DOI: http://dx.doi.org/10.26483/ijarcs.v14i2.6969

ISSN No. 0976-5697

Volume 14, No. 2, March-April 2023

International Journal of Advanced Research in Computer Science

SURVEY REPORT

Available Online at www.ijarcs.info

A SURVEY ON PREDICTION OF AUTISM SPECTRUM DISORDER USING DATA SCIENCE TECHNIQUES

R. Ramya
Research Scholar, Department of Computer Science,
Periyar E.V.R College (Autonomous), Tiruchirappalli,
Tamil Nadu, India.

Dr. S. Panneer Arokiaraj
Associate Professor, Department of Computer Science,
Periyar E.V.R College (Autonomous), Tiruchirappalli,
Tamil Nadu, India.

Abstract: Autism Spectrum Disorder is a lifelong brain developmental disorder. Diagnosing the level of Autism and predicting the severity of the same are too complex, and it requires a depth analysis of the historical data on the autism patient. Nowadays, Data science techniques play a vital role in diagnosing autism. Decision Tree, Random Forest, Logistic Regression, Adaboost, Naïve Bayse, K-Nearest Neighbour, Support Vector Machine and etc., are the few techniques labeled under the roof of data science are used to predict such disorders. The paper aims to present a survey on the various models proposed by various researchers to predict the severity of autism using data science techniques.

Keywords: Autism Spectrum Disorder (ASD), Accuracy Parameters, ASD Dataset, Data Science Techniques and Data science Platform.

I. INTRODUCTION

Autism is an abnormality in the brain structure or function. Autism appears before a child ages three because the brain tends to grow too fast during the first three years of life [1]. The following are the Autism patient-related symptoms:

- Complicated with verbal and non-verbal communication.
- Inability to participate in a conversation.
- Difficult with Social interaction.
- Difficult to develop friends with others, and prefer to play alone.
- Difficult to adjust to changes in the common surroundings environment.
- Tedious body movements such as spinning, head banging and flapping.
- Preoccupation with familiar objects.
- Make a little or inconsistent eye contact.

The cause of Autism is uncertain, but possible factors include hereditary data and difference in the development of certain brain functions, leading to impairment in cognitive and social aspects [2].

A child with Autism has three different levels such as low-functioning or Serve Autism, Moderate Autism and High-Functioning or Mild Autism [3]. Figure 1 represents the Autism types and their subtypes.

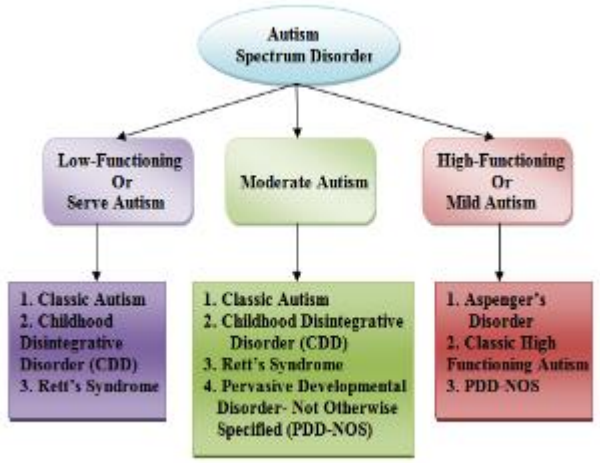

Figure 1: Types of Autism

Diagnosis of Autism is complicated in the early stages because of the more common early symptoms. There is no particular medical test, and diagnosis is made by observing patient symptoms and behavior through a historical dataset [4]. Hence, it is essential to have such data containing the details of Autism patients, including their symptoms, history, lab test data, vaccination, etc. Nowadays, there are too many automated techniques to diagnose Autism Spectrum Disorder like data science, machine learning, Artificial Intelligence, deep learning, etc. So, in this paper, we will briefly introduce Autism diagnoses using various data science techniques such as Decision Tree, Random Forest, Support Vector Machine, Adaboost, Glmboost, Convolutional Neural Network (CNN), Rule-Machine Learning (RML), Logistic regression, K-Nearest Neighbors.

This article is structured as, under the section literature review, discusses the various proposed models based on data science techniques linked with Autism. The

methodology presents an overview of various data science techniques and a comparative study of different existing models based on features. The result and discussion section analyze the efficacy of the various current models used for predicting Autism. Finally, the conclusion section highlights the features and limitations of the enlisted techniques used for this study.

