Using Machine Learning Techniques for Autism Spectrum Disorder Analysis and Detection in Children

Norah Alshahrani, Abdulaziz Almaleh

Abstract—Autism Spectrum Disorder (ASD) is a condition related to issues with brain development that affects how a person recognises and communicates with others which results in difficulties with interaction and communication socially and it is constantly growing. Early recognition of ASD allows children to lead safe and healthy lives and helps doctors with accurate diagnoses and management of conditions. Therefore, it is crucial to develop a method that will achieve good results and with high accuracy for the measurement of ASD in children. In this paper, ASD datasets of toddlers and children have been analyzed. We employed the following machine learning techniques to attempt to explore ASD: Random Forest (RF), Decision Tree (DT), Naïve Bayes (NB) and Support Vector Machine (SVM). Then feature selection was used to provide fewer attributes from ASD datasets while preserving model performance. As a result, we found that the best result has been provided by SVM, achieving 0.98% in the toddler dataset and 0.99% in the children dataset.

Keywords—Autism Spectrum Disorder, ASD, Machine Learning, ML, Feature Selection, Support Vector Machine, SVM.

I. Introduction

AUTISM Spectrum Disorder (ASD) is a disease that affects children from a young age and it leads to a disorder in their mental development which affects their level of intelligence and prevents them from communicating with the outside world [1]. Some ASD symptoms, such as lack of eye contact or a lack of responsiveness to their name, are present in children in early childhood [1]. Children grow normally during the first few months or years of life, but they suddenly become introverted, aggressive, or lose the language skills they have already acquired, and ASD symptoms usually appear at the age of one-year [2]. However, previous research which was carried out by Baxter et al. [3] where they have stated that males were three times more likely than females to have ASD with a frequency of 7.6 per 1000, or one person in 132 people over the world. While there is no single known cause of Autism Spectrum Disorder, there might be many causes for ASD such as genetic factors which may be linked to a genetic disorder like Fragile X syndrome and Rett syndrome. In addition, environmental factors can be considered as one of the possible causes like complications during pregnancy, air pollutants and medications [4].

The world has noticed a significant increase in the number of children with ASD of different races and it is rising a concern [5]. It is not clear whether this is due to genetic or environmental factors. However, ASD affects children of different races, but some factors increase a child's risk of autism and they may include gender, jaundice, and family history. Based on that, in this study, we will analyze the selected data to investigate whether there is a relationship between gender, jaundice, and family history and the presence of a child with ASD. In addition, the result of this study will enable doctors and specialists with accurate results regarding the root causes of such disease.

This study is aimed to analyze and detect the data of children with ASD and predict the upcoming cases with high accuracy. The aim of the study can be divided into several objectives:

1) Building a model that helps doctors to diagnose whether children have autism or not.
2) Working on early detection of children with autism to reduce the spread of autism and provide treatment.

II. Related Work

SVM, RF and kNN are the three models Erkan and Thanh have used in their study [6]. On data sets, all of the three models were trained and tested with various alpha values, regardless of whether the dataset was complete (i.e., no missing values) or incomplete (i.e., missing data). In this literature, it has been reported that the RF method identified data with 100% accuracy for all datasets. However, if the model achieved 100% accuracy means that the model is overfitted.

Akter et al. [7] have relied on several steps for early detection of ASD. Firstly, they have worked on the transformations feature and used three methods which are Log, Z-score and Sine. Then, four methods of feature selection techniques have been used. In addition, they have selected six evaluations for metrics: AUROC, Log-loss, Accuracy, Sensitivity, Kappa statistics and Specificity. The findings were given in tables, and each model was evaluated using transformations and a random sample distribution with summary statistics (Maximum, Mean, and Median). As a result, SVM performed best for the toddler dataset, while Adaboost performed best for the children and adult datasets. Furthermore, Glmboost had the best adolescent results. The feature transformations with the best classification results were Sine function for toddlers and Z-score for children and adolescents. However, it was unclear which of the results

N Alshahrani is with Department of Information System and Data Science, King Khalid University, Abha, Saudi Arabia (e-mail: 442813667@kku.edu.sa).

A Almaleh is with College of computer science, King Khalid University, Abha, Saudi Arabia (e-mail: ajoyrulah@kku.edu.sa).

were taken into account and no one strategy has shown to be successful and simple for all datasets.

