



Performance Analysis of various ML Algorithms in Detecting ASD

¹Goutham M, ²Shambulinga M, ³K. Viswavardhan Reddy,

¹gouthamm ldc19@rvce.edu.in, ²shambulingam@rvce.edu.in, ³viswavardhan@rvce.edu.in,

^{1,2} Department of Electronics and Telecommunication Engineering, RV College of Engineering, Bengaluru, India

ABSTRACT

Autism Spectrum Disorder (ASD) is gaining momentum faster. Detecting autism symptoms through screening tests can be very costly and time consuming. With advances in Artificial Intelligence and Machine Learning (ML), autism can be diagnosed at a very early stage. Although there have been many studies using different methods, these studies have not provided any definitive conclusions about how autism symptoms should be assessed in the context of different age groups. Therefore, the aim of this paper is to propose an effective assessment model based on ML technology and to develop a user interface for ASD assessment for those of any age. As a result of this research, the Autism Prediction Model was developed by combining the user interface based on the Random Forest-Cart (Classification and Regression Tree) and Random Forest-ID3 (Interactive Dichotomizer 3) and the proposed prediction model. Developed. The proposed model AQ10 dataset was evaluated with 250 actual datasets collected from individuals with and without autistic features. The evaluation results showed that the proposed prediction model would provide better results in terms of accuracy, precision, sensitivity, accuracy and false positive rate (FPR).

Keywords—Autism Spectrum disorder, Machine Learning, Performance analysis, Random forest, Decision Tree.

Introduction

Autism Spectrum Disorder (ASD) is a mental disorder that causes severe social, communication and behavioral challenges that often appear in the first two years of life and gradually develop over time. People with autism experience a variety of focus conflicts, learning disabilities, mental health problems such as depression, anxiety, motor problems, neurological problems and more. Tests show that both genes and nature play an important role. The current autism prevalence rate around the world is widespread and is growing at an alarming rate. According to the WHO, 1 in 160 people has autism spectrum disorder. Some people with this condition are able to live independently, while others need lifelong support and care. Diagnosis of autism requires considerable expense and time. Autism Spectrum Disorder is a neurodevelopmental disorder that affects a person's communication, communication skills and learning skills. Although autism is diagnosed at any time, its symptoms usually appear within the first two years of life and develop over time. Autism patients experience a wide variety of challenges such as concentration difficulties, learning disabilities, mental health problems such as anxiety, depression, motor problems, neurological problems and more.

Autism spectrum disorder can be controlled by counseling people with appropriate medications if detected early. , Which can prevent the patient's condition from deteriorating further and reduce the long-term costs associated with late diagnosis. Therefore there is a great need for a useful, accurate and simple diagnostic tool to identify the symptoms in a person and determine if a person definitely needs an autism syndrome test. In this paper we will use machine learning to find a comprehensive set of conditions to diagnose autism spectrum disorder. Considering the data of the past decades, several studies have been published in support of the hypothesis of increased ASD cases. The prevalence in some population studies in North America varies, from 1 in 68 children in 2012 or 1 in 59 in 2014. According to a recently published study, the prevalence of ASD was 13.4 per 1000 children under the age of 4 in 2010, 15.3 in 2012 and 17.0 in 2014. It is proposed that the median age of diagnosis is three or four years of age or later. For children with low socioeconomic status with or without a previous family history of ASD.

