like age, ethnicity, chils born with ASD and Q-chart.

Attributes are:
Case_No, Age_Mons, Family_mem_with_ASD, Sex, Ethnicity, Qchat-10-Score, Jaundice, Who finished the test, ASD traits by class? Toddlers (Q-CHAT) screening approach; Q-CHAT-10 is a condensed form. The responses are converted into binary values that represent class types: "Yes" is given if the score is higher than three, and "No" is given if no signs of an ASD are discovered.

IV. *Methodology*

The main objective of this research is to employ machine learning methodologies for predicting Autism in children.

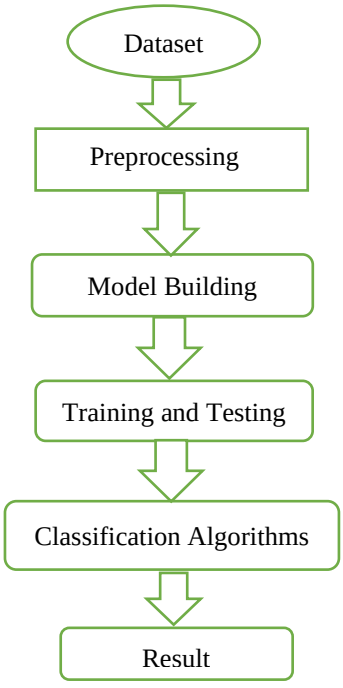

Fig. 1. Proposed System

A . Data preprocessing

Data preprocessing is crucial for ensuring the dataset's quality and relevance in predicting autism in Childrens.

a. Data Cleaning:

We deal with missing values to improve its suitability for training and analysis by cleaning up unprocessed or noisy data. The non-contributing characteristics, such as "Case_No," and "Who completed the test" columns were eliminated from dataset.

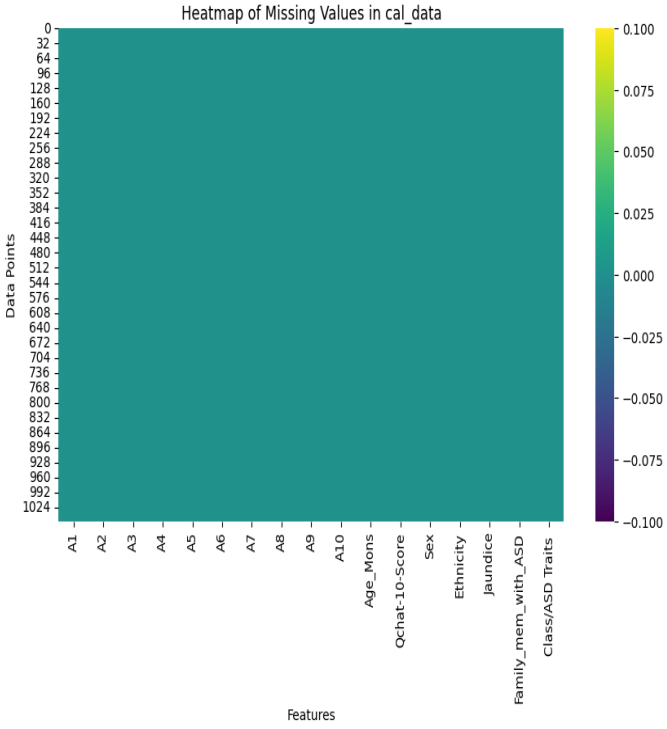

Fig. 2. Heatmap for Missing values

In Fig 2 we take features on x-axis and data points on y-axis. The map identify missing values with high concentration . The above figure does not contain any missing values.

b. Outliers:

Data points that substantially deviate from the bulk of observations in a dataset are outliers**.** In my study that I used interquartile range (IQR) is used to detect the outliers in my dataset. The IQR which extends from the first quartile (Q1) to the third quartile (Q3).