II.LITERATURE REVIEW

Autism is a wrong connection between human brain cells. The efficacy of data science techniques is quite commendable in predicting the different types of autism disorders based on the syndrome. This section briefly presents the works related to the data science techniques used to predict the level of Autism.

C.S.Kanimozhiselvi et al. developed "Grading Autism children using machine learning techniques", They categorized Autism Spectrum Disorder patients into three different levels such as, high, moderate and low, based on Autism patient-related symptoms. The implementation work was carried out in different stages: the first step of implementation is collecting Childhood Autism Rating Scale (CARS) based on case histories for preparing a real-world dataset. The second step is predicting the grade of Autism using some classification models consistent with CARS diagnostic criteria. Finally, evaluating the performance and predicting the different levels of Autism [5].

Kazi Shahrukh Omar et al. proposed a paper titled "A Machine Learning Approach to Predict Autism Spectrum Disorder", proposed an effective prediction model based on data science technique, and developed a mobile application for predicting Autism for the people of different age groups. For this, the researcher has merged the Random Forest-CART and Random Forest-ID3 techniques used to analyze the Autism dataset [6].

Tania Akter et al. proposed a "Machine Learning Based Models for Early Stage Detection of Autism Spectrum Disorders" model. It gathered the early detection of ASD dataset in different stages of life (toddler, child, adolescent and adult) and analyzed the results using a range of different classifiers to explore the significant features of Autism [7].

Suman Raja and Sarfaraz Masood proposed "Analysis and Detection of Autism Spectrum Disorder using Machine Learning Techniques". This research work aimed to build a machine learning model that predicts Autism using supervised machine learning algorithms. The steps involved were pre-processing of data, training, testing with specified models, evaluation of results, and predicting Autism [8].

FadiThabtah et al. developed "A New Machine Learning Model based on induction of rules for Autism detection", in which this article proposed a new machine learning method called Rule-Machine Learning (RML) that only detects autistic traits of different cases and controls but also offers users knowledge bases that can be utilized by domain experts in understanding the reasons behind the classification [9].

Jaber Alwidian et al. "Predicting Autism Spectrum Disorder using Machine Learning Technique". This prediction model utilized the Association Classification (AC) of data mining to predict whether an individual has an Autism disorder. They conducted a Comparative performance analysis on the seven Association Classification (CMAR, CBA, FACA, MCAR, FCBA, ECBA, and WCBA) techniques [10].

Roopa B. S and R. Manjunatha Prasad developed" Identification of Best Fit Learning Models Based on Calibration for Better Classification of Autism". Various supervised learning models were first tested on 1101 subjects with 530 ASD subjects and 571 Normal subjects. The performance was calibrated in terms of the brier score, which is a measure to predict autism in a probabilistic way [11].

Charlotte Küpper et al. proposed "Identifying Predictive Features of Autism Spectrum Disorder in a Clinical Sample of Adolescents and Adults using Machine Learning, "focusing on adolescents and adults as assessed with the ASDS module 4. It used the SVM classifier to examine whether ASD detection can be improved by identifying a subset of behavioral features from the ADOS module 4 in a routine clinical sample of N=673 high-functioning adolescents and adults with ASD(n=885) and non-ASD(n=288) [12].

Lakshmi B and Kala A, developed "Prediction of Autistic Spectrum Disorder Based on Behavioural Features Using Machine Learning," in which Neural Networks were chosen for implementing the Autism predictive system. The Neural Networks of the proposed system is a combination of linear and nonlinear functions that take up the vectors comprising various attributes defined in the Autism dataset [13].

Devika Varshini G and Chinnaiyan R, proposed "Optimized Machine Learning Classification Approaches for Prediction of Autism Spectrum Disorder" in which this paper the effectiveness of various machine learning algorithm and pre-processing techniques for the task of classification for medical dataset that are used for predicting the early Autism traits in toddlers and adults is evaluated [14].

The literature review discusses from various authors the techniques which can be utilized to determine Autism risk factors and in the section of methodology, a detailed study about Autism prediction will be carried out.

III. MATERIALS AND METHODS

Autism is a pervasive developmental disorder. A variety of data science techniques that aid in the classification of different types of autism and their severity levels. The methodology section briefly outlines different classes of data science techniques used in predicting autism.

