Autistic Spectrum Disorder Screening: Prediction with Machine Learning Models

Astha Baranwal

School of Information and Technology Engineering
Vellore Institute of Technology
VIT-Vellore, Tamil Nadu, India

asthabaranwalabc@gmail.com

***Abstract* - Autistic Spectrum Disorder (ASD) is a developmental disorder that can be observed in all age groups. This paper uses ASD screening dataset for analysis and prediction of probable cases in adults, children and adolescents. The dataset for each of the age groups are analyzed and inferences are drawn from them. Machine learning algorithms like Artificial Neural Networks (ANN), Random Forest, Logistic Regression, Decision Tree and Support Vector Machines (SVM) are used for prediction.**

Index Terms – Autism Spectrum Disorder, Machine Learning, Classification, Medical, Diagnosis

I. INTRODUCTION

Autism Spectrum Disorders bring about certain challenges in social, behavioral, communication and emotional understanding in an individual. People diagnosed with ASD have a range of symptoms. That is why it is termed as a 'spectrum' disorder. Since ASD is a neurological developmental disorder, there is no specific medical test for it, thus making the diagnosis of ASD an arduous task. Although now these disorders can be observed in early childhood, there are some cases in which the symptoms are not diagnosed until adolescent or adulthood. ASD currently has no standard treatment. An early diagnosis and a head start in therapies can potentially lead to better results.

This paper focuses on proposing a model which would assist in prediction of ASD in an individual so that diagnosis can be done and further treatments may be followed. The dataset used is the Autistic Spectrum Disorder Screening Data. The datasets provide insights into various factors affecting the prediction of the disorder. Machine learning algorithms like decision tree, random forest, logistic regression, support vector classifier and artificial neural networks are used for finding out the optimal model for each dataset. Several performance metrics are used in order to analyze and compare each model from every angle possible.

II. RELATED WORK

Various methods for determining ASD have already been used. From employing image processing techniques to gather abnormalities in brain structure which may indicate ASD, to observing genetic makeup of individuals, the previous researches on the topic have paved way for better diagnosis of ASD. Anibal et al [1] used a large brain imaging dataset to identify ASD patients with deep learning algorithms. They showed the anterior-posterior underconnectivity autistic brains. Hassan et al [2] utilized the decision tree algorithm for analysis of National Database for Autism Research (NDAR) dataset. Thabtah et al [3] aimed at extracting the most influencial features which contribute in ASD prediction. For this purpose, they used Variable Analysis which extracted features from child, adolescent and adult dataset. Choudhery et al [4] utilized gene expression dataset. K-means clustering was used to cluster genes and then a support vector machine model classified functional connectivity changes associated with ASD. Pagnozzi et al [5] investigated brain changes linked to ASD patients through MRI image data. They identified certain biomarkers by studying brain morphology. This paper aims to create an optimal model for autism spectrum disorder prediction based on the autism screening datasets for three age groups, viz., child, adolescent and adult, contributed by Fadi Fayez Thabtah [3].

III. METHODOLOGY

A. The Dataset

The datasets used are the Autism Screening Datasets for adult, adolescent and child age groups. The data has been collected through a mobile application called ASDTests, which aims at self-assessment for a potential patient. There are 20 attributes in each dataset having continuous, categorical and binary values. The dependent attribute is Class_ASD which determines if an individual has ASD (1) or not (0). The adult dataset has 704 records, the adolescent dataset has 104 records and the child dataset has 292 records. Clearly, the adult dataset is more suited for building machine learning models.

TABLE I
DATASET DESCRIPTION

Serial no.	Attribute	Description	Data type
1-10	A1_Score to A9_Score	Answer code of the corresponding question.	Binary
11	age	Age of the individual	Integer
12	gender	Gender of the individual	String (f, m)
13	ethnicity	Ethnic group the individual belongs to	String
14	jaundice	If the person had jaundice at birth	String (no, yes)
