Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777753

IDENTIFYING ASD BY EARLY SIGNS OF BRAIN DISRUPTIONS

[1]Rajendra C J, [2]D S Suresh
[1]Research Scholar, Department of ECE, Channabasaveshwara Institute of Technology, Gubbi, Tumkur, Affiliated to Visveswaraya Technological University, Belagavi, India
[2]Professor, Department of ECE, Channabasaveshwara Institute of Technology, Gubbi, Tumkur, Affiliated to Visveswaraya Technological University, Belagavi, India

Abstract:-
Autism is a neurological disorder that has long-term consequences for a person's interactions and communiqué with others. Autism is a "behavioral disease" which can be identified at any age. Symptoms usually arise within the first 2 years of life. According to the ASD issue, it begins in childhood and progresses through puberty and maturity. With the growing usage of ML approaches in medical diagnosis research, this work investigates the feasibility of employing SVM, LR, KNN, Decision Tree, Random Forest, and XGBoost for predicting and assessing ASD difficulties in adults. On a publicly available non-clinically ASD dataset, the offered approaches are tested. The ASD screening in adolescent dataset has 801 cases and 20 attributes. Following the application of several machine learning approaches and the treatment of missing values, the findings strongly imply that Logistic Regression-based prediction models perform better on this datasets with an accuracy of 88%.
Keywords: Autism Spectrum Disorder (ASD), K-Nearest Neighbor (KNN), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), XGBoost (XGB).

I. INTRODUCTION

ASD is becoming increasingly prevalent among people of all ages. Early diagnosis of this neurological condition can substantially aid in the subject's mental and physical well maintenance. With the increased use of machine learning-based replicas in the estimation of numerous human diseases, primary diagnosis built on multiple health and physiological parameters appears to remain a possibility. This reason prompted us to become more interested in the diagnosis and investigation of ASD disorders in command to enhance treatment methods. ASD detection becomes tough because there are various different mental conditions with few symptoms that are extremely similar to those with ASD symptoms, making this task challenging.

Autism Spectrum condition is a developmental condition of the human brain. A individual with Autism Spectrum Disorder is often unable of social engagement and message with other people [1]. Here, a person's life is frequently influenced for the rest of their life. It's interesting to note that both ecological and hereditary reasons may have a role in the development of this condition. This problem's symptoms can appear as early as three years old and can last a generation. It is not feasible to completely treat a patient suffering from this condition;

however, the effects can be mitigated for a period of time if the signs are discovered early.

Scientists have yet to classify the precise origins of ASD by presuming that human genetic factor are to blame. Human genes have an impact on development via altering the environment. There are various risk factors that influence ASD, such as low birth weight, having a brotherly with ASD, and having elderly parents, among others. Instead, there are about social interaction and communication issues such as untimely laughing and giggling, lack of pain sensitivity, inability to create adequate eye contact, lack of engagement with people, desire to live alone, and so on.

Individuals with ASD struggle with limited interests and regular repetition of actions. The list that follows provides concrete instances of the categories of actions.

• Excessive repetition of specific actions, such as repeating words or phrases.
• The individual will be upset if their routine changes.
• Taking a passing interest in specific aspects of the issue, such as numbers, facts, and so on.
• Less susceptible to light, noise, and other stimuli than another individual.

The most crucial things to take to reduce the signs of ASD and improve the quality of life of ASD sufferers are primary detection and treatment. Though, there is no medical test process for detecting autism. Observation is generally used to identify ASD symptoms. ASD signs are typically detected by parents and instructors in older and adolescent students who attend school. Following that, the school's special tutoring team evaluates ASD symptoms. This school staff suggested that these youngsters consult their doctor for any necessary tests. Adults have a more difficult time detecting ASD symptoms than grownup kids and adolescents since certain ASD symptoms overlap with other mental health conditions. It is easier to identify behavioral changes in a kid through observation because they can be noticed as early as six months of age, as opposed to Autism precise brain imaging, which can be seen after two years of age.