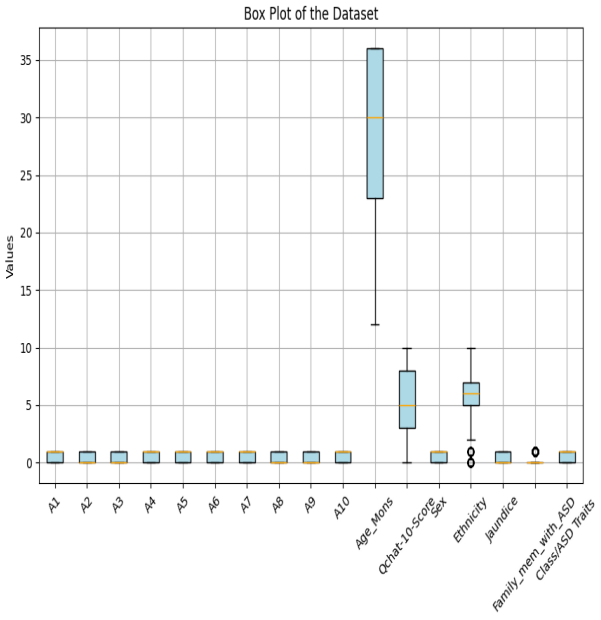

Fig. 3. Identifying the outliers

In Fig 3 shows that all features data points are grouped around the median, with very few outliers is shown in box plot. This range is usually 1.5 times the IQR by default. Any data points beyond the whiskers are considered as outliers.

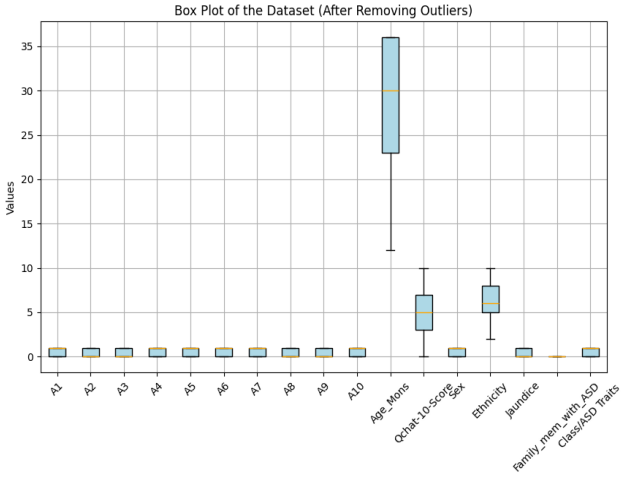

Fig. 4. After removing outliers

In Fig 4 shows the Box plot of data points after removing outliers. The data points which are beyond the median are removed.

c. Visualization

The model is overfitting the data if it has a low training error but a high testing error. Conversely, a large training and testing error indicates that the model is underfitting the data. A decent model will not cause the data to be overfit or underfit.

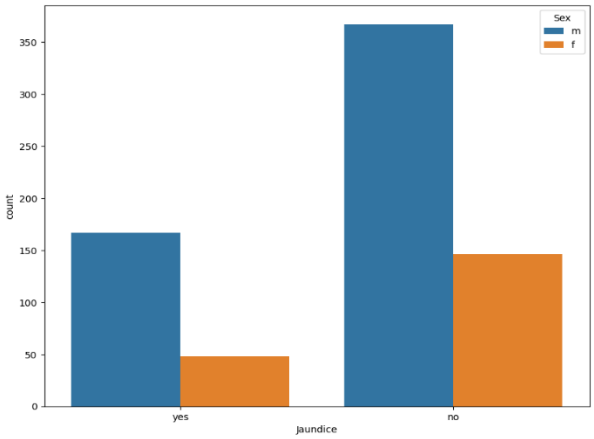

Fig. 5. The children born with jaundice having ASD positive based gender

One in every 68 children between the ages of two and three has autism, according to the gender distribution graph in Figure 5, which indicates that ASD is more common in men than in girls. According to a gender distribution graph, one in every two to three-year-old kid has autism, with a higher prevalence of ASD in men than in women.