Raj and Masood [8] have presented methods for detecting autism by using behavioural disease. Three datasets of children, adolescents, and adults with autism were used in this work and multiple models, such as LR and SVM have been applied. Although the model with higher accuracy in all datasets is Convolutional Neural Network (CNN), there is a shortage in explaining how to work on CNN with non-image data. Moreover, kNN models have assigned k to 5 and there are not enough details added on how the value was chosen. kNN is determined by the value of k which is the number of nearest neighbours. It is the calculation of the euclidean distance between the test points and the training points. Then, they determine the best k-value with the lowest mean error value.

Chowdhury and Iraj [9] have worked to predict autism by using ML algorithm. Data were collected from a survey conducted by the Center for Autism Research at the University of Cambridge, UK [10]. In addition, the missing data were handled by the deletion method. In fact, choosing a feature selection is very important as a Chi-square or Fisher's Score. However, in this work it has not been mentioned that they will use any techniques of Feature Selections. Moreover, the accuracy of the SVM was 0.95% which is the best result.

Hossain et al. [11] have used five different ways for feature selection important to facilitate the detection of autism. They have explained that Relief F feature selection technique was the best way. Eight models have been applied such as Logistic Regression (LR), Iterative Classifier Optimizer (ICO) and others based on Relief F. They found that all the models that were applied achieved high accuracy. Also, the results showed that MultiLayer Perceptron (MLP) achieved 100% accuracy with minimum attributes for all datasets. However, if a model accuracy achieves 100% that means the model is overfitted.

Much of the work discussed above uses machine learning and deep learning techniques, and the majority of the findings were labeled as overfitting. As a result, there is a clear need to investigate the feasibility of using machine learning-based models to diagnose ASD in children via a speedy and effective technique for early detection and diagnosis.

III. Datasets

In this study, two ASD datasets have been selected for the purpose of this study, both acquired through the open-source bank Kaggle [12]. The first data set of toddlers between 12 and 36 months of age consisted of 19 columns and instances of 1188 [13]. The second data set for children aged 4 to 11 years consisted of 20 columns and instances of 801 [14]. The two datasets contain ten attributes (A1 to A10) representing answers to screening questions as well as categorical variables such as gender, ethnicity, jaundice, family ASD, who completed the test, and ASD class. Also, they contain numerical variables such as age and Q-chat-10-Score. The attributes description are shown in Tables I and II.

TABLE I
Details of Variable Mapping

Variable	Q-CHAT-10 Toddler features (12-36) months	Q-CHAT-10 Child features (4–11) years
Q1	Does your child look at you when you call his/her name?	S/he often notices small sounds when others do not
Q2	How easy is it for you to get eye contact with your child?	S/he usually concentrates more on the whole picture rather than the small details
Q3	Does your child point to indicate that s/he wants something?	In a social group, s/he can easily keep track of several different people's conversation
Q4	Does your child point to sharing interest with you?	S/he finds it easy to go back and forth between different activities
Q5	Does your child pretend?	S/he does not know how to keep a conversation going with his/her peers
Q6	Does your child follow where you're looking?	S/he is good at social chit-chat
Q7	If you or someone else in the family is visibly upset, does your child show signs of wanting to comfort them?	When s/he is read a story, s/he finds it difficult to work out the character's intentions or feelings
Q8	Would you describe your child's first words as:	When s/he was in preschool, s/he used to enjoy playing pretending games with other children
Q9	Does your child use simple gestures?	S/he finds it easy to work out what someone is thinking or feeling just by looking at their face
Q10	Does your child stare at nothing with no apparent purpose?	S/he finds it hard to make new friends

IV. Proposed Solution

Our methodology technique for detecting ASD is shown in Fig. 1 and is detailed further below.

A. Data Preprocessing

Various data preprocessing methods are used, such as dealing with missing values, normalization, etc. The children's dataset contains records with missing values. There are numerous approaches to deal with missing values, including replacing missing values with interim values or removing instances that contain missing values. To prevent providing biased estimations that might lead to incorrect results, each record having a missing value in the chosen dataset was deleted.

B. Exploratory Data Analysis

For two ASD datasets, the correlation of gender, jaundice, family ASD, and ethnicity in ASD cases were investigated by using the Pearson correlation coefficient (PCC).

