PREDICTION OF AUTISTIC SPECTRUM DISORDER BASED ON BEHAVIOURAL FEATURES USING MACHINE LEARNING

Lakshmi B[1], Kala A[2]

[1] Student, Department of Information Technology, Sri Venkateswara College of Engineering, TamilNadu, India
[2]Asst. Professor, Department of Information Technology, Sri Venkateswara College of Engineering, TamilNadu, India

---***---

Abstract - *Autistic Spectrum Disorder is a neurodevelopmental disorder that affects a person's interaction, communication, learning skills and it is gaining momentum faster than ever. Detecting autism traits through screening tests is time-consuming and very expensive. With the advancement of machine learning and it's algorithms, autism can be predicted at an early stage. Although there are a lot of studies using different techniques, these studies did not provide any definitive conclusion about the prediction of autism in terms of different age groups. Therefore, this project aims at building a machine learning model that predicts the disorder using supervised machine learning algorithms.*

Key Words*: *ASD, Autism, Neural networks, machine learning, Adam

1. INTRODUCTION

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. As it is evident from the name, it gives the computer that which makes it more similar to humans: ***The ability to learn***. Machine learning is actively being used today, perhaps in many more places than one would expect. This project is about the application of machine learning in the field of health care, to identify ASD.

Autistic spectrum disorder is a neurodevelopmental disorder that affects a person's interaction, communication, learning skills and it is gaining its momentum faster than ever. Detecting autism traits through screening tests is time-consuming and very expensive. With the advancement of machine learning and its algorithms, autism can be predicted at an early stage.

Current explosion rate of autism around the world is numerous and it is increasing at a very high rate. According to the WHO, about 1 out of every 160 children has ASD. Some people with this disorder can live independently, while others require life-long care and support. Diagnosis of autism requires a significant amount of time and cost. Earlier detection of autism can come to a great help by prescribing patients with proper medication at an early stage. It can prevent the patient's condition from deteriorating further and would help to reduce long term costs associated with delayed diagnosis.

Thus, a time-efficient, accurate and easy screening test tool is very much required which would predict autism traits in an individual and identify whether or not they require comprehensive autism assessment. Therefore, this project aims at building a machine learning model that predicts the disorder using Neural networks.

The objective of this work is to propose an autism prediction model using Neural Networks that could effectively predict autism traits of an individual of any age. To be more precise, the focus of this work is on developing an autism screening application for predicting the ASD traits among people of age groups 4-11 years, 12-17 years and for people of age 18 and more.

The rest of the paper is organized as follows. Section II discusses the related research done in this area previously. Section III presents the research methodology. Section IV elaborates the detailed implementation of the proposed system and the implemented system is evaluated in Section V. Finally, Section VI concludes the paper by highlighting the research contributions, limitations and future plans to extend this work further.

2. LITERATURE SURVEY

This section briefly describes the works related to prediction techniques of ASD. For example, Kazi aims to propose an effective prediction model based on the ML technique and to develop a mobile application for predicting ASD for people of any age. The autism prediction model was developed by merging Random Forest-CART (Classification and Regression Trees) and Random Forest-Id3 (Iterative Dichotomiser 3). The proposed model was evaluated with AQ-10 dataset and 250 real datasets collected from people with and without autistic traits. The evaluation results showed that the proposed prediction model provides better results in terms of accuracy, specificity, sensitivity, precision and false positive rate (FPR) for both kinds of datasets.[2]

Kayleigh provides a comprehensive review of 45 papers utilizing supervised machine learning in ASD, including algorithms for classification and text analysis. In these 35 reviewed ASD research studies, the most commonly used supervised machine learning algorithms were SVM and ADtree. Supervised machine learning algorithms were used to identify candidate ASD genes, and to investigate obscure links between ASD and other domains.[3]

Milan used six personal characteristics age, sex, handedness, and three individual measures of IQ from 851 subjects in the Autism Brain Imaging Data Exchange (ABIDE) database to predict the model's performance. While [1] Daniel analyzed an eye movement dataset from a face recognition task, to classify children with and without ASD to obtain an accuracy of 88.51%. [4]

D.P.Wall is The Autism Diagnostic Interview-Revised (ADI-R) is one of the most commonly used instruments for assisting in the behavioural diagnosis of autism. The exam consists of 93 questions that must be answered by a care provider within a focused session that often spans 2.5 hours. Machine learning techniques were used to study the complete sets of answers to the ADI-R available at the Autism Genetic Research Exchange (AGRE) for 891 individuals diagnosed with autism and 75 individuals who did not meet the criteria for an autism diagnosis. The analysis showed that 7 of 93 items contained in the ADI-R were sufficient to classify autism with 99.9% statistical accuracy.[5]