Applying the Wisdom of Reducing the Diagnosis of Autism [1] Use ML technique to study the genetic research of autism. Their analysis showed that 93 out of 7 factors are sufficient for ADI-R to diagnose autism with 99.9% accuracy. Autism Diagnostics Machine Learning [2] provides a way to build a vector support device (SVM) that can help diagnose autism. Their differences were between 85.6% and 94.3% sensitivity and between 80.9% and 89.3%. Diagnosis of Autism Spectrum Disorder Using Practice [3] The purpose of this paper is to develop an autism prediction model using ML techniques and to develop a web system that can accurately assess individual autism symptoms. Using ML technique to identify patterns of people with ASD. [4] Describes the reliance on technology and smartphones that are essential to having a technology identity that can control industrial and domestic applications using IoT. ASD. Spectrum Disorder (ASD) is a common growth disorder that affects people with varying degrees of disability. Currently, there is a lot of research on overweight in children with autism. A New Approach to Diagnosis of Autism [6] Trained physicians diagnose autism in children as young as two years of age, and proven effective treatment begins immediately. Unfortunately, he points out that the average age of diagnosis of autism in the United States is 4.3 years. Automatic identification and recording of behavioral patterns in children with autism spectrum disorder [7] Infrastructure is designed to record, identify, and label behavioral disorders in children with autism spectrum disorder (ASD). The system consists of 2 wearable and stable sensor forums. Depression response in parents of children with autism spectrum disorder [8] Books on this topic suggest that the abilities of parents with children with autism differ from those of children affected by certain age or other conditions.

Since early intervention provides the best opportunity for healthy growth and lifelong benefits, this paper can help parents determine if their child has ASD at an early age. It is also important in the health sector as there is no good medical cause for this disease. Based on the Autism Speaks organization in the United States, ASD can be diagnosed by applying a behavioral test or questionnaire, which requires more time and effort from parents and physicians. Therefore, this work demonstrates the power of machine learning algorithms in identifying individuals with specific autistic spectrum disorder symptoms [9]. Furthermore, this paper aims to turn autism diagnosis into a faster process that will enable treatment to be provided in earlier and more effective stages of a child's development using machine-learning algorithms.

The organization of the rest of the paper is as follows. Section II displays related tasks. Section III briefly describes the proposed procedure. Section IV discusses ASD screening system implementation. In Section V, it outlines the conclusions and future scope from the proposed work.

METHODOLOGY

Various autism rating criteria have been developed over the past 30 years to diagnose autism spectrum disorder (ASD) in the previous stage. There are various tools and devices developed by psychologists and neuroscientists to confirm this at an early stage. Most tools focus on diagnosing users through screening methods.

A. Technical KDD Framework

The proposed cognitive innovation in the database framework of ASD detection features is shown in Figure 1. The Knowledge Discovery Process (KDD) process begins with understanding the symptoms and factors of the autism spectrum disorder discussed in the second section. The next three processes are related to data collection, data preprocessing and data conversion (190).