- Correlation between gender and ASD cases:
  Researchers have found a correlation between autism and gender. They stated that males are more probable

TABLE II
FEATURES DESCRIPTION

Feature	Type	Description
A1: Answer for question 1	Binary	The answer code of the question based on the screening method used
A2: Answer for question 2	Binary	The answer code of the question based on the screening method used
A3: Answer for question 3	Binary	The answer code of the question based on the screening method used
A4: Answer for question 4	Binary	The answer code of the question based on the screening method used
A5: Answer for question 5	Binary	The answer code of the question based on the screening method used
A6: Answer for question 6	Binary	The answer code of the question based on the screening method used
A7: Answer for question 7	Binary	The answer code of the question based on the screening method used
A8: Answer for question 8	Binary	The answer code of the question based on the screening method used
A9: Answer for question 9	Binary	The answer code of the question based on the screening method used
A10: Answer for question 10	Binary	The answer code of the question based on the screening method used
Age	Number	Toddlers (months) , Children (year)
Q-chat10-Score	Number	Toddler (Less than or equal 3 no ASD traits ; greater than 3 ASD traits) , children (Less than or equal 6 no ASD traits ; greater than 6 ASD traits)
Gender	Character	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean	Whether the case was born with jaundice
Family member with ASD history	Boolean	Whether any immediate family member has a PDD
Who is completing the test	String	Parent, self, caregiver, medical staff, clinician, etc.
Used the screening app before	Boolean	Whether the user has used a screening app
Class variable	Boolean	Toddler and Children diagnosed with ASD

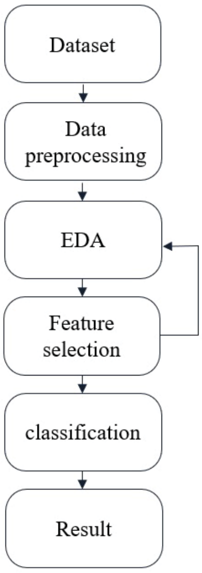

Fig. 1 Methodology for detection of ASD

TABLE III
MEASURING ASSOCIATION USING PEARSON CORRELATION COEFFICIENT

Variables	Dataset	PCC	P-value	Interpretation of association
Gender	Toddler	0.108	0.0001	Significant
Gender	Child	0.025	0.491	No significant
Jaundice	Toddler	0.0795	0.006	Significant
Jaundice	Child	-0.0006	0.985	No-significant
Family ASD	Toddler	-0.0198	0.4991	No-significant
Family ASD	Child	-0.0229	0.529	No-significant
Ethnicity	Toddler	-0.146	4.302e-07	Significant
Ethnicity	Child	-0.138	0.00017	Significant

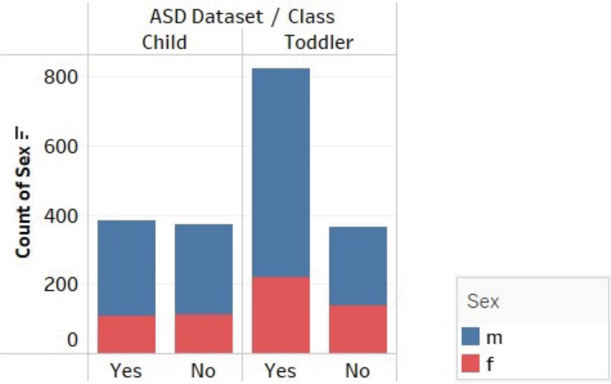

Fig. 2 Distribution of the ASD dataset by gender

to contract the ASD than a female [3]. The distribution of the ASD dataset by gender is shown in Fig. 2 and it has been noticed that ASD is more frequent in males than females. In addition, the p-value in Table III shows a correlation significant in the toddler dataset and no significant correlation in the children dataset.

- Correlation between the family ASD and ASD cases:

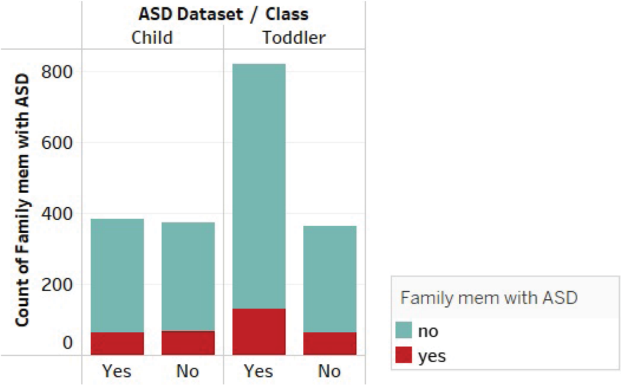

Fig. 3 Correlation between the family ASD and ASD cases

In Fig. 3, the correlation can be described as relatively weak between ASD cases and family ASD for two data sets. Based on the p-value, the correlation study findings in Table III reveal that there is no significant correlation between family ASD and ASD cases for toddlers and children.