A. Decision Tree Algorithm

A decision tree is the most powerful decision support tool that uses a tree-like model of decisions. The decision tree model has two steps in the process: the learning step and the prediction step.

- The first step is the learning step; the model is developed based on Autism training data.
- The second step is the prediction step; the model is used to predict the response for Autism data.

The decision tree algorithm has applied the Real-time Childhood Autism Rating Scale (CARS) dataset, and it gives the greatest accuracy of 1.00 for training data and 0.96 for test data. One of the major limitations of this algorithm is that it is not suitable for a large dataset [5], [15].

B. Random Forest

A Random Forest is an ensemble learning method in which many decision trees are constructed and combined to get a more accurate prediction. The steps involved in the Random Forest algorithm are:

- Random Forest n random numbers of records are taken from the autism data set having k number of records.
- Individual decision trees are constructed for every sample, and every decision tree will produce an output.
- Finally, the result is measured based on Averaging for Classification or Majority Voting and Regression.

A prediction model is proposed to improve the performance that merges the Random Forest- CART model with the Random Forest model - ID3. The outcome of these merged models provides an effective and efficient approach to discover Autism, and it reduces the time and cost of Autism prediction. The drawback of this model is that it is only suitable for small datasets [6].

C. Support Vector Machine (SVM)

The Support Vector Machine works by mapping to a high dimensional feature space so that data points can be grouped, even when the data are not otherwise linearly separable. A separator between the categories is found, and then data are transformed in such a way that the separator could be drawn as a hyper plane. This SVM model gives numerous features of data analysis such as better handling of missing values, high accuracy and consumes the minimum time of prediction [7],[12],[13].

D. Adaboost

Adaboost is a popular boosting technique. It is a short form of adaptive boosting, which uses the same training set over and thus need not be large, but the classifiers should be simple so that they do not overfit. It measures the importance of features by calculating the increase in the model's prediction error after permuting the feature [7].

E. Glmboost

The Boosted General Linear Model by allowing the linear model to be related to the response variable through a link function and by allowing the magnitude of the variance of each dimension to be a function of its predicted value. [7].

F. Convolutional Neural Network (CNN)

The Convolutional Neural Network is an interconnected network architecture for deep learning. Convolution is a mathematical operation that allocates the integration of two sets of information. Convolution is applied to the Autism dataset to filter the dataset and then produce a feature map. CNN produces highly accurate results for autism prediction and consumes the minimum time for prediction [8].

G. Rule-Machine Learning (RML)

RML is based on covering classification, employing a search technique for rule discovery. The RML then evaluates the exposed rule and deletes any redundancies. A rule is represented as $(A1,v1)\wedge(A2,v2)\wedge\ldots.\wedge(An,vn)\rightarrow Cn$ where the predecessor is a combination of variable values, and the subsequent is a category value (ASD, No ASD). The merits of the RML algorithm are that it produces accurate results, consumes less time to predict the results, and is better at handling noisy data [9].

H. Logistic Regression

Logistic regression is statistical software to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation. This type of analysis is then used for predicting the event and the overall performance of this technique shows good efficiency [11], [16].

I. K-Nearest Neighbours (K-NN)

K-NN models are very useful in handling both regression and classification problems. The KNN model uses the feature of resemblance to predict the value of some new data points. The allocation of values to the new data points is how strongly it resembles the points in the training set. The steps involved in K-NN are:

1. Choose the K number of the neighbours.
2. Compute the Euclidean distance of K neighbours.
3. Take the K-NN as per the designed Euclidean distance.
4. Among these neighbours, count the data points in every category.
5. Allocate the new data points to that category holding the maximum neighbours.
6. Finally construct the K-NN model.

The outcome of this KNN model provides an effective and efficient approach to detect Autism, and it reduces the time and cost of predicting Autism [17].

Table 1 presents an overview of various existing techniques used for identifying the Autism disorder. It tabulates the study based on the parameters, platform, dataset, data science techniques used, and their features and limitations.