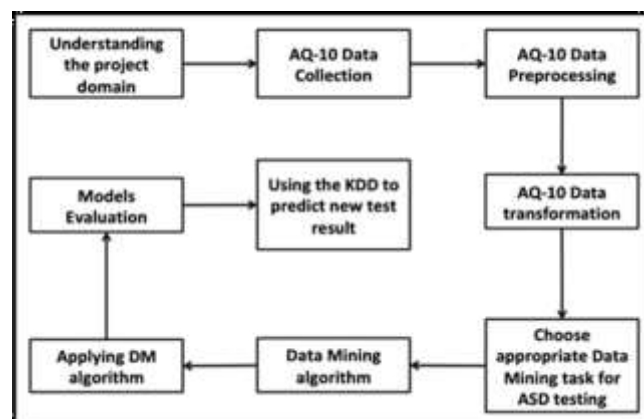


Fig. 1. KDD Process in Database

- 1) *Understanding the application domain*: This process defines the goals of the end-user and the environment in which the KDD process will take place with a full understanding of what should do
- 2) *Creating a data set*: This process checks the available data and then integrates it with additional obtained data into one data set for the knowledge discovery.
- 3) *Preprocessing*: The available data goes to the preprocessing step, which includes handling missing values and removing noise or outliers or duplicated data to enhance the reliability of the data.
- 4) *Data Transformation*: This step prepares and develops better data to create the best possible model. It includes customizing the data dimension by feature selection and record sampling..
- 5) *Choosing the appropriate data mining task*: The main goal of this process is to decide on which type of data Mining to use. Data mining types include clustering, classification, and regression, depending on the DM goal, either prediction or description.
- 6) *Choosing the data mining algorithm*: This step includes selecting the appropriate searching patterns to use. Each algorithm has parameters and tactics of machine learning.
- 7) *Applying the Data Mining algorithm*: To get a satisfying result in this process, it might require applying it several times.
- 8) *Evaluation*: The model with the found patterns is evaluated and interpreted concerning the goals mentioned in the first process. This step focuses on the usefulness and comprehensibility of the induced model.
- 9) *Using the discovered knowledge*: The final step is to try the knowledge into other systems for further action and make changes to the system and measure the effects.

B. Data collection

The data used in this paper are secondary data; AQ-10 ASD data as shown in Table 1. Data was collected through a mobile app from patient using the ASD AQ-10 diagnostic method and health experts relied on the final diagnosis. The primary purpose of the app is to collect useful data about ASD cases.

This data is later stored in MySQL database to investigate the key features that affect ASD diagnostics using data analysis. In addition, the data are

divided into three databases, as each age group has a specific diagnostic questionnaire [9].

C. Data preprocessing

The collected data usually cannot be used directly in performing the analysis process. Therefore, the raw data needs to be cleaned and performed in a usable format. Cleaning the data includes replacing or removing missing values and discretization for certain continuous variables such as the age of individuals; this step is called the data pre-processing. Before using the available data, all the redundant or unnecessary variables must be removed from the Database. The Null values, the columns that do not have unique values, must be removed too. This process will improve the model's prediction and raise the accuracy rate. The variables that have been removed in the ASD databases are as listed.

- ASD Screening type: this variable was added to split the data into spirit databases based on age categories.
- Reasons for taking the screening: this variable contains texts, and it did not add any value to the data analysis. Also, it will negatively affect the prediction model results.
- Language: since the app is available for people worldwide, the diagnostic test is also available with the most common languages.
- User: this variable represents the person who answers the screening instead of the child.
- Used app before: the users were asked this question to avoid attribute duplication.

D Data conversion

Data conversion, on the other hand, optimizes data size by feature selection to suit the needs of the model [10]. Both of these functions are required to achieve better performance and accuracy in machine learning prediction models. The data mining process is the next step in the KDD process, which involves data analysis and discovery algorithms that produce a specific calculation of the model on the data. To find the right machine-learning algorithm that provides the best ASD outcome assessment and strict search rules, the user must specify the type of data mining for the database. The rule step is searched by a classification system used to estimate the value of invisible cases. All possible machine-learning algorithms are evaluated using confusing metrics to ensure accuracy, sensitivity and specificity. Sharing the knowledge found with health professionals is the final process so that it can be used professional way of serving the community and helping parents to know the status of their children from an early age. [11]

E Data Description

In this paper, databases belonging to all age groups, i.e. infants, children and adults, are used. The dataset can be divided into ten behavioral questions for each age group and several variables that affect the final evaluation of the condition is used in clinical databases. [12]

Affecting variables include age, gender, race, jaundice and family history. Table 1 shows these variables along with the data type and description of each variable. The next three tables show ten variable details in 2 toddler, adolescent and child screening methods. There are ten attributes in three databases, the answer to which is yes or no. These ten questions are the most commonly seen symptoms in diagnosing people with ASD. [13]

Table 1 AQ-10 Children Screen features

SL No	Attributes	Values
1	Often notice noise when others cannot	0-1
2	Recognise the picture / direction / type of image	0-1
3	Social Interactions	0-1
4	Able to perform more than once atleast	0-1
5	Can study while speaking with other	0-1
6	Get information about any sources	0-1
7	Difficult to trust people	0-1
8	Easily identify wheather a person in good mood while speaking	0-1
9	Difficult in speech	0-1
10	Lack of concentration	0-1