- Correlation between jaundice and ASD cases:

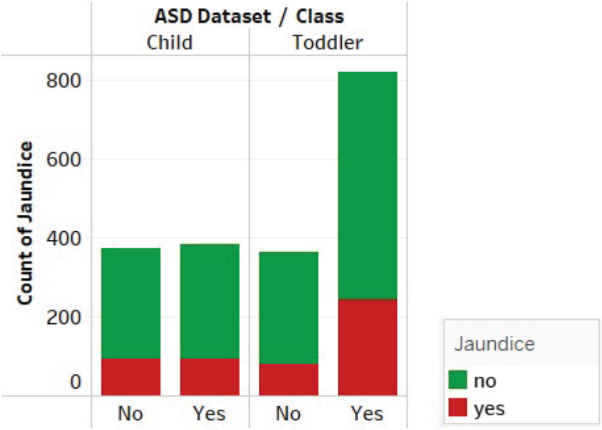

Fig. 4 Correlation between jaundice and ASD cases

Fig. 4 describes the correlation between jaundice and ASD cases for two datasets. For the children dataset, our correlation analysis results in Table III reveal that there is no significant connection between jaundice and ASD cases. However, the results for the toddler dataset have shown a correlation in p-value.

- Correlation between the ethnicity and ASD cases:
  Fig. 5 describes the correlation between ethnicity and ASD cases for two datasets. According to the p-value, there is a significant correlation between ethnicity and ASD cases in both datasets, as shown in Table III. We noted white Europeans, Asians, and then the Middle East had the most increase in autism cases in both datasets.

C. Feature Selection

The purpose of using feature selection is to improve classification accuracy by selecting the most significant set of features and identifying the models that outperform in both selected ASD datasets. To ensure the quality of the characteristics employed, Pearson's correlation coefficient (PCC) was used. The correlation and p-values for each column of the target column were determined. The result is summarized in Table IV. As a result, if this feature (Q-chat-10-Score) is used in classification, it means that the classification algorithm already knows the target's outcome. In order to ensure that there is no bias in the data needed for the models, this attribute has been removed from the two datasets during the analysis.

TABLE IV
FEATURE SELECTION

Feature	Description
A1 – A10	The most important features to detect cases of autism.
Case No, that completed the test and used app before	The p-value is higher than 0.05 in both dataset so it's no significant.
Q-chat-10-Score	Using this attribute means that the model already has the outcome of the target.

D. Model

Both datasets were divided into two sections. The first part involves training datasets at a rate of 80%, followed by testing datasets at a rate of 20%. We implemented machine learning classification models to obtain the findings after separating the datasets. There are many different classification models that can be used. However, in this study, we relied on four classifications:

- RF algorithm is a supervised machine learning technique built on decision tree algorithms. It is a classifier made up of a group of tree-structured classifiers that vote for the most popular class using identically distributed independent random vectors [15].
- DT is a supervised machine learning technique in which data are constantly partitioned according to a parameter. Decision nodes and leaves are two entities that may be used to explain the tree. The leaves indicate the final results, and the data are split at the decision nodes [15]. The Information Gain approach was utilized in this study to divide a DT.
- The supervised machine learning algorithm Naive Bayes (NB) is one of the most efficient and successful ways of categorization accessible. It is a generative model with a joint probability distribution [16].
- SVM are supervised machine learning algorithms that may be used to address issues like classification and regression. However, it is the most widely used in classification problems [17].