Table 1: Overview of Various Existing Models

Year	Parameters / Platform	Dataset / Techniques	Features	Limitations
2019	1. Accuracy 2.Precision 3.Recall **Platform:** Artificial Intelligence	**Dataset:** Real time Childhood Autism Rating Scale (CARS) data **Techniques:** 1.Naive Bayesian, 2.Decision Tree,	1. It identifies the early stages of Autism. 2. The decision tree has the highest degree of accuracy.	1. Only useful for small datasets. 2. It has a limited set of classificatio n models.

		3.K Nearest Neighbor 4.Support Vector Machine.		
2019	1.Accuracy 2.Specificity 3.Sensitivity 4.Precision 5.False Positive Rate(FPR) **Platform:** Android application with the help of Amazon Web Service (AWS)	**Dataset:** AQ-10 dataset **Techniques:** 1.Merging Random Forest-CART 2.Merging Random Forest-ID3	1. Offers an effective and efficient method for detecting Autism traits in different age groups. 2. It reduces time and cost.	1. Less suitable for large datasets. 2. Inspection applications are not designed for the age group of below three.
2019	1.Accuracy 2.Kappa Statistics 3.AUROC 4.Sensititvity 5.Specificity 6.Logloss **Platform:** R Programming	**Dataset:** 1.Kaggle 2. UCI ML Repository **Techniques:** 1.Adaboost 2. FDA 3. C5.0 4.Gimboost 5.LDA 6.MDA 7.PDA 8.SVM 9.CART	1. Using a variety of feature selection and ranking methods, it is highly predictive for Autism.	1. Some of the classifiers produced inconsistent results. 2. The currently offered ASD data was insufficient to fully resolve this ASD prediction.
2020	1.Specificity 2.Sensitivity 3.Accuracy **Platform:** Python	**Dataset:** UCI ML Repository **Techniques:** 1.Logistic Regression, 2.SVM, 3.Naive Bayes, 4.KNN, 5.ANN, 6.CNN	1. It is preferable to handle missing values. 2. It has a high degree of accuracy. 3. It takes the least amount of time.	1. When comparing SVM, ANN, and CNN. SVM and ANN have lower accuracy.
2020	1. Accuracy 2. Sensitivity 3. Specificity 4. Precision 5.F1-Measure **Platform:** WEKA	**Dataset:** Data is collected by a mobile application **Techniques:** 1. RIPPER 2. RIDOR 3. Nnge 4. Bagging 5.CART 6. RML 7.PRISM 8.Adaboost 9. C4.5	1. Overall, Machine Learning techniques performed well.	1. One of the limitations of this work is that it does not include instances involving toddlers, which are uncommon and difficult to obtain.
2020	1. Accuracy 2.F-Measure 3. Recall 4. Precision	**Dataset:** UCI ML Repository	1. It demonstrates a high level of	1. It is not appropriate for other types of

	Platform: Java combination with the WEKA tool	**Techniques:** 1.CMAR 2.CBA 3.FACA 4.MCAR 5.FCBA 6.ECBA 7.WCBA	classification accuracy. 2. It shortens the screening process. 3.It identifies the fewest number of ASD codes, reducing the problem's complexity.	Autism medical data sets.
2020	1.Brierscore 2.Precision 3.Recall 4.F-Score **Platform:** 1.Scikit 2.Tensorflow	**Dataset:** Autism Brain Imaging Data Exchange Repository **Techniques:** 1.Decision Tree 2.Random Forest 3.SVM 4.K-NN 5.Naïve Bayes 6.Logistic Regression	1. Logistic and SVM models have the lowest Brierscore. 2. Obtained the greatest accuracy (AUC Score of 94.52).	1. It consumes more time.
2020	1.Sensitivity 2.Specificity **Platform:** R version 3.5.1 in Rstudio 1.1.456	**Dataset:** Data came from four specialized ASD outpatient clinics in Germany **Techniques:** 1.SVM model 2.ADOS Algorithm	1. It demonstrated high specificity and sensitivity. 2. It greatly aided in the complicated diagnostic process of Autism.	1. The ADOS was usually considered in clinical decision-making, but it did not always determine the diagnosis.
2020	1.Accuracy 2.Precision 3.F1 score 4.Recall **Platform:** Python with Keras data processing package	**Dataset:** UCI ML Repository **Techniques:** Adam algorithm is used for this neural network.	1. Effective in handling the noise data.	1. It is less suitable for large datasets. 2. It provided less accuracy when using a large dataset.
2020	1.Precision 2.F1 Score 3.Accuracy **Platform:** WEKA	**Dataset:** UCI ML Repository **Techniques:** 1.Logistic Regression 2. KNN 3.Random Forest	1. Effective for preprocessing methods.	1. It provided less precision.