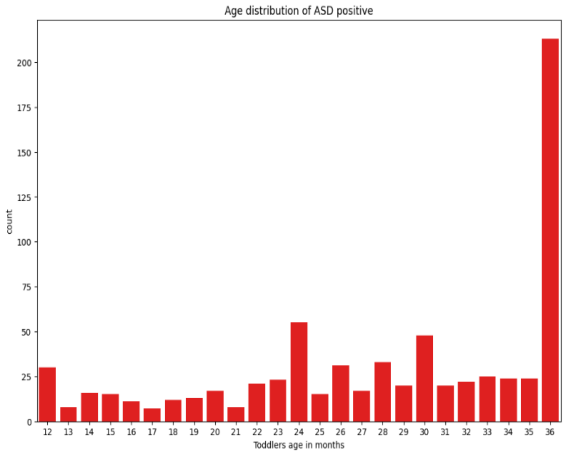

Fig. 6. ASD positive based on their age in months

In Fig 6, it is about how many Childrens have ASD positive based on their age in months, with most cases occurring around 36 months of age. Significant signs of autism occur at 3 years, with one out of every 68 children aged 2-3 years having autism.

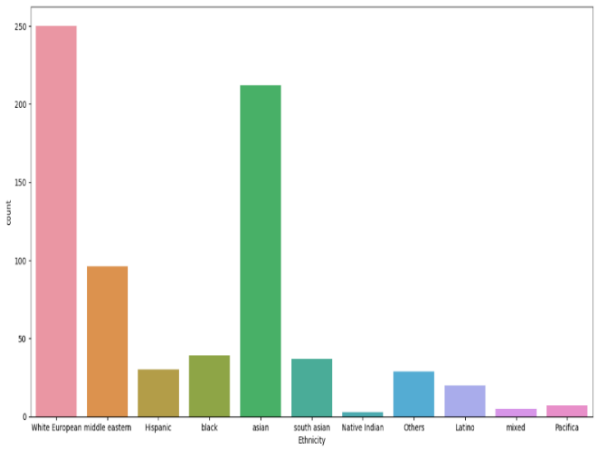

Fig. 7. Distribution of ethnicity among childrens with ASD

In Fig 7, it describes about distribution of ethnicity among Childrens with autism spectrum disorder (ASD). According to graph white Europeans and Asian are with most cases having ASD.

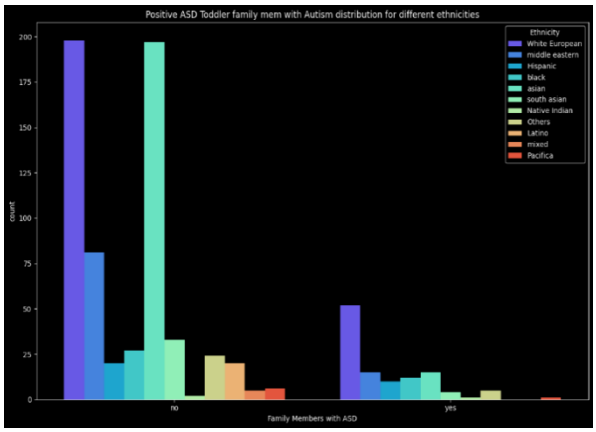

Fig. 8. Distribution of autism in families across various ethnicities

In Fig 8, it count plot and illustrating the distribution of autism in families across various ethnicities within a dataset of Childrens with Autism Spectrum Disorder (ASD).

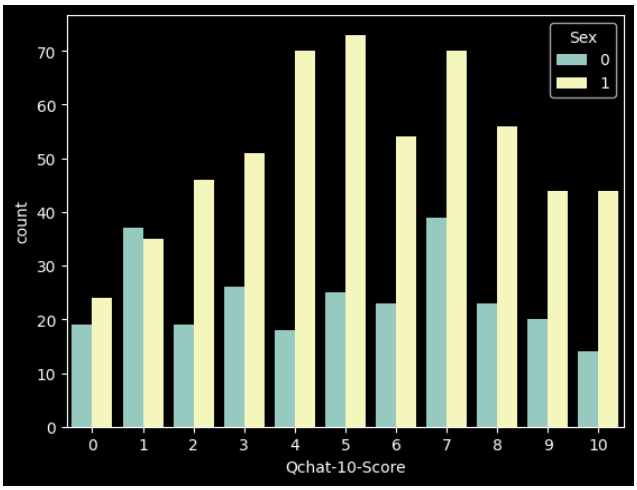

Fig. 9. Qchant-10-score for male and female

In Fig 9 describe about Qchat-10-Score differentiates between meals and females and the score range is from 0 to 10. In the graph 0 means females and 1 means males compared to females males have highest possibility having ASD.