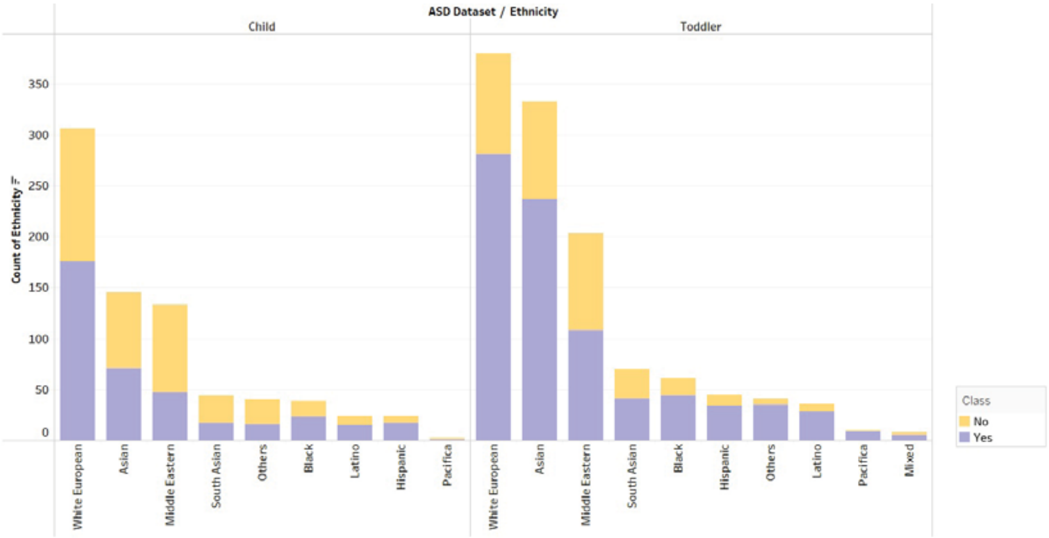

Fig. 5 Correlation between ethnicity and ASD cases

E. Evaluation

Accuracy, F1 Score, AUROC, log-loss, and Cohen's Kappa are some of the classification metrics that have been used to describe the results of classifiers and evaluate their performance. Their formulation is as follows:

- Accuracy:
$$\frac{TP+TN}{TP+FP+FN+TN}$$
- F1 score:
$$2\times\frac{Precision \times Recall}{Precision + Recall}$$
- AUROC:
True Positive Rate (TPR):
$$\frac{TP}{TP+FN}$$
False Positive Rate (FPR):
$$\frac{FP}{FP+TN}$$
- log-loss:
$$\frac{-\sum_{y=1}^{j}\sum_{x=1}^{n} f(x,y)log(p(x,y))}{n}$$
- Cohen's Kappa :
$$1-\frac{1-p_o}{1-p_e}$$

The following evaluation metrics are computed by using four different groups of predicted and actual values:

- True Positive (TP): the toddler/child is having ASD and is correctly predicted as having ASD.
- True Negative (TN): the toddler/child is not having ASD and is correctly predicted as not having ASD.
- False Positive (FP): the toddler/child is not having ASD and is incorrectly predicted as having ASD.
- False Negative (FN): the toddler/child is having ASD and is incorrectly predicted as not having ASD.

V. Result and Discussion

In this study, we relied on the PCC in feature selection. In addition, four models were applied to two datasets which are RF, DT, NB, and SVM. The results were presented and the comparison between the classifiers through different evaluation metrics such as Accuracy, F1 Score, AUROC, log-loss, and Cohen's Kappa has been considered.

The findings for the two datasets are shown in Table V. The best model is SVM classifiers, which have 0.98% accuracy for the toddler dataset and 0.99% for the children dataset. Furthermore, in the toddler dataset, RF had an accuracy of 0.97%, and in the children dataset, it had an accuracy of 0.98%. The next model is DT, which scored about 0.93% for toddler datasets and 0.98% for children datasets. While the NB was the least accurate model, it achieved an accuracy 0.91% in the toddler dataset and 0.90% in the children dataset.

TABLE V
Accuracy of Different Classifiers

Dataset	RF	DT	NB	SVM
Toddler	0.979	0.933	0.916	0.983
Children	0.987	0.987	0.908	0.993

In the AUROC metric, a perfect classifier is the one whose score is 1. Table VI shows the result for two datasets. Accordingly, the best classifiers are SVM and RF where the result for SVM of the dataset of toddler was 0.97% and children 0.99%. In addition, the RF model achieved 0.97% for the toddler and 0.98% for the children. The performance of the DT model was good, as it achieved 0.94% for the toddler and 0.98% for the children. The last classifier is NB, the score for toddler is 0.89% and for children 0.90% and it is considered as the worst result compared to the other classifiers even though its performance is good. The AUC Receiver Operating

Open Science Index, Health and Medical Engineering Vol:17, No:2, 2023 publications.waset.org/10012974.pdf

Characteristics (ROC) curves for the two datasets are shown in Figs. 6 and 7.