IV. RESULT AND DISCUSSION

This section intends to list the performance comparison of various models used in this study proposed by various researchers down the years based on their accuracy.

Table 2: Comparison of Enlisted Techniques Based on Accuracy

Proposers of The Model	Data Science Techniques Used	Accuracy
C.S.Kanimozhiselvi et al.	1. Decision tree algorithm	The training set had the highest accuracy of 100 percent, and the test set had the highest accuracy of 96 percent.
Kazi Shahrukh Omar et al.	1. Merged Random Forest-CART and Decision Tree-CART algorithm	It predicted Autism, in the case of children, adolescents, and adults, with 92.26%, 93.78%, and 97.10% accuracy respectively.
Tania Akter et al.,	1.SVM (toddler dataset) 2.Adaboost (children dataset) 3.Glmboost (adolescent dataset)	All of these algorithms provided 100% accuracy.
Suman Raja and Sarfaraz Masood	1.Convolutional neural network (CNN)	It performed best for 99.56 %, 98.30 %, and 96.88 % accuracy in adults, children, and adolescents.
FadiThabtah et al.,	1.Rule-Machine Learning (RML)	It obtained 95% accuracy.
Jaber Alwidian et al.,	1.Weighted Classification Based on Association Rules (WCBA)	The accuracy rate was 86 %.
Roopa B. S and R. Manjunatha Prasad,	1. SVM 2.Logistic Regression	The classification accuracy was 88%.
Charlotte Küpper et al.,	1.SVM model	The rate of the accuracy of the model was 90%
Lakshmi B and Kala A,	1.Neural networks model	It provided 90% of accuracy.
Devika Varshini G and Chinnaiyan R,	1. KNN 2.Logistic regression 3. Random forest	The accuracy of KNN was 69.2%, while the accuracy of logistic regression and random forest classifiers was 68.601%.

According to the table above, the model proposed by Tania Akter et al. provided the highest accuracy when compared to others. The Support Vector Machine, Adaboost, and Glmboost models continue to outperform in terms of autism prediction.

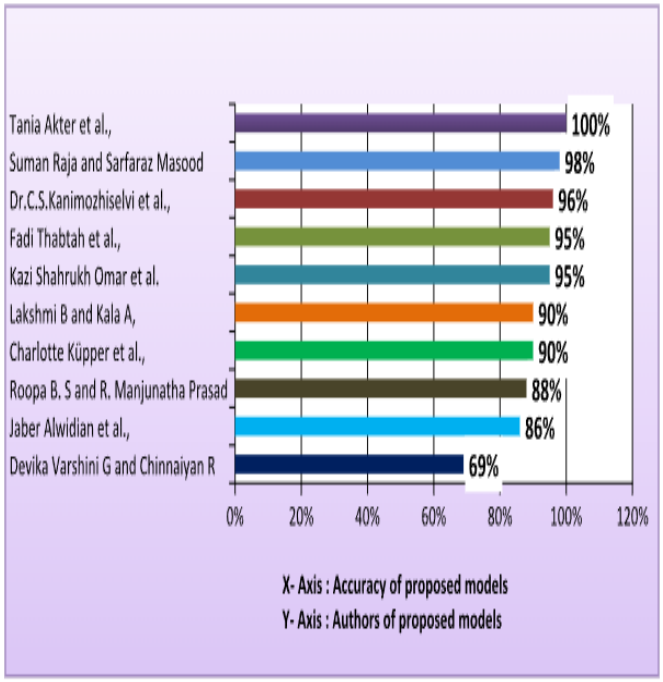

Figure 2: Performance comparison of the various enlisted techniques.

Finally, all of the proposed techniques are compared based on accuracy, which shows that the SVM, Adaboost, and Glmboost outperform with accuracy of 100 %, while the Convolutional neural network (CNN) came in second with accuracy of 98 percent. Finally, the Decision Tree Algorithm and Rule-Machine Learning (RML) came in last with an accuracy of 95%.