B. Model Building

a. Data Splitting:

We separated our dataset into two subsets: the training set, which contained 60% of the total data, and the testing set, which contained 40% of the total data.

b. Standard Scaling:

Our dataset was split into two sections: a training. We used a technique known as standard scaling to scale the characteristics. The process of placing every feature in a dataset on an equal scale is known as standard scaling. With this modification, all features are guaranteed to have a mean of 0 and a standard deviation of 1. To put it another way, it helps level the playing field for all features, which facilitates accurate comparison and analysis. By ensuring that every characteristic is on the same scale, our analysis will be made easier.

c. Training and Testing

In addition, we used five classification models—Naive Bayes, Support Vector Machine, K-Nearest Neighbours, Random Forest Classifier, and Logistic Regression—to train our data. For every model, we determined its testing and training accuracies. Additionally, we produced a classification report for each specific model that included the accuracy, recall, and f1 score for the training and testing sets.

d. Choosing the best classifier

In our autism disorder prediction research, the logistic regression model and random forest stood out with an impressive testing accuracy of 100%.

V. MACHINE LEARNING ALGORITHMS

In our study, we used machine learning algorithms to make predictions. The algorithms are:

Logistic Regression (LR) : A logistic function is used in logistic regression, a statistical technique, to determine which model best fits the connection between a binomial character and independent variables.

Naive Bayes (NB): Assuming conditional independence of all input characteristics, the Naive Bayes (NB) model of conditional probability is based on counting and the Bayes theorem. As a result, training data is used less often since the convergence rate is faster than with discriminative models like logistic regression. On the other hand, NB performs best when there are few characteristics and a large bias.

Support Vector Machine (SVM): A classification technique called Support Vector Machine (SVM) looks for the optimal hyperplane, or margin, to split a data set into two groups. It finds ways to maximize the margin of the training data. the ideal hyperplane for separation. Better results were obtained while using a linear RBF kernel for the training procedure.

K-Nearest Neighbors (KNN): KNN is a machine learning technique that is easy to use and effective at grouping objects or predicting values. The basic concept of KNN is as follows: it uses the average of its closest neighbours in the feature space or the majority vote to classify or forecast a data item. The number of neighbours taken into account while making decisions is determined by the parameter 'k'.

Random Forest Classifier (RFC): A versatile approach that may be applied to regression, classification, and other

applications is the random forest classifier. Using random data points, it generates many decision trees in order to function. Voting is used to determine which of each tree's predictions is the best.

VI. Results

A. Precision and recall

By counting the number of correctly anticipated positive points, precision calculates the accuracy of positive predictions.

$$\text{Precision} = 2 + \frac{Precision*Recall}{Precision+Recall}$$

$$\text{precision} = \frac{TP}{TP+FP}$$

Recall compares to sensitivity, which gauges how well our algorithm detects positive points, to determine what proportion of positives were accurately predicted.

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

The likelihood of the classifier's accurate predictions, or the percentage of right predictions among all the predictions, is known as accuracy.

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

Five machine learning models were employed, namely: Random Forest Classifier (RFC), K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), and Logistic Regression (LR). Confusion matrix and F1 score were used to assess performance; a comparison is show.

The model's accuracy, precision and recall values were assessed, and the F1 score was calculated by calculating the harmonic mean of these values. A score of 1 indicates the best model, with a higher score indicating better performance.