This paper's contents are categorized in to following sections: Section 1 provides an impression of the ASD condition then the obstacles that the patients confront. Section 2 is a review of recent literature in which certain representations for ASD identification have been created. Section 3 defines the datasets used, and Section 4 describes individual module of the approach. Section 5 presents and discusses the outcomes of numerous experiments and section 6 presents the conclusion.

II. Literature Survey

An approach for identifying Autism using optimal behavior sets is proposed [3]. Binary firefly feature range wrapper based on swarm intelligence was tested in this article using an ASD dataset through 21 cases from UCI ML library. The alternate theory of the trial asserts that a machine learning model can achieve improved classification accuracy with fewer feature subsets. It was observed that 10 of the 21 features in the ASD dataset are enough to differentiate between ASD and non-ASD cases using a single-objective binary firefly feature assortment framework based on Swarm intelligence. The results produced with this method validate the theory by providing an average exactness in the range of 92.12%-97.95% with best feature subsets, which is almost equivalent to the average exactness obtained with the entire ASD dataset.

[8] presents a DSM-5-based ASD screening technique based on Machine Learning Adaption. A screening instrument was used in ASD to achieve 1 or more goals. In this effort, the investigator conferred the ASD ML classification and its merits and downsides. By employing the DSM-IV manual rather than the DSM-5 text, the researcher intended to emphasize the issues connected with present ASD screening approaches, as well as the reliability of such tools.

The authors [13] presented an ASD study utilizing Classification Methods. The primary area of this research was to classify the problem and levels of autism. To examine the student's actions and social interaction, SVM and Fuzzy methods with WEKA tools were used.

In article [14], it proposes a method for classifying autism by searching for the smallest group of traits. Here, the scientists cast-off a ML method to assess the clinical calculation of Autism syndrome. The ADOS was castoff to a subgroup of children with autism-related behaviour. ADOS is made up of four modules. 8 different machine learning methods were utilized in this work to discover stepwise backward characteristics on score pages from 4540 persons. It indicates ASD risk using 9 of 28 actions from Module 2 & 12 of 28 actions from Module 3 with 98.27% and 97.66% accuracy, respectively.

Through mimicry, ML classifiers were utilized to distinguish autistic adults [11]. This study sought to examine the fundamental issue of discriminative test settings and kinematic characteristics. The collection includes the hand movements of 16 ASC patients. Machine learning algorithms were utilized to extract 40 kinematic constraints from a set of eight imitation scenarios. The feasibility of employing machine learning algorithms to investigate high-dimensional facts and the investigative classification of autism in a small model size is demonstrated in this study. The sensitivity rates estimated by RIPPER on the AQ-Adolescent dataset using the features Va (87.30%), CHI (80.95%), IG (80.95%), Correlation (84.13%), CFS (84.13%), and "no feature selection" (80.00%).

There is a clear need to discover the feasibility of using ML based representations for the recognition of Autism in the human inhabitants from the preceding section. The majority of the effort covered above use traditional machine learning algorithms, which limit its performance. For this reason, the performance of numerous machine learning models with separate models have been developed and compared for each population set (explained further below).

III. DATASET

The dataset for this study was found in the UCI Repository, which remains open to the public [12]. Dataset details are mentioned below in table 1.

Table 1: Details of dataset

Sl. No.	**Name of Dataset**	**Source**	**Attribute Type**	**No. of Attributes**	**No. of Cases**
1	ASD adolescent Screening Data	UCI ML Repository	Categorical and binary	20	801

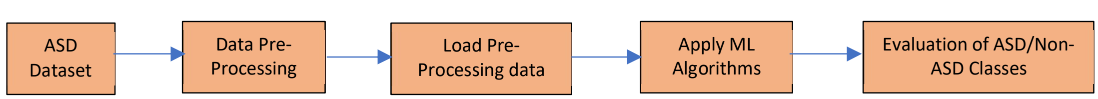
The dataset has 20 attributes which is used for estimation. The list of attributes are as follows:

Table 2: List of Attributes

Attribute Id	Description
1	ID
2-10	Based on the screening technique answers of 9 questions
11	Age
12	Gender
13	Anyone in the family had jaundice?
14	Any of the family member earlier diagnosed with autism
15	The user's country of residence
16	Is the user familiar with the screening application?
17	Screening Application Result
18	Age Discrimination
19	Relationship
20	Screening Points

IV. Methodology

The proposed workflow, which include data pre-processing, training/testing through specified models, results evaluation, and ASD prediction as been shown in figure 1.



(I/p) (O/p)

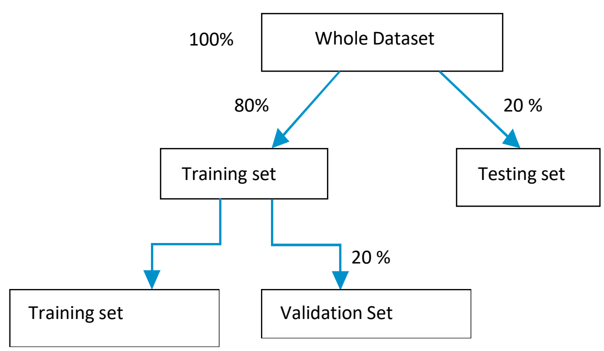
Figure. 1. ASD detection steps

A. Data Pre-Processing

It is a process of converting raw data into a understandable and intelligible format. Because of faults and null values, everyday data is frequently partial and inconsistent. A decent pre-processed data set will always produce a good outcome. To handle imperfect and unpredictable data, several data pre-processing approaches are utilized, such as treating missing values, outlier identification, data discretization, data drop and so on. The imputation approach was used to deal with missing values in these datasets.

B. Training / Testing Model

Complete dataset has been divided into two sections, one for training and one for testing, with a ratio of 80:20. Again, training data has been divided into two sections for cross-validation purposes. The training dataset is split into two halves, with an 80:20 split for the validation dataset. Figure 2 depicts the final training, testing, and validation sets used for classification.



80 %

Figure. 2. Final Training / Testing and Validation Sets

i) Support Vector Machine (SVM)

It is a classification and regression approach that uses linear administered ML. It solves pattern recognition problems. It does not cause the overfitting problem. SVM distinguishes between programs by significant a choice boundary [19]. A simple SVM diagram is shown in figure 3.

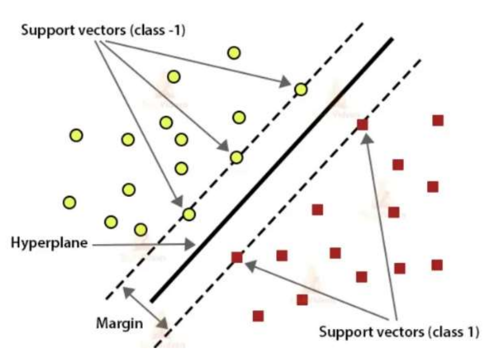


Figure. 3. SVM classifier

ii) Decision Tree (DT)

A decision tree is a graphical representation of a decision-making process that is castoff in ML and data mining. It is a tree-like structure that represents several possible outcomes based on a series of actions or occurrences. The root node, which symbolizes the initial decision or event, is at the top of the decision tree. Each node that follows represents a decision or event that can result in a variety of outcomes. These nodes are linked by branches that represent the many courses that can be taken in response to the decision or occurrence. Each decision node has a set of alternative outcomes, and the probability of each option is estimated based on the available data. The final decision or conclusion is chosen by following the path through the tree with the best likelihood of success as shown in figure 4.