IMPLEMENTATION

This section begins by describing the technical framework of the paper and identifies the data-mining objective of this paper. The following sections describe the data collection procedures, DM approach and specificity of the algorithm, the appropriate machine learning model to be used with the cases being monitored and how it is evaluated. The pre-defined goal is to develop a model that can identify a person with specific ASD symptoms using

specific machine learning techniques. The next practical work steps in this paper.[14]

1. To study the autism spectrum disorder, its symptoms, age group affecting it and various ML algorithms.
2. Collect the required data sets of ASD from various online sites.
3. To optimize the ML algorithms used for ASD assessment.
4. To evaluate the performance of various ML algorithms in the assessment of ASD

For test and training data sets applicable to various machine learning algorithms, the algorithms are discussed as follows

A.Implementing Algorithm (Algorithm-1)

Initially, the Cart Tree Certificate section was selected to create a predictive model. Initially, the root of the tree contains the entire database. The data is then sorted using the best feature. The partitioning process is repeated until the node contains data for a different label range. The sequence selection method is solved by Gini impurity and Information gain (IG) .

ALORITHIM 1: Decision Tree CART Classifier

```

1: features → {AQ - 10 questions; gender; inheritance}
2: classes→ {yes(autistic traits); no(no autistic traits)}
3: procedure BUILD TREE(rows)
4: for each possible features do
5:   calculate max gain
6: end for
7: if max gain = 0 then
8:   return leaf
9: end if
10: TrueRows; FalseRows→ Partition(rows)
11: TrueBranch→ Build Tree(TrueRows)
12: FalseBranch→ Build Tree(FalseRows)
13: return DecisionNode(TrueBranch; FalseBranch)
14:
15: procedure CLASSIFY(row,node)
16: if node = leaf then
17: return node.predictions
18: else
19: Iterate_Tree
20:endif

```

In the random forest, each node is classified using the best ones in a subset of randomly selected predictors. This particular strategy is more effective and powerful against overbalancing than many other classifiers, including discrimination analysis, vector support mechanisms and neural networks. To make the prediction model more accurate, a random forest cart classifier [Algorithm 2] is assigned. Here again the algorithm can be divided into two categories: creating a random forest [line no. 1-10] and splitting test data [line no. 12-28].

ALORITHIM 2: Random Forest CART Classifier

```

1: Same as Line 1-13 of Algorithm 1
2: procedure BUILD FOREST(rows,p,train_ratio)
3: tree_array→{ }
4: while p≠0 do
5: train→random(train_ratio * len(rows))
6: tree→ BUILD TREE(train)
7: tree_array:append(tree)
8: p→ p-1
9: end while
10: return tree_array
11:
12: procedure CLASSIFY(row; tree_array[ ]; p)
13: i→0; vote_yes→0; vote_no→ 0
14: while i ≠ p do
15: tree→ tree_array(i)
16: node→ root(tree)
17: if node = leaf then
18: if leaf.prediction = "Y es" then
19: vote_yes→ vote_yes + 1

```

```

20: else if leaf:prediction = "No" then
21: vote_no → vote_no + 1
22: end if
23: else
24: Iterate_tree
25: end if
26: i → i + 1
27: end while
28: return vote_yes > vote_no

```

To improve performance, a predictive model has been proposed combining the concept of random forest - CART with random forest - ID3 [Algorithm 3]. The proposed prediction model algorithm can be divided into two stages as before: generating structured integrated forest and segment test data. Its difference from 2 is that the randomness here is greatly enhanced by the production and addition of ID3 decision trees to the random forest [rows 3-13]. Algorithm 3 works better than Algorithm 2 because the increase in ID3 tree resolution is severely curtailed and thus further reduces the error compared to Algorithm 2.