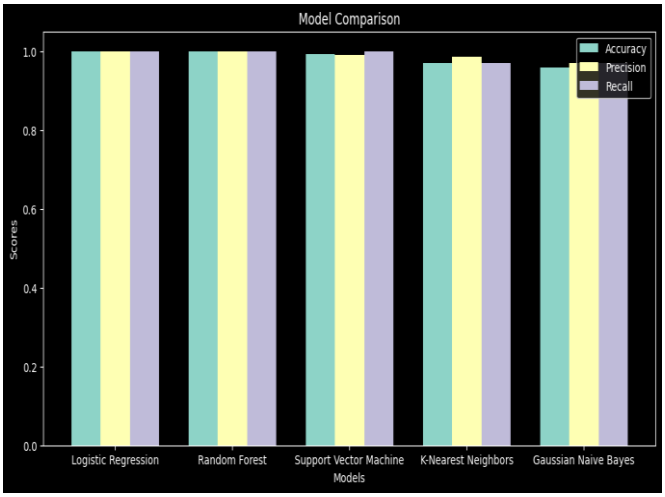

Fig. 10. Comparison of metrics

In fig 10 that represents the comparison of accuracy precision and recall score of each model. It clearly depicts that LR and RF has highest accuracy of and precision followed by Logistic Regression, Random forest , KNN and then decision tree.

B. Comparison of Accuracies

We've compared the testing accuracies of every model of our proposed system with the accuracies recorded in existing system in the Table 1.

Table 1. Accuracy Comparison

Model	Existing Model Accuracy [19]	Proposed System Accuracy
Logistic Regression	97.15	100
Random Forest	81.51	100
SVM	93.84	97.2
KNN	90.52	96.86
Naive Bayes	94.79	97.34

From table 1 we can see the spike of testing accuracies from existing model to our proposed system. Logistic Regression and Random Forest has the highest testing accuracy of 100% and KNN stays at the bottom with 94.3. The improvement in accuracy is the result of data balancing techniques as well as eliminating the collinear features in our proposed system.

C. K-fold Cross Validation:

We also applied K-fold cross validation in our study on every model with 5 folds and compared our testing accuracy with K fold accuracy in fig.11, We can see that the accuracies we get from k-fold cross-validation are bit lower than those from. This is because k-fold cross-validation involves training and testing the model multiple times on different parts of the data, providing a more thorough evaluation.

Fig. 11. Comparison of k-fold and testing accuracy.

D.ROC and AUC Curves

Table 2. Area under Curve

Algorithms	AUC (%)
Logistic Regression	1.00
Random Forests	1.00
SVM	1.00
KNN	0.99
Naive Bayes	0.98

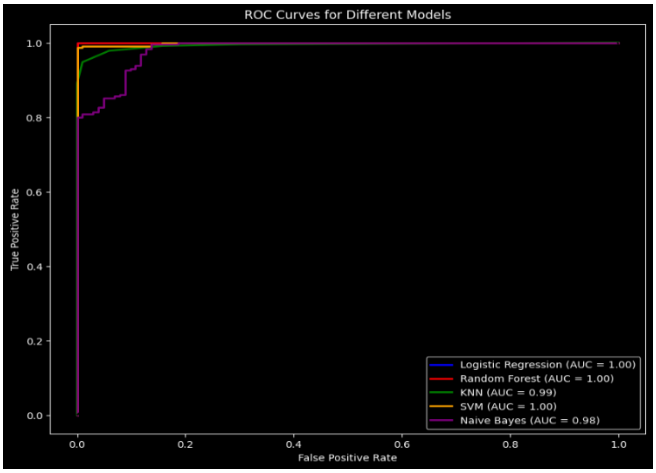

Fig. 12. ROC and AUC curves

Fig. 12 the ROC curves of each machine learning algorithm are illustrated, providing insights into their classification performance. The Area Under the Curve (AUC) serves as a pivotal metric, where higher values denote superior classifier performance. Logistic Regression and Random Forest achieves the highest AUC score of 1.00%, indicating its robust discriminative ability. Conversely, the Naïve Bayes demonstrates the lowest AUC score of 0.98%.

E. Confusion Matrix:

We created confusion matrices for each model using the testing data. These matrices help us see how well our models classify data by showing the number of correct and incorrect predictions. By analyzing these matrices, we can understand how accurate our models are in making predictions and identify areas for improvement. This process allows us to select the most effective model for our analysis and improve the overall reliability of our predictions. Here are the confusion matrices generated for every model. confusion matrices generated for every model.