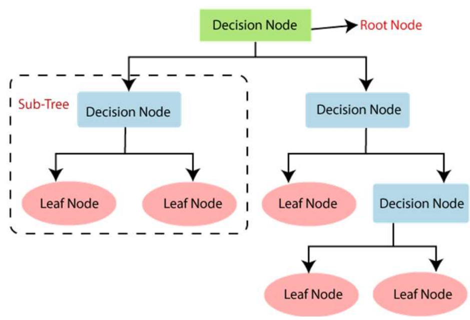


Figure. 4. Basic Decision Tree structure

iii) Random Forest (RF)

Random Forest is a machine learning algorithm that is used for classification, regression, and other tasks involving supervised learning. It is an ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of predictions. In a random forest, multiple decision trees are created using a random selection of features and data samples from the training set. Each decision tree is trained independently on a different subset of the data, and the final prediction is made by combining the predictions of all the individual trees. One of the key advantages of random forests is that they are less prone to overfitting than individual decision trees. This is because each decision tree is trained on a different subset of the data, which helps to reduce the bias and variance in the model. Additionally, the random selection of features helps to reduce the correlation between the trees, which further improves the accuracy and robustness of the model. A simple diagram of RF is shown in figure 5.

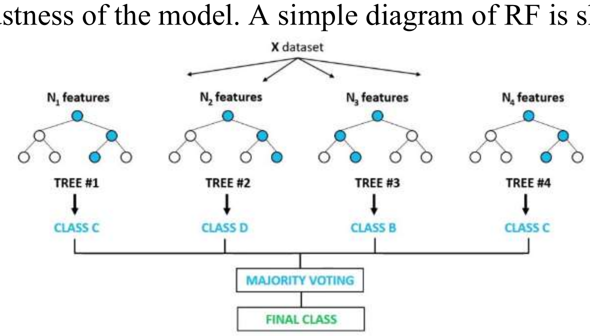


Figure. 5. Random Forest Structure

iv) Logistic Regression (LR)

It is a reversion method for analyzing binary reliant on variables. Its o/p value can remain in either zero or one form. It is employed in the case of the continuous value dataset. It describes the link between a single dependent binary variable and a single nominal or ordinary variable. The sigmoidal function can be used to express it.

v) K- Nearest Neighbour (KNN)

It is the most straightforward supervised learning technique and used to solve classification and reversion problems, and assumed that equivalent data is available nearby. The 'K' component denotes the number of seed points to be chosen. To reduce error, it must be properly picked. As a result, it is founded on the concept of resemblance, which can be measured in terms of distance, closeness, or propinquity. Euclidean distance is the most often used distance measurement.

vi) XGBoost

The XGBoost (Extreme Gradient Boosting) ensemble learning method combines numerous weak prediction models into a single powerful predictive model. To iteratively increase the performance of a predictive model, XGBoost employs a gradient boosting architecture. It starts by fitting an initial model to the data and then iteratively adds further models to the ensemble that correct for errors committed by prior models. Each new model is trained on the residuals (the difference amongst the predicted and actual values) of the preceding models, with the goal of lowering the ensemble's total error. One of XGBoost's primary advantages is that it is very scalable and capable of handling big datasets with high-dimensional features. A simple diagram of XGBoost is shown in figure 6.

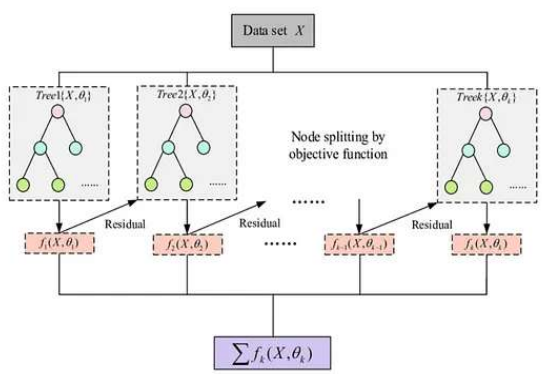


Figure. 6. An XGBoost Classifier

V. Result & Discussion

Precision, recall, and accuracy are calculated using the classification report and confusion matrix. How thoroughly the model is trained determines the outcome.