Bram is about Predicting if a child has Autism Spectrum Disorder proved possible by using developmental delay, learning disabilities and speech or other language problems. Two methods were used to identify the severity of the ASD. The 1-away method improved the accuracy from 54.1% to 90.2%, which is a significant increase. This and the fact that the severity was based on input from just the caretakers of the children, prompts the need for further research in this matter.[6]

Wenbo identifies autism using Support Vector Machine (SVM) which provided accuracy up to 89% whereas [8] Jianbo used Natural Language Processing (NLP) for autism detection based on information extracted from medical forms of potential ASD patients. The proposed system achieves it an 83.4% accuracy and 91.1% recall, which is very promising.[7]

Chua combined a deep learning method with SVMRFE to improve the classification accuracy of ASD based on the whole ABIDE dataset. A total of 501 subjects with autism and 553 subjects with typical control across 17 sites were involved in the study. The state-of-the-art average accuracy of 93.59%.[9] Anibal used deep learning techniques to classify autism classes using clinical datasets. [10]

From the literature review, it is evident that, though many types of research has been carried out in this field, the researchers did not come to a decisive conclusion on using ML approach to predict autism for different age groups. Different tools and methods were employed for autism screening tests, but none concentrated on different age groups.

3. RESEARCH METHODOLOGY

The research was carried out in four phases: Data Set collection, Data synthesis, Developing the prediction model, Evaluating the predicted model. The phases are briefly discussed in the following subsections:

A. Data Set collection

To develop an effective predictive model, AQ-10 dataset was used which consists of three different datasets based AQ-10 screening tool questions. These three data sets contain data of age groups of 4-11 years (child), 12-17 years (adolescent) and 18 years plus (adults). AQ-10 or Autism spectrum Quotient tool is used to identify whether an individual should be referred for a comprehensive autism assessment. These questions mainly focus on domains like attention switching, communication, imagination, and social interaction. Since the actual collection of data from patients would be quite difficult, the data is collected from the UCI Machine Learning Repository as well, which is depicted in Fig 1.

```
@relation child

@attribute A1_Score {0,1}
@attribute A2_Score {0,1}
@attribute A3_Score {0,1}
@attribute A4_Score {0,1}
@attribute A5_Score {0,1}
@attribute A6_Score {0,1}
@attribute A7_Score {0,1}
@attribute A8_Score {0,1}
@attribute A9_Score {0,1}
@attribute A10_Score {0,1}
@attribute age numeric
@attribute gender {m,f}
@attribute ethnicity {Others,'Middle Eastern ',White-European,Black,'South Asian',Asian,Pasifika,Hispanic,Turkish,Latino}
@attribute jundice {no,yes}
@attribute austim {no,yes}
@attribute contry_of_res {Jordan,'United States',Egypt,'United Kingdom',Bahrain,Austria,Kuwait,'United Arab Emirates',Europe,Malta,Bulgaria,'South
Africa',India,Afghanistan,Georgia,'New Zealand',Syria,Iraq,Australia,'Saudi Arabia',Armenia,Turkey,Pakistan,Canada,Oman,Brazil,'South Korea','Costa
Rica',Sweden,Philippines,Malaysia,Argentina,Japan,Bangladesh,Qatar,Ireland,Romania,Netherlands,Lebanon,Germany,Latvia,Russia,Italy,China,Nigeria,'U.S. Outlying
Islands',Nepal,Mexico,'Isle of Man',Libya,Ghana,Bhutan}
@attribute used_app_before {no,yes}
@attribute result numeric
@attribute age_desc {'4-11 years'}
@attribute relation {Parent,Self,Relative,'Health care professional',self}
@attribute Class/ASD {NO,YES}

@data
1,1,0,0,1,1,0,1,0,0,6,m,Others,no,no,Jordan,no,5,'4-11 years',Parent,NO
1,1,0,0,1,1,0,1,0,0,6,m,'Middle Eastern ',no,no,Jordan,no,5,'4-11 years',Parent,NO
1,1,0,0,0,1,1,1,0,0,6,m,?,no,no,Jordan,yes,5,'4-11 years',?,NO
0,1,0,0,1,1,0,0,0,1,5,f,?,yes,no,Jordan,no,4,'4-11 years',?,NO
1,1,1,1,1,1,1,1,1,1,5,m,Others,yes,no,'United States',no,10,'4-11 years',Parent,YES
0,0,1,0,1,1,0,1,0,1,4,m,?,no,yes,Egypt,no,5,'4-11 years',?,NO
1,0,1,1,1,1,0,1,0,1,5,m,White-European,no,no,'United Kingdom',no,7,'4-11 years',Parent,YES
1,1,1,1,1,1,1,1,0,0,5,f,'Middle Eastern ',no,no,Bahrain,no,8,'4-11 years',Parent,YES
1,1,1,1,1,1,1,0,0,0,11,f,'Middle Eastern ',no,no,Bahrain,no,7,'4-11 years',Parent,YES
0,0,1,1,1,0,1,1,0,0,11,f,?,no,yes,Austria,no,5,'4-11 years',?,NO
1,0,0,0,1,1,1,1,1,1,10,m,White-European,yes,no,'United Kingdom',no,7,'4-11 years',Self,YES
0,1,0,0,1,0,0,0,0,1,5,f,?,no,no,Kuwait,no,3,'4-11 years',?,NO
0,1,1,1,1,1,1,1,1,1,4,m,White-European,yes,no,'United States',no,9,'4-11 years',Parent,YES
1,0,0,0,0,0,1,0,0,0,4,f,Black,no,no,'United Arab Emirates',no,2,'4-11 years',Parent,NO
1,1,1,1,1,1,1,1,1,1,6,m,White-European,no,no,Europe,no,10,'4-11 years',Parent,YES
1,1,1,1,1,1,1,1,1,1,8,m,White-European,no,no,Malta,no,10,'4-11 years',Parent,YES
1,1,1,1,1,1,0,1,1,1,4,m,'South Asian',no,no,Bulgaria,no,9,'4-11 years',Parent,YES
0,0,0,0,0,0,1,0,0,0,7,m,Others,no,no,'United States',no,1,'4-11 years',Parent,NO
1,0,1,1,1,0,1,1,1,1,11,m,White-European,no,yes,'United States',no,8,'4-11 years',Parent,YES
1,1,1,1,1,1,0,1,0,1,5,m,?,no,no,Egypt,no,8,'4-11 years',?,YES
1,1,1,1,1,1,1,0,1,0,5,m,White-European,yes,no,'South Africa',no,8,'4-11 years',Parent,YES
0,0,1,1,0,1,0,1,1,0,9,f,?,no,no,Egypt,no,5,'4-11 years',?,NO
1,1,0,1,0,0,0,0,0,0,4,m,Asian,no,no,India,no,3,'4-11 years',Parent,NO
1,0,1,1,0,1,0,0,1,0,6,f,'South Asian',no,no,India,no,5,'4-11 years',Parent,NO
```


Fig 1. Sample data set from the UCI Machine Learning Repository

B. Data synthesis

The collected data were synthesized to remove irrelevant features. For example, ID was irrelevant to develop a prediction model, hence it was removed. Further, unnecessary fields were deleted using pandas. This is done in order to increase the accuracy in classification. Summary of synthesized datasets is shown in Table 1.

Table 1. Summary of the chosen dataset

Age Group	Total Cleaned Instances	% of Male-Female	Average Age
4-11 years	248	70.16% male, 29.84% female	6.43 years
12-16 years	98	50% male, 50% female	14.13 years
18 and more	608	52.7% male, 47.3% female	29.63 years

C. Developing the prediction model

To generate a prediction of autism traits, algorithms had been developed and their accuracy was tested. After attaining results from various types of machine learning techniques such as Linear regression, SVM, Naive Bayes; Neural networks were found to be highly feasible with higher accuracy than the other algorithms. So, Neural Networks was proposed for implementing the ASD predictive system. Further modifications were made to the algorithm to get better results.