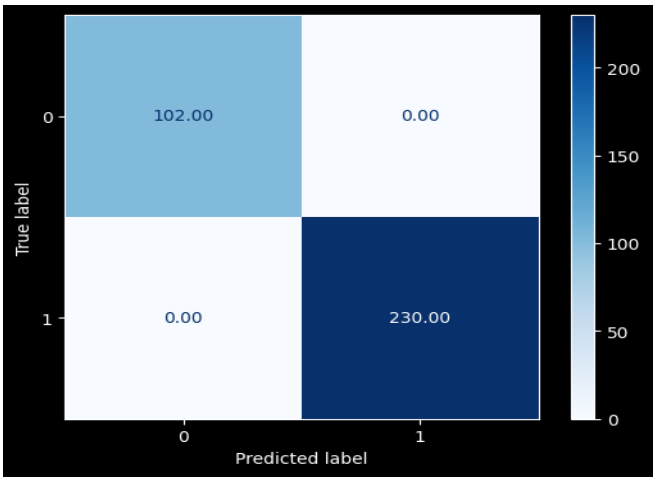

Fig. 13. Logistic Regression Confusion Matrix

Fig 13 shows that the Logistic Regression model accurately identified 10z positive cases and 230 negative cases. However, it correctly classified 0 negative cases as positive and missed 0 positive cases

.

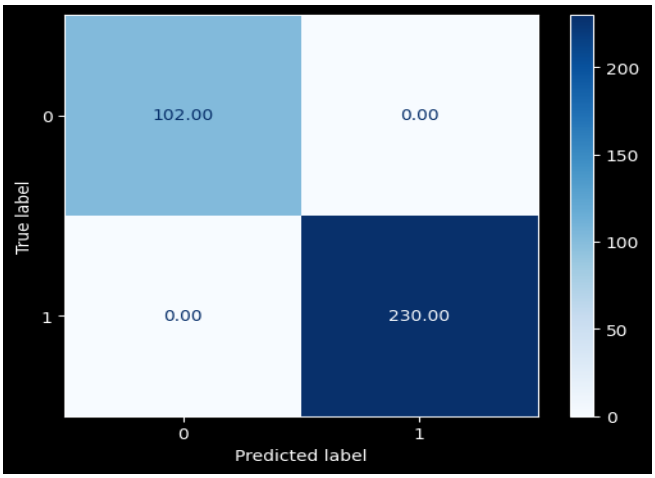

Fig. 14. Random Forest Confusion Matrix

Fig.14 shows that the Random Forest model accurately identified 102 positive cases and 230 negative cases. However, it correctly classified 0 negative cases as positive and missed 4 positive cases.

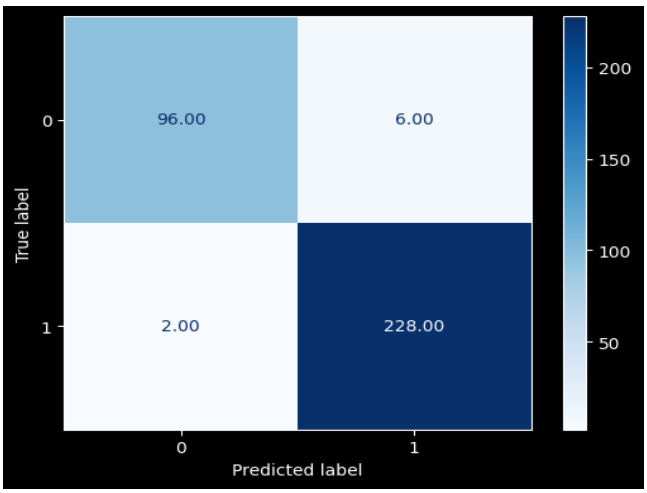

Fig. 15. SVM Confusion Matrix

Fig.15 shows that the SVM model accurately identified 96 positive cases and 228 negative cases. However, it incorrectly classified 6 negative cases as positive and missed 2 positive cases.