A. Performance Evaluation metrics

Performance measurement is essential for determining how well a categorization model works to achieve a goal. Performance evaluation measures are used to assess the classification model's efficacy and performance on the test dataset. It is critical to select the appropriate metrics to assess model performance, such as the confusion matrix, accuracy, precision, and recall. The performance measurements are calculated using the following formulas:

Table 3: Confusion Matrix Elements

$$Actual\ ASD\ Values = \frac{True\ Positive\ (TP) \qquad\qquad False\ Positive\ (FP)}{False\ Negative\ (FN) \qquad\qquad True\ Negative\ (TN)}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$\text{Accuracy} = \frac{TP + TN}{(TN + TP + FP + FN)} \tag{3}$$

For ASD, screening data for adolescent's, the experimental outcomes of some ML algorithms approach with all features range were shown. All 20 features are chosen to determine the precision, recall, and accuracy of the predicted model. The overall performance actions of all the ML classifiers are detailed below:

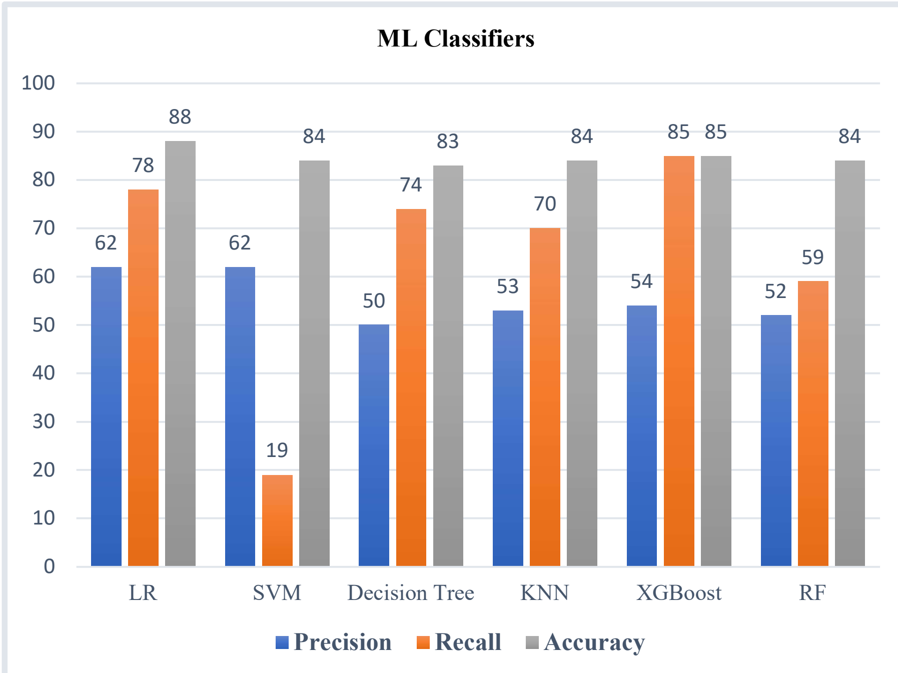


Figure 7: Adolescent Autistic Spectrum Disorder Screening Data Overall Results

On the novel dataset, the accuracy of multiple machine learning representations on the ASD adolescent's diagnosis dataset ranged from 83% to 88%. The Decision Tree classifier had the lowest accuracy of 83%. On the novel dataset, Logistic Regression gave a prediction accuracy of 88%. All ML algorithms' learning curves describe the prediction model's results. The Evaluation of model results with existing method [27] on ASD Screening Data for adolescent's is shown in figure 8.