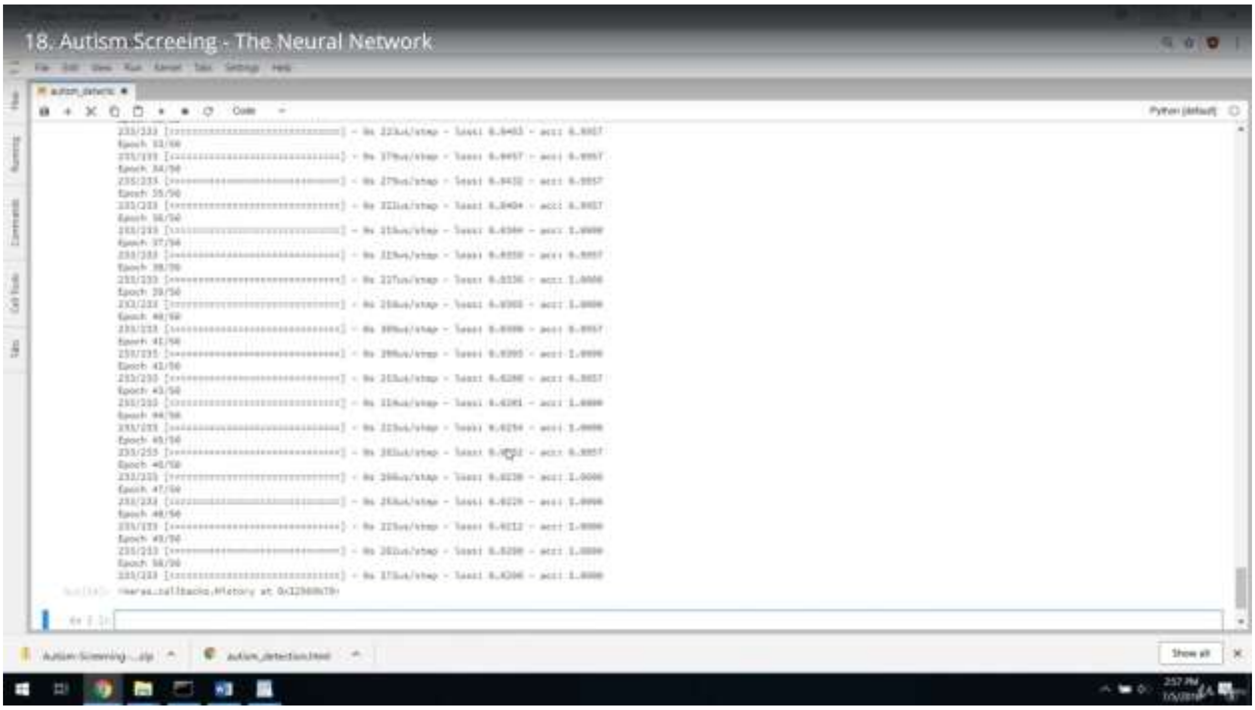

Fig 2. Development and Prediction of the model

D. Evaluating the predicted model

The model is tested with data that has been trained with the help of neural networks. This is used in fine-tuning the prediction. Of all the data taken from the UCI Machine Learning Repository, 80% of it is used for training the model and the remaining 20% of the data is used for testing. Testing helps us to fine-tune the model further to increase the accuracy in prediction. With this, a 90% accuracy was achieved.