ALORITHIM 3: Merged Random Forest Classifier

```

1: features → { AQ10 questions; gender; inheritanceg
2: classes → {yes(autistic traits); no(no autistic traits)g
3: procedure BUILD TREE ID3(rows)
4: for each possible features do
5: calculate max gain
6: end for
7: if max gain = 0 then
8: return leaf
9: end if
10: TrueRows , FalseRows → Partition(rows)
11: TrueBranch → Build Tree ID3(TrueRows)
12: FalseBranch → Build Tree ID3(FalseRows)
13: return DecisionNode(TrueBranch; FalseBranch)
14:
15: procedure BUILD TREE CART(rows)
16: for each possible features do
17: calculate max gain
18: end for
19: if max gain = 0 then
20: return leaf
21: end if
22: TrueRows , FalseRows → Partition(rows)
23: TrueBranch → Build Tree CART(TrueRows)
24: FalseBranch → Build Tree CART(FalseRows)
25: return DecisionNode(TrueBranch; FalseBranch)
26:
27: procedure BUILD FOREST(rows; p; train_ratio)
28: tree_array → {}
29: while p ≠ 0 do
30: train → random(train_ratio _ len(rows))
31: tree1 → BUILD TREE ID3(train)
32: tree2 → BUILD TREE CART(train)
33: tree_array:append(tree1)
34: tree_array:append(tree2)
35: p → p - 1
36: end while
37: return tree_array
38:
39: procedure CLASSIFY(row; tree_array[ ]; p)
40: i → 0; vote_yes → 0; vote_no → 0
41: while i ≠ p do

```

```

42: tree → tree_array(i)
43: node → root(tree)
44: if node = leaf then
45: if leaf:prediction = "Yes" then
46: vote_yes → vote_yes + 1
47: else if leaf:prediction = "No" then
48: vote_no → vote_no + 1
49: end if
50: else
51: Iterate_tree
52: end if
53: i → i + 1
54: end while
55: return vote_yes > vote_no

```

Evaluating the results of all appropriate algorithms will lead to selecting the appropriate machine-learning model. The data used in this paper are primarily focused on the diagnosis of individuals with ASD symptoms, with a number of variables based on AQ-10 that usually affect the outcome of the diagnosis. Therefore, the prediction model is considered a classification problem that may or may not lead to ASD. Therefore, a number of models that are appropriately monitored for a given task have been applied, the results are analyzed and evaluated. Before implementing these machine learning models, a feature selection process must be implemented to support evaluation results and model accuracy by removing vulnerable variables from the database.

EXPERIMENTAL RESULTS

This section provides an overview of the results obtained, the process to solve the research question and the visibility of those results. It focuses on feature selection methods and outcomes, measures to evaluate machine-learning models based on feature selection results, and standard rules for ASD detection derived from the best machine-learning models.

Accuracy and performance analysis

The performance of algorithm is better understood by knowing its performance parameters such as accuracy, sensitivity, precision, specificity scores from the confusion matrix generated. The following section gives a brief description's about the calculations and plotting a graph using a software packages in the research work.

The confusion matrix generated for naïve bayes is :

$$\begin{bmatrix} 460 & 10 \\ 30 & 420 \end{bmatrix}$$

True positive = 460

False positive = 10

False negative = 30

True negative = 420

Now, will generate a performance parameter table which consists the values of accuracy, precision, sensitivity and specificity as shown in table 2

Table 2 Performance parameter

Performance parameters	Result	Data set used
Accuracy	95%	All age groups
Precision	97%	All age groups
Sensitivity	93%	All age groups
Specificity	97%	All age groups

The confusion matrix generated for decision tree and random forest is

$$\begin{bmatrix} 564 & 0 \\ 0 & 316 \end{bmatrix}$$

True positive = 564

False positive = 0

False negative = 0

True negative = 316

Now, will generate a performance parameter table which is as shown in table 3

Table 3. Performance parameter

Performance parameters	Result	Data set used
Accuracy	100%	All age groups

B. Simulation and implementation

Python execution assesses the strengths of the statistical process and identifies the strengths and weaknesses of machine learning models using confusing matrices that lead to simulation results. In RStudio, the machine-learning model provides a confusing matrix as a model evaluation tool. Define the set of mathematical values used to apply the model and to determine the model's ability to select the best model to summarize and evaluate database set results. There are sets of values taken into account and that accuracy