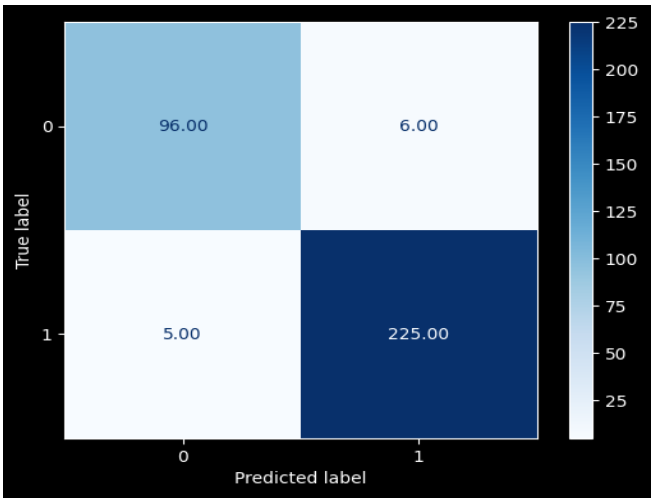

Fig. 16. KNN Confusion Matrix

Fig 16 shows that the KNN model accurately identified 96 positive cases and 225 negative cases. However, it incorrectly classified 6 negative cases as positive and missed 5 positive cases.

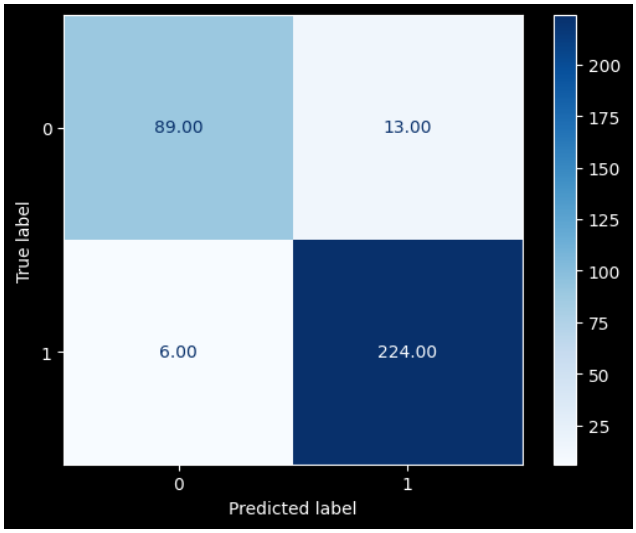

Fig. 17. NB Confusion Matrix

Fig 17 shows that the NB model accurately identified 89 positive cases and 224 negative cases. However, it incorrectly classified 13 negative cases as positive and missed 6 positive cases.

V. CONCLUSION

The assessment of Autism Spectrum Disorder (ASD) behavioural traits is a time-consuming process with overlapping symptoms. There is no diagnostic test or screening tool for detecting ASD. An automated ASD prediction model was designed, with Logistic Regression showing the highest accuracy out of five models applied to the dataset. This research focuses on developing an automated model for detecting autism in children using large and open-source ASD datasets. The current dataset lacks sufficient instances, but the findings provide valuable insights for future development. The researchers plan to use a larger dataset and deep learning techniques to improve system robustness and performance. The study analyses various classification models for accurately detecting ASD in children based on behavioural and medical information, which can be used by other researchers for further exploration.

VI. REFERENCES:

[1] Landa, R.J Gross, A.L Stuart, E.A. and Faherty, A, "Developmental trajectories in children with and without autism spectrum disorders: The first 3years", Child Dev 2019, 84, 429–442.

[2] Belmonte, M.K Allen, G Beckel-Mitchener, AMBoulanger, L.M Caper, R.A Webb, and S.J, "Autism and abnormal development of brain connectivity," J. Neurosis, 2018, 24, 9228–9231.

[3] Sahin, M. A., and Sahin, S. "Investigation of autism spectrum disorder with machine learning techniques: A review study," Computers in Biology and Medicine,2021, 10.1016/j.compbiomed.2021.10449.

[4] Usta, M.B Karabekiroglu, K Sahin, B Aydin, M Bozkurt, A Kara Osman, T Aral, A Cobanoglu, C Kurt, A.D Kesim, N et al, "Use of machine learning methods in prediction of short-term outcome in autism spectrum disorders," *Psychiatry Clin ;*Psychopharmacology, vol *29*, pp. 320–325, 2019.