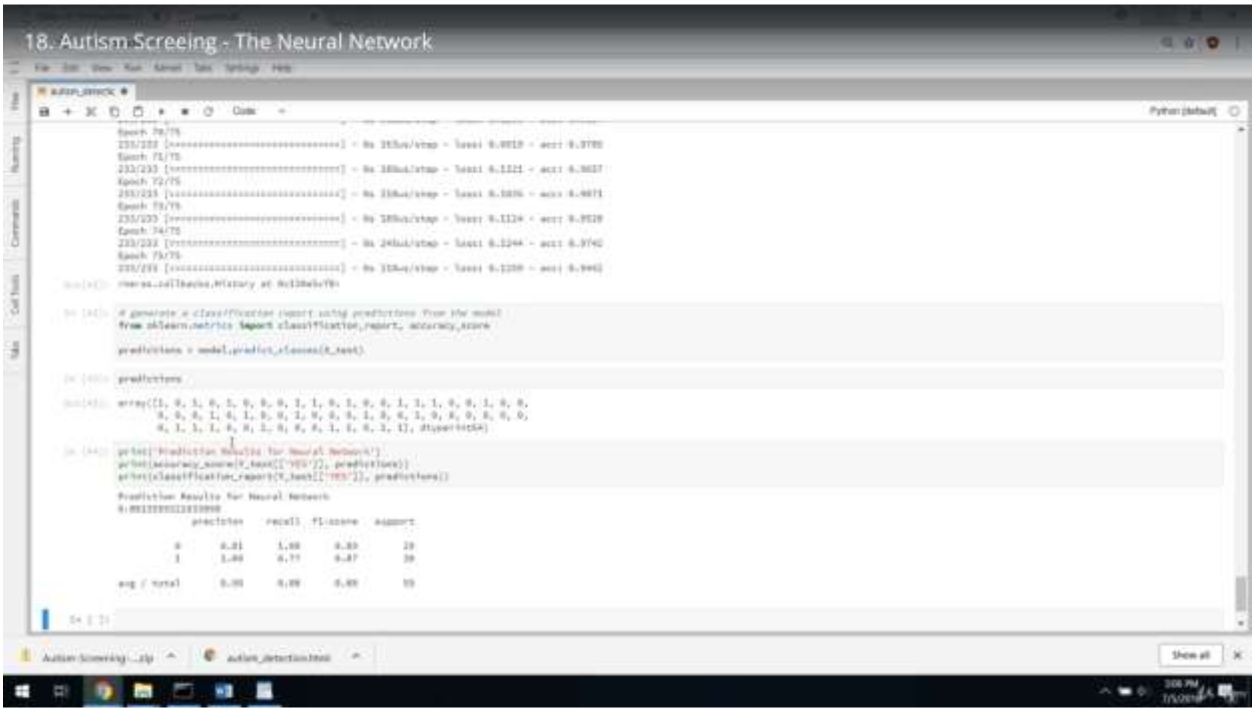

Fig 3. Evaluation and accuracy results

4. IMPLEMENTATION OF PROPOSED SYSTEM

The model is trained using the data set that has been pre-processed. This is used for proper prediction later. Hence, the data set is used to train multiple models. A neural network is a series of algorithms that endeavours to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature. Neural networks can adapt to changing input; so the network generates the best possible result without needing to redesign the output criteria.

The neuron of the proposed system is a combination of a linear and a nonlinear function that takes up vectors comprising the various attributes that are defined in the AQ-10 dataset. The importance of various attributes is defined in the weight function. This linear combination can be depicted as:

$$f(x1, x2) = w1x1 + w2x2 \text{ --- } (1)$$

Equation 1. A linear function of the neural network

The nonlinear part of the model also called the activation function, is represented using the ReLU function. A neural network without an activation function is essentially just a linear regression model. The activation function does the non-linear transformation to the input making it capable of learning and performing more complex tasks.

$$\mathrm{ReLU}(x) = \max(x,0) \text{ --- } (2)$$

Equation 2. The activation function of the neural network

Thus, the model can be represented as:

$$f(x1, x2) = \max(0, w1x1+w2x2) \text{ --- } (3)$$

Equation 3. Proposed neural network model

Adam algorithm is used for this neural network.

```
for t in range(num_iterations):
    g = compute_gradient(x, y)
    m = beta_1 * m + (1 - beta_1) * g
    v = beta_2 * v + (1 - beta_2) * np.power(g, 2)
    m_hat = m / (1 - np.power(beta_1, t))
    v_hat = v / (1 - np.power(beta_2, t))
    w = w - step_size * m_hat / (np.sqrt(v_hat) + epsilon)
```

Fig 4. Basic Adam algorithm functionality

5. EVALUATION OF PROPOSED SYSTEM

The prediction is the actual accurate identification of the autism data based on the input given. Owing to the data given the model is trained in a better way. More data, more fine-tuning. Hence, bigger datasets give more accuracy. The advantage of the usage of neural networks for prediction is that they are able to learn from examples only and that after their learning is finished, they are able to catch hidden and strongly non-linear dependencies, even when there is significant noise in the training set. The disadvantage is that neural networks can learn the dependency valid in a certain period only. The error of prediction cannot be generally estimated. However, the accuracy was close to 90%. This was implemented using Python with Keras data processing package. With multiple epochs and batch processing, the accuracy was found to be close to 90%. This can still be fine-tuned, which is covered in future work.

```
In [21]: from sklearn.metrics import classification_report, accuracy_score
         predictions = model.predict_classes(X_test)

In [22]: predictions

Out[22]: array([1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1,
                0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0,
                0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1], dtype=int64)

In [23]: print(accuracy_score(Y_test[["YES"]], predictions))

         0.8983050847457628

In [24]: print(classification_report(Y_test[["YES"]], predictions))

                       precision    recall  f1-score   support

                    0       0.90      0.90      0.90        29
                    1       0.90      0.90      0.90        30

             accuracy                           0.90        59
            macro avg       0.90      0.90      0.90        59
         weighted avg       0.90      0.90      0.90        59
```

Fig 5. Accuracy of the proposed model

6. CONCLUSION AND FUTURE WORK

Autism is quite common, and with the results, one can find out which is the major contributing factor towards autism. Since the data set is quite comprehensive in terms of the factors, one can easily scrutinize such pregnant mothers and take care in the initial stages. Also, this will help the health care providers to split the funding and care accordingly.