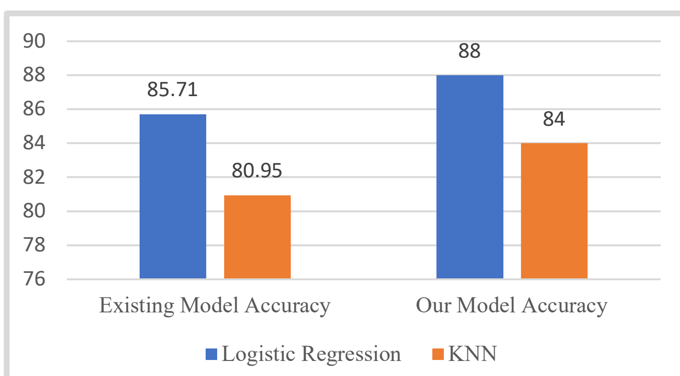


Figure 8: Evaluation of model results with existing method [27] on ASD Screening Data for adolescent's

VI. CONCLUSION

Using several machine learning techniques, this study aimed to detect autism spectrum disease via early symptoms of brain disturbances. Numerous performance evaluation measures were utilized to examine the performance of the ASD detection models developed on non-clinical datasets of adolescents. When compared to another recent work [27] on this problem, the Logistic Regression classifier with all of its feature attributes performed better after managing missing values. After accounting for missing values, both the LR and KNN-based models demonstrate prediction accuracy of roughly 88% and 84% for the ASD teenage dataset, respectively.

In this proposed work, recognition of ASD by early signs of brain disruptions was tried using several machine learning techniques. Several performance evaluation measures were employed to examine the performance of the models constructed for ASD identification using teenage non-clinical datasets. When compared to alternative recent work [27] on this problem, the Logistic Regression classifier with all of its feature properties performed better after handling missing values. After accounting for missing values, both the LR and KNN-based models show a prediction accuracy of roughly 88% and 84% for the ASD teenage dataset, respectively. The findings strongly suggest that, rather than the other conventional ML classifiers provided in previous works, an LR-based model can be utilized to diagnose autism spectrum disorder in early symptoms of brain problems.

Acknowledgment
The authors would like to acknowledge Department of Electronics and Communication Engineering Research Centre, Channabasaveshwara Institute of Technology affiliated to Visveswaraya Technological University, Belagavi for their extended support and cooperation to carry out this research work.

REFERENCES

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

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

[3] Vaishali, R., and R. Sasikala. "A machine learning based approach to classify Autism with optimum behaviour sets. (2018) " International Journal of Engineering & Technology **7(4)**: 18.

[4] Constantino, John N., Patricia D. Lavesser, Y. I. Zhang, Anna M. Abbacchi, Teddi Gray, and Richard D. Todd. (2007) "Rapid quantitative assessment of autistic social impairment by classroom teachers." Journal of the American Academy of Child & Adolescent Psychiatry **46(12)**: 1668-1676.

[5] Daniel Bone, Matthew S. Goodwin, Matthew P. Black, Chi-Chun Lee, Kartik Audhkhasi, and Shrikanth Narayanan. (2015) "Applying machine learning to facilitate autism diagnostics: pitfalls and promises." *Journal of autism and developmental disorders* **45(5)**: 1121-1136.

[6] Dennis Paul Wall, J. Kosmicki, T. F. Deluca, E. Harstad, and Vincent Alfred Fusaro. (2012) "Use of machine learning to shorten observation- based screening and diagnosis of autism." *Translational psychiatry*, ***2*(4):** e100.

[7] Dennis P. Wall, Rebecca Dally, Rhiannon Luyster, Jae-Yoon Jung, and Todd F. DeLuca. (2012) "Use of artificial intelligence to shorten the behavioral diagnosis of autism." *PloS one*, ***7*(8)**: e43855.

[8] Fadi Thabtah. (2017). “Autism spectrum disorder screening: machine learning adaptation and DSM-5 fulfillment.” In *Proceedings of the 1st International Conference on Medical and Health Informatics*, pp. 1-6. ACM.

[9] Daniel Bone, Chi-Chun Lee, Matthew P. Black, Marian E. Williams, Sungbok Lee, Pat Levitt, and Shrikanth Narayanan.(2014). “The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody.” *Journal of Speech, Language, and Hearing Research*, ***57*(4)**: pp.1162-1177.

[10] Fadi Thabtah. (2017) “ASD Tests. A mobile app for ASD screening.” www.asdtests.com [accessed December 20th, 2017].

[11] Baihua Li, Arjun Sharma, James Meng, Senthil Purushwalkam, and Emma Gowen. (2017) “Applying machine learning to identify autistic adults using imitation: An exploratory study.” *PloS one*, ***12*(8)**: e0182652.

[12] Fadi Fayez Thabtah (2017), “Autistic Spectrum Disorder Screening Data for Adult”., https://archive.ics.uci.edu/ml/machine-learning- databases/00426/.

[13] M. S. Mythili, and AR Mohamed Shanavas. (2014) “A study on Autism spectrum disorders using classification techniques.” *International Journal of Soft Computing and Engineering (IJSCE)*, ***4***: 88-91.

[14] J. A. Kosmicki, V. Sochat, M. Duda, and D. P. Wall. (2015) “Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning.“ *Translational psychiatry*, **5(2)**: e514.

[15] Fadi Fayez Thabtah (2017), “Autistic Spectrum Disorder Screening Data for children,” https://archive.ics.uci.edu/ml/machine-learning- databases/00419/ ,2017

[16] Fadi Fayez Thabtah (2017), “Autistic Spectrum Disorder Screening Data for Adolescent”, https://archive.ics.uci.edu/ml/machine-learning- databases/00420/.

[17] John, George H., and Pat Langley. (1995). “Estimating continuous distributions in Bayesian classifiers.” In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* (pp. 338-345). Morgan Kaufmann Publishers Inc.

[18] Quinlan, J. R. (1993) “Program for machine learning.” *C4. 5*.

[19] Keerthi, S. Sathiya, Shirish Krishnaj Shevade, Chiranjib Bhattacharyya, and Karuturi Radha Krishna Murthy. (2001) “Improvements to Platt's
SMO algorithm for SVM classifier design. “ *Neural computation*, ***13*(3):**637-649.

[20] J. C. Platt (1999) “12 fast training of Support Vector Machines using sequential minimal optimization,” Adv. Kernel method, pp. 185-208,

[21] Sankar K Pal., and Sushmita Mitra. (1992) “Multilayer perceptron, fuzzy sets, and classification.” *IEEE Transactions on neural networks*, ***3*(5):** 683-697.

[22] David W Aha., Dennis Kibler, and Marc K. Albert. (1991) “Instance-based learning algorithms. *Machine learning*,” ***6*(1)**: 37-66.

[23] Katherine Gotham, Susan Risi, Andrew Pickles, and Catherine Lord. (2007) “The Autism Diagnostic Observation Schedule: revised algorithms for improved diagnostic validity.” *Journal of autism and developmental disorders* ***37*(4)**: p.613.

[24] Sarfaraz Masood, Abhinav Rai, Aakash Aggarwal, Mohammad Najmud Doja, and Musheer Ahmad. (2018) “Detecting distraction of drivers using convolutional neural network.” *Pattern Recognition Letters*.

[25] Sarfaraz Masood, Adhyan Srivastava, Harish Chandra Thuwal, and Musheer Ahmad.

(2018). “Real-time sign language gesture (word) recognition from video sequences using CNN and RNN.” In *Intelligent Engineering Informatics* (pp. 623-632). Springer, Singapore.

[26] Sarfaraz Masood, Harish Chandra Thuwal, and Adhyan Srivastava. (2018). “American Sign Language character recognition using convolution neural network. “In *Smart Computing and Informatics* (pp. 403-412). Springer, Singapore.

[27] Suman Raj and Sarfaraz Masood, “Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques” in International Conference on Computational Intelligence and Data Science (ICCIDS 2019)