[5] Hyde, K.K Novack, M.N LaHaye, N Parlett-Pellerito, C Anden, R Dixon, D.R. and Linstead, E, "Applications of supervised machine learning in autism spectrum disorder ," A review. *Rev. J. Autism Dev.* Disorder, vol *6*, pp. 128–146 2019.

[6] Al Banna MH, Ghosh T, Taher KA, Kaiser MS and Mahmud M, "A monitoring system for patients of autism spectrum disorder using artificial intelligence," International conference on brain Infor mastics, pp. 251–62, 2020

[7] Bala, Mousumi, Mohammad Hanif Ali, Md. Shahriar Satu, Khondokar Fida Hasan, and Mohammad Ali Moni, "Efficient Machine Learning Models for Early Stage Detection of Autism Spectrum Disorder, "Algorithms15,pp.5:166,2020, https://doi.org/10.3390/a15050166Abidin.

[8] Uddin MJ, Alyami SA, Ali S, Azad A, and Moni MA, Akter T, Ali MH, Khan MI, Satu MS, Uddin MJ, Alyami SA, Ali S, Azad A, and Moni MA, "Improved Transfer-Learning-Based Facial Recognition Framework to Detect Autistic Children at an Early Stage," Brain Sciences,2019 https://doi.org/10.3390/brainsci11060734.

[9] Akter, T.; Ali, M.H.; Khan, M.I.; Satu, M.S.; and Moni, M.A, "Machine learning model to predict autism investigating eye-tracking dataset. In Proceedings of the 2021 2nd International Conference on Robotics," Electrical and Signal Processing Techniques (ICREST), pp. 383–387,2021.

[10] Hossain, M.D.; Kabir, M.A.; Anwar, A.; and Islam, M.Z, "Detecting Autism Spectrum Disorder using Machine Learning," arXiv:2009.14499,2020.
[11] Raj S and Masood S, "Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques," Procedia Compute, vol 167, pp.994–1004,2020.
[12] Thabtah, F and Peebles, D, " A new machine learning model based on induction of rules for autism detection," *Health Inform. J*, vol *26*, pp. 264–286,2020.
[13] Cavus, Nadire, Abdulmalik A. Lawan, Zurki Ibrahim, Abdullahi Dahiru, Sadiya Tahir, Usama Ishaq Abdulrazak, and Adamu Hussaini ,"A Systematic Literature Review on the Application of Machine-Learning Models in Behavioral Assessment of Autism Spectrum Disorder," Journal of Personalized Medicine, pp.4:299, 11,2021, https://doi.org/10.3390/jpm11040299.
[14] Wang, W., et al. "An overview of machine learning in autism spectrum disorder: A systematic literature review",IEEE Access 2020, 10.1109/ACCESS. 2020 .3013657
[15] Chen, J., et al. "A Review of Machine Learning Methods for Autism Spectrum Disorder Prediction," Journal of Healthcare Engineering , 2021 , 10.1155/2021/9989573
[16] Li, L., et al. "A review of machine learning methods for diagnosing autism spectrum disorder," Frontiers in Psychology, 2021, 10.3389/fpsyg.2021.631284
[17] Zhang, C., et al. "A review on the application of machine learning methods for diagnosing autism spectrum disorder," Computer Methods and Programs in Biomedicine, 2021, 10.2174/2666826622999210601142558.
[18] Kumar, A., and Kumar, V. "Machine Learning Based Autism Spectrum Disorder Prediction Models: A Review," International Journal of Research in Engineering, Science and Management, 2020, 10.47577/IJRESM.2020.7224.
[19] Varadkar, K., Purkayastha, D. and Krishnan, D," Detection of Autism Spectrum Disorder in Children Using Machine Learning Techniques," *SN COM PUT.SCI*,2021,https://doi.org/10.1007/s42979-021-00776-5.
[20] Dataset:https://www.kaggle.com/fabdelja/autism screening-for toddlers. Accessed 1 Oct 2019.