The primary limitation of the study is the lack of sufficiently large data to train the model. Another limitation is that the screening application is not designed for the age group below 3 years as open-source data was not available. Our future work will focus on collecting more data from various sources to improve the accuracy of the proposed system to take it to a higher level.

REFERENCES

[1] Kazi Shahrukh Omar, Prodipta Mondal, Nabila Shahnaz Khan, Md. Rezaul Karim Rizvi, 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

[2] Kayleigh K. Hyde, Marlena N. Novack, Nicholas LaHaye, Chelsea Parlett-Pelleriti, Raymond Anden, Dennis R. Dixon, Erik Linstead, “Applications of Supervised Machine Learning in Autism Spectrum Disorder Research: a Review”, Review Journal of Autism and Developmental Disorders (2019) 6:128–146, https://doi.org/10.1007/s40489-019-00158-x

[3] https://ctb.ku.edu/en/table-of-contents/evaluate/evaluate-community-interventions/information-gathering-synthesis/main

[4] D. P. Wall, R. Dally, R. Luyster, J.-Y. Jung, and T. F. DeLuca, “Use of artificial intelligence to shorten the behavioral diagnosis of autism,” PloS one, vol. 7, no. 8, p. e43855, 2012.

[5] Bram van den Bekerom, “Using Machine Learning for Detection of Autism Spectrum Disorder”, 2017

[6] Wenbo Liu, Ming Li, and Li Yi, “Identifying Children with Autism Spectrum Disorder Based on Their Face Processing Abnormality: A Machine Learning Framework”, Autism Research 00: 00–00, 2016

[7] Jianbo Yuan, Chester Holtz, Tristram Smith, Jiebo Luo, “Autism spectrum disorder detection from semi-structured and unstructured medical data”, EURASIP Journal on Bioinformatics and Systems Biology (2017) 2017:3 DOI 10.1186/s13637-017-0057-1

[8] Daniel Bone, Matthew S. Goodwin, Matthew P. Black, Chi-Chun Lee, Kartik Audhkhasi, Shrikanth Narayanan, "Applying Machine Learning to Facilitate Autism Diagnostics: Pitfalls and Promises", Journal of Autism and Developmental Disorders, May 2015, Volume 45, Issue 5, pp 1121–1136

[9] Chua Wang, Zhiyong Xiao, Baoyu Wang, Jianhua Wu, "Identification of Autism Based on SVM-RFE and Stacked Sparse Auto-Encoder", IEEE Access ISSN-2169-3536, 21 August, 2019

[10] Jared A. Nielsen, Brandon A. Zielinski, P. Thomas Fletcher, Andrew L. Alexander, Nicholas Lange, Erin D. Bigler, Janet E. Lainhart and Jeffrey S. Anderson, "Multisite functional connectivity MRI classification of autism: ABIDE", Front. Hum. Neurosci., 7 (September)(2013), pp. 1-12

[11] Plitt M., Barnes K.A., Martin A. "Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards" NeuroImage: Clinical, 7 (2015), pp. 359-366

[12] Anibal Sólon Alexandre Rosa Franco, R. Cameron Craddock, Augusto Buchweitz, Felipe Meneguzz, "Identification of autism spectrum disorder using deep learning and the ABIDE dataset", NeuroImage: Clinical Volume 17, 2018, Pages 16-23, https://doi.org/10.1016/j.nicl.2017.08.017

[13] Milan N. Parikh, Hailong Li1 and Lili He, "Enhancing Diagnosis of Autism With Optimized Machine Learning Models and Personal Characteristic Data", Front. Comput. Neurosci.,15 February 2019

[14] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[15] Yuji Roh, Geon Heo, Steven Euijong Whang, "A Survey on Data Collection for Machine Learning A Big Data - AI Integration Perspective", IEEE, August 2019

[16] Ibrahim M. Nasser, Mohammed O. Al-Shawwa, Samy S. Abu-Naser, "Artificial Neural Network for Diagnose Autism Spectrum Disorder", International Journal of Academic Information Systems Research (IJAISR), February 2019