TABLE VI
AUROC of Different Classifiers

Dataset	RF	DT	NB	SVM
Toddler	0.970	0.940	0.897	0.973
Children	0.987	0.987	0.908	0.993

The F1 Score is another metric that has been used to assess the quality of models. The F1 score represents the balance between precision and recall. Table VII shows the result F1 score where higher F1 scores are better. SVM and RF show the best results where SVM achieved 0.98% for toddlers and 0.99% for children and RF for both datasets achieved approximately 0.98%. In DT, the model achieved good results for toddlers where the score is 0.95% and for children 0.98%. In addition, NB scored 0.94% for toddlers and 0.91% for children.

TABLE VII
F1 Score of Different Classifiers

Dataset	RF	DT	NB	SVM
Toddler	0.985	0.950	0.940	0.988
Children	0.987	0.987	0.910	0.994

The log-loss indicates how close the expected values and actual values are. The larger the log-loss, the greater the difference between the expected and actual likelihood. A classifier with a lower log-loss score is better. Therefore, a log-loss was used to compare the classifiers, and we found that SVM achieved the best value, as the result for toddlers was 0.58 and for children 0.22. Also, RF achieved a good result with the toddler being 0.72 and the children 0.45. Moreover, the DT model achieved a good result in children 0.45 and the result was higher in toddlers achieving 2.32. However, for the NB model, it achieved a worse result in both groups as it is shown in Table VIII.

TABLE VIII
Log-Loss of Different Classifiers

Dataset	RF	DT	NB	SVM
Toddler	0.726	2.322	2.902	0.580
Children	0.454	0.454	3.181	0.227

Cohen's Kappa metrics has been used as a classification evaluation. Cohen's kappa can be used to compare any model used for the same classification task. Furthermore, a Cohen's kappa higher value means perfect agreement and it is better. Cohen's Kappa results are shown in Table IX. The results of this test showed that the SVM model was the best classifier as it achieved 0.96% for toddler and 0.98% for children. Then, RF achieved 0.95% for toddler and 0.97% for children. Regarding DT, it achieved a good result in children 0.97% and the result was reduced to 0.84% for toddler. Moreover, The NB result was approximately 0.80% for both datasets.

TABLE IX
Cohen Kappa of Different Classifiers

Dataset	RF	DT	NB	SVM
Toddler	0.950	0.848	0.801	0.960
Children	0.974	0.974	0.816	0.987

TABLE X
Comparison of Our Model with Other Research

Dataset	Author	Accuracy	AUROC	Log-loss	kappa
Toddler	Akter et al. [7]	98.77	99.98	3.01	97.10
	This reseach	98.3	97.3	0.58	96.0
Children	Akter et al. [7]	97.20	99.89	9.62	94.41
	This reseach	99.3	99.3	0.227	98.7

Although several researchers have conducted studies on ASD, we still need more improvement. Our results were compared with the results of [7] as shown in Table X. We found that they used different classifiers for the given values, the best for toddler model was SVM, and the best child model was Adaboost. However, in our research SVM achieved all good results for the two datasets.

By using machine learning techniques, healthcare professionals can detect autism spectrum disorder in children faster and with high accuracy. In addition, they will be able to help the patients of ASD by making the necessary decisions to limit the development of this disease. In fact, by using our methodology, specifically the SVM model, they will be able to detect ASD in children with high accuracy.

VI. Conclusion

In this study, we have analyzed the ASD datasets of toddler and children. We applied a variety of machine learning approaches to detect ASD, including RF, DT, NB, and SVM. Feature selection and specifically PCC were applied to present fewer features from the ASD datasets while maintaining model performance. The achieved results will improve the doctors' capacity to detect ASD. We found that SVM outperforms the other classifications, indicated by Accuracy, F1 Score, AUROC, log-loss, and Cohen's Kappa. However, the study's main limitation is that datasets for teenagers and adults were not accessible for inclusion in the study. In future work, we want to acquire datasets for all ages and analyze them using different machine learning techniques and deep learning approaches.