B.Feature selection results and analysis

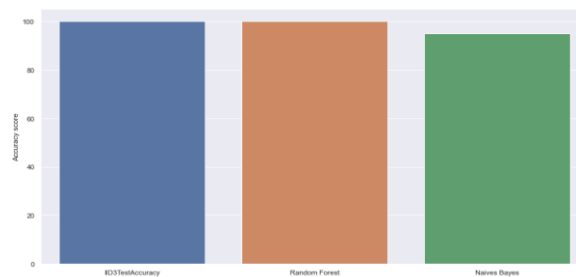
The predictive model of the ASD database is considered to be the classification-assessment problem, which maintains the variable labeled with the classification input variable. Feature selection algorithms basically estimate the relationship between ASD test results independently of each other variable in the database using two filter-based methods: chi-squared and mutual information. Code lines related to mutual information techniques, such as the Information Gain method, and Chi-Square were applied to estimate autistic characteristic features in all available datasets, and then compared the performance of each technique. Showed high correlation with AQ questions and ASD diagnosis results. The only variables in the feature selection techniques are the variables used in the machine Model learning to improve model performance. The accuracy achieved by each algorithm is shown in Figure [2].

```
scores = [ID3TestAccuracy,score_rf,score_nb]
algorithms = ["ID3TestAccuracy","Random Forest","Naives Bayes"]

for i in range(len(algorithms)):
    print("The accuracy score achieved using "+algorithms[i]+" is: "+str(scores[i])+" %")

The accuracy score achieved using ID3TestAccuracy is: 100.0 %
The accuracy score achieved using Random Forest is: 100.0 %
The accuracy score achieved using Naives Bayes is: 95.0 %

sns.set(rc={'figure.figsize':(15,8)})
plt.xlabel("Algorithms")
plt.ylabel("Accuracy score")
sns.barplot(algorithms,scores)
```



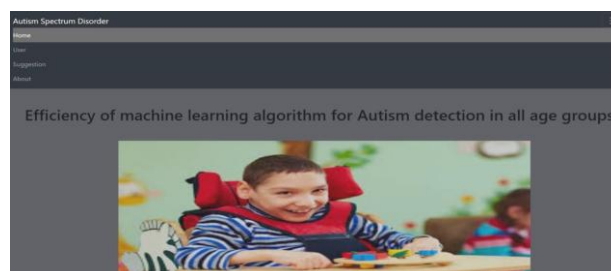
Fig[2]:Accuracy Analysis

The following images are a visual user interface, where patients can provide input with their points and other parameters such as medical assistant or directions.

C.Website for Diagnosis of ASD patients

The following images are a visual aid where participants can provide their points with other structures such as input with the help of a medical assistant or guidelines.

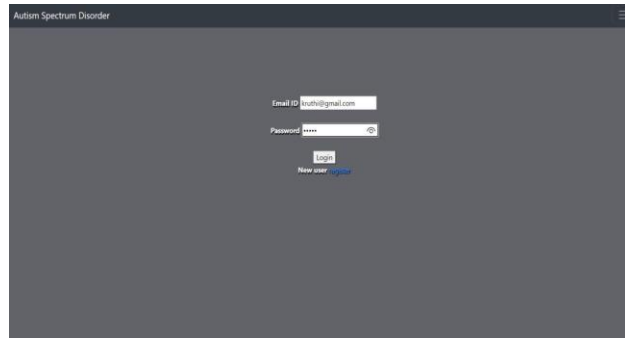
Home page



Fig[3]: Home Page

B. Log in page

The user needs to login by giving his registered email id and password. If the user is new, registration is first.