V. CONCLUSION

This paper summarizes various data science techniques for the prediction of Autism Spectrum Disorder. SVM, Adaboost and Glmboost outperformed well than the other data science techniques. It was examined using six commonly used statistical measures, including Accuracy, Kappa Statistics, AUROC, Sensitivity, Specificity, and Logloss. The performance of these techniques was high in all of the proposed models, providing a strong indication of the potential power of data science techniques in serving such a critical domain. According to the findings of the study, one may strongly accept the importance of early detection of

Autism Spectrum Disorder with the least amount of time, money, and complexity.

VI. REFERENCES

[1] Lord, Catherine et al., "Autism spectrum disorder", Lancet (London, England) vol. 392,10146 (2018): 508-520. doi:10.1016/S0140-6736(18)31129-2.

[2] Lisa Campisi, Nazish Imran, Ahsan Nazeer, Norbert Skokauskas, Muhammad Waqar Azeem, "Autism spectrum disorder. *British Medical Bulletin*" Volume 127, Issue 1, September 2018.

[3] Autism, Available at: http://www.Autismspeaks.org/what-Autism.

[4] Thabtah, Fadi, FiruzKamalov, and Khairan Rajab, "A new computational intelligence approach to detect autistic features for Autism screening", International journal of medical informatics 117: 112-124(2018).

[5] Dr.C.S.Kanimozhiselvi, Mr.D.Jayaprakash and Ms.K.S.Kalaivani, "Grading Autism Children Using Machine Learning Techniques", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 5 (2019) pp. 1186-1188.

[6] Kazi Shahrukh Omar, Prodipta Mondal, Nabila Shahnaz Khan, Md. Rezaul Karim Rizvi and Md Nazrul Islam, " A Machine Learning Approach to Predict Autism Spectrum Disorder" 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

[7] Tania Akter, Md. Shahriare Satu, Md. Imran Khan , Mohammad Hanif Ali , Shahadat Uddin, Pietro Lió, Julian M. W. Quinn , and Mohammad Ali Moni, "Machine Learning-Based Models for Early Stage Detection of Autism Spectrum Disorder", IEEE Access, Received October 10, 2019, accepted October 30, 2019, date of publication November 11, 2019.

[8] Suman Raja, Sarfaraz Masood, "Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques" Published by Elsevier B.V, International Conference on Computational Intelligence and Data Science (ICCIDS 2019).

[9] FadiThabtah and David Peebles, "A new machine learning model based on induction of rules for Autism detection" Health Informatics Journal 2020, Vol. 26(1) 264–286.

[10] Jaber Alwidian, Ammar Elhassan, Rawan Ghnemat, "Predicting Autism Spectrum Disorder using Machine Learning Technique", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-5, January 2020.

[11] Roopa B. S., R. ManjunathaPrasad, "Identification of Best Fit Learning Models Based on Calibration for Better Classification of Autism" International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-3, February, 2020.

[12] Charlotte Küpper, SannaStroth, NicoleWolf, Florian Hauck, Natalia Kliewer, Tanja Schad-Hansjosten, Inge Kamp-Becker, LuisePoustka, VeitRoessner, Katharina Schultebraucks & Stefan Roepke, "Identifying predictive features of Autism spectrum disorders in a clinical sample of adolescents and adults using machine learning", Scientific Reports | (2020) 10:4805 | https://doi.org/10.1038/s41598-020-61607-w.

[13] Lakshmi B, Kala A, "Prediction of Autistic Spectrum Disorder Based on Behavioural Features Using Machine Learning", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072.

[14] Devika Varshini G, Chinnaiyan R, "Optimized Machine Learning Classification Approaches for Prediction of Autism Spectrum Disorder" Ann Autism Dev Disord. 2020; 1(1): 1001. Copyright © 2020.

[15] N. S. Khan, M. H. Muaz, A. Kabir, and M. N. Islam, "Diabetes predicting health application using machine learning", In 2017 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE). IEEE, 2017, pp. 237–240.

[16] Thabtah, Fadi, "Machine learning in autistic spectrum disorder behavioral research: A review and ways forward" (2018). Informatics for Health and Social Care : 1-20.

[17] Mujthaba G.M, Abdalla Al, MajurKolhar, Mohamed Rahmath, "Data Science Techniques,Tools and Predictions", International Journal of Recent Technology and Engineering(IJRTE), volume-8, issue-6, March 2020.