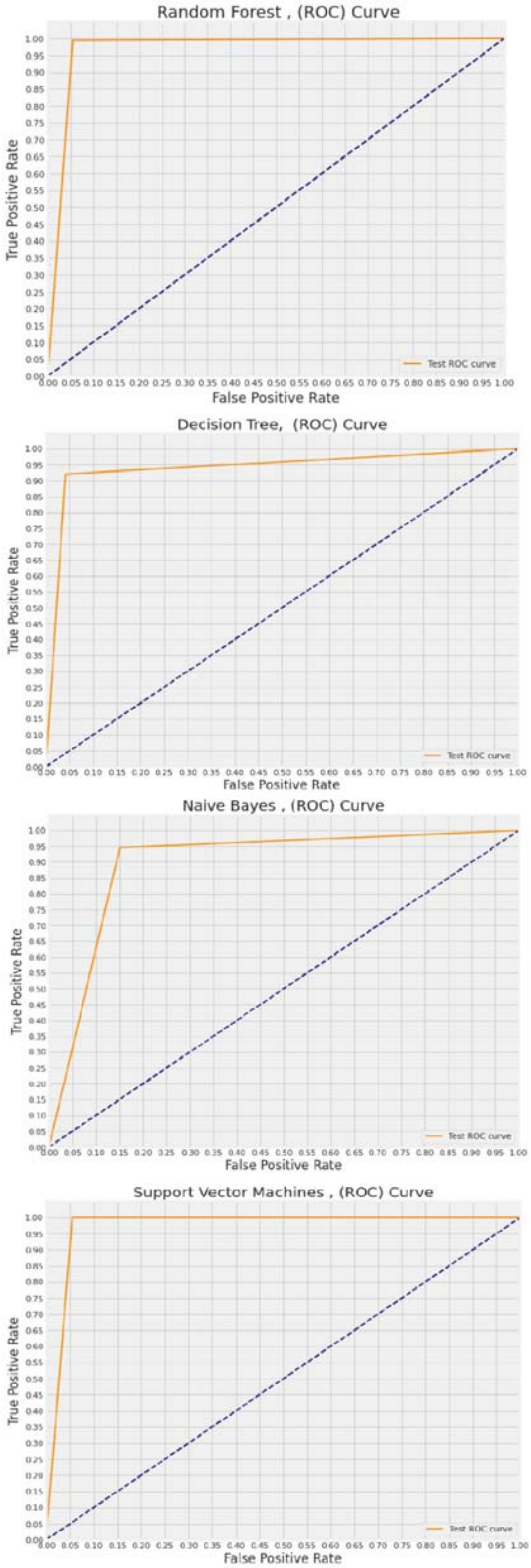

Fig. 6 AUROC Curve for Toddler dataset

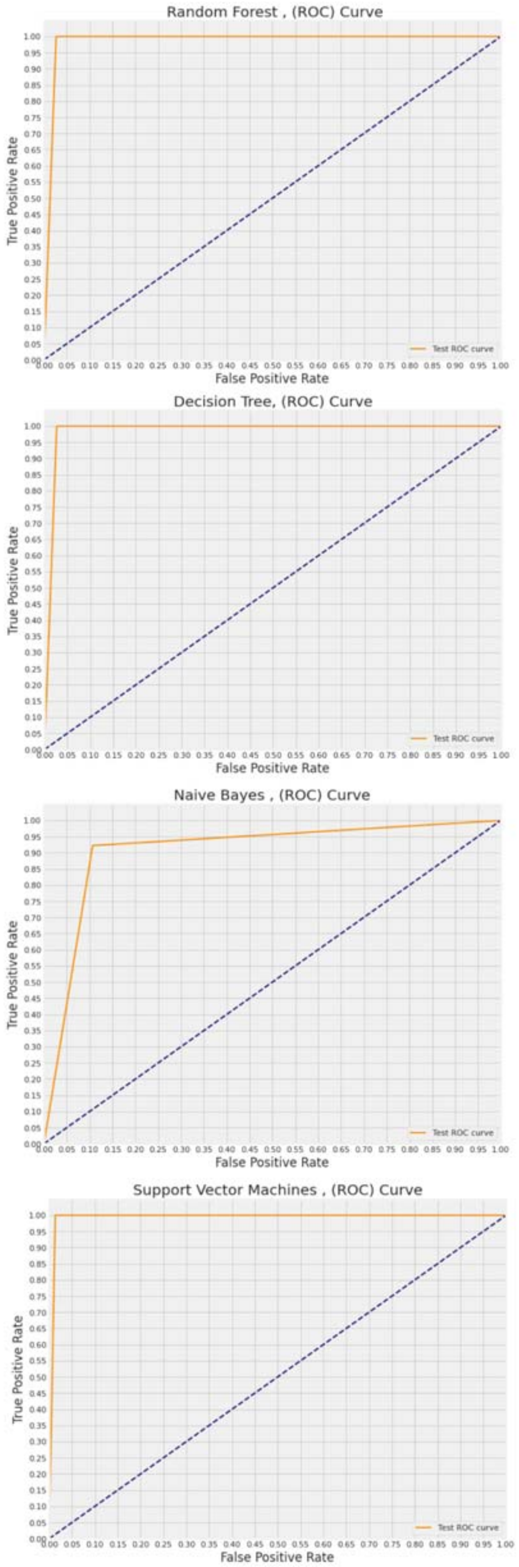

Fig. 7 AUROC Curve for Children dataset

References

[1] Naila Z Khan, Lilia Albores Gallo, Aurora Arghir, Bogdan Budisteanu, Magdalena Budisteanu, Iuliana Dobrescu, Kirsty Donald, Samia El-Tabari, Michelle Hoogenhout, Fidelie Kalambayi, et al. Autism and the grand challenges in global mental health. 2012.

[2] Fabio Apicella, Valeria Costanzo, and Giulia Purpura. Are early visual behavior impairments involved in the onset of autism spectrum disorders? insights for early diagnosis and intervention. *European Journal of Pediatrics*, 179(2):225–234, 2020.

[3] Amanda J Baxter, TS Brugha, Holly E Erskine, Roman W Scheurer, Theo Vos, and James G Scott. The epidemiology and global burden of autism spectrum disorders. *Psychological medicine*, 45(3):601–613, 2015.

[4] L Mercer, S Creighton, JJA Holden, and MES Lewis. Parental perspectives on the causes of an autism spectrum disorder in their children. *Journal of Genetic Counseling*, 15(1):41–50, 2006.

[5] Michael D Kogan, Bonnie B Strickland, Stephen J Blumberg, Gopal K Singh, James M Perrin, and Peter C van Dyck. A national profile of the health care experiences and family impact of autism spectrum disorder among children in the united states, 2005–2006. *Pediatrics*, 122(6):e1149–e1158, 2008.

[6] Uğur Erkan and Dang NH Thanh. Autism spectrum disorder detection with machine learning methods. *Current Psychiatry Research and Reviews Formerly: Current Psychiatry Reviews*, 15(4):297–308, 2019.

[7] Tania Akter, Md Shahriare Satu, Md Imran Khan, Mohammad Hanif Ali, Shahadat Uddin, Pietro Lio, Julian MW Quinn, and Mohammad Ali Moni. Machine learning-based models for early stage detection of autism spectrum disorders. *IEEE Access*, 7:166509–166527, 2019.

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

[9] Koushik Chowdhury and Mir Ahmad Iraj. Predicting autism spectrum disorder using machine learning classifiers. In *2020 international conference on recent trends on electronics, information, communication & technology (RTEICT)*, pages 324–327. IEEE, 2020.

[10] National Institue of Health Research. Autism research centre. *http://docs.autismresearchcentre.com/tests/AQ10.pdf*, Website.

[11] Md Delowar Hossain, Muhammad Ashad Kabir, Adnan Anwar, and Md Zahidul Islam. Detecting autism spectrum disorder using machine learning techniques. *Health Information Science and Systems*, 9(1):1–13, 2021.

[12] kaggle. Datasets. *https://www.kaggle.com/datasets*, Website.

[13] kaggle. behavior analysis of autism. *https://www.kaggle.com/iashiqul/behavior-analysis-of-autism*, Website.

[14] Kaggle. autism screening child. *https://www.kaggle.com/datasets/basmarg/autism-screening-child-two-version?select=Child-Data2017.csv*.

[15] Nasiba Mahdi Abdulkareem, Adnan Mohsin Abdulazeez, et al. Machine learning classification based on radom forest algorithm: A review. *International Journal of Science and Business*, 5(2):128–142, 2021.

[16] Sayali D Jadhav and HP Channe. Comparative study of k-nn, naive bayes and decision tree classification techniques. *International Journal of Science and Research (IJSR)*, 5(1):1842–1845, 2016.

[17] Batta Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9:381–386, 2020.