Fig[4]: Login Page

E. User information page

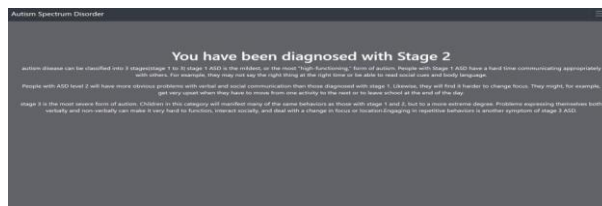
Participants should fill the form and click submit button.



Fig[5]: Information Page

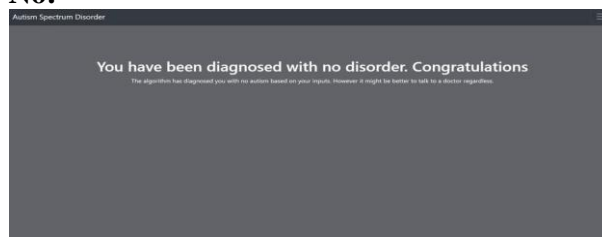
F. Result page.

Yes:



Fig[5]: ASD Diagonised page

No:



References

- [1] American Psychiatric Association . Retrieved from American Psychiatric Association 2012
- [2] Ben-David, S. S.-S. Understanding Machine Learning: From Theory to Algorithms. -: Cambridge University Press.2014
- [3] Becerra-Culqui, T. A., Lynch, F. L., Owen-Smith,A. A., Spitzer, J., & Croen, L. A. (2018). Parental first concerns and timing of Autism Spectrum Disorder diagnosis. *Journal of autism and developmental disorders*, 48(10), 3367-3376 2016.
- [4] Bishop-Fitzpatrick, L., Movaghar, A., Greenberg, J. S., Page, D., DaWalt, L. S., Brilliant, M. H., & Mailick, M. R. (2018). Using machine learning to identify patterns of lifetime health problems in decedents with autism spectrum disorder. *Autism Research*, 11(8), 1120-1128.
- [5] Bravo Oro A., N.-C. M. (2014). *Autistic Behavior Checklist (ABC) and Its Applications*. New York: Springer.
- [6] Carla A. Mazefsky, R. A. (2011, March -). PubMed Central. Use of artificial intelligence to shorten the behavioral diagnosis of autism,” Retrieved May 30, 2012
- [7] Duvekot, J., van der Ende, J., Verhulst, F. C., Slappendel, G., van Daalen, E., Maras, A., & Greaves-Lord, K. (2017). Factors influencing the probability of a diagnosis of autism spectrum disorder in girls versus boys. *Autism*, 21(6), 646-658. 2015
- [8] Gandhi, R. he short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls,” *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 51, 2018
- [9] Greenhalgh, T. How to read a paper. Papers that report diagnostic or screening tests. London: University College London Medical School 2017
- [10] Grove, J., Ripke, S., Als, T. D., Mattheisen, M., Walters, R. K., Won, H & Awashti, S. (2019). Identification of common genetic risk variants for autism spectrum disorder. *Nature genetics*, 51(3), 431-444.
- [11] Hershy, A. Gini Index vs Information Entropy. Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion,”2019
- [12] Hyde, K. K., Novack, M. N., LaHaye, N., Parlett-Pelleriti, C., Anden, R., Dixon, D. R., & Linstead, E. Applications of supervised machine learning in autism spectrum disorder research: a review. *Review Journal of Autism and Developmental Disorders*, 6(2), 128-146.2019
- [13] Ibrahim, S., Djemal, R., & Alsuwailem, A. Electroencephalography (EEG) signal processing for epilepsy and autism spectrum disorder diagnosis. *Biocybernetics and BiomedicalEngineering*, 38(1), 16-26. 2018
- [14] Jung, H. F. Meneguzzi, “Identification of autism spectrum disorder using deep learning and the abide dataset,” *NeuroImage: Clinical*, vol. 17, 2018
- [15] LeBarton, E. S., & Landa, R. J. Infant motor skill predicts later expressive language and autism spectrum disorder diagnosis. *Infant Behavior and Development*, 54, 37-47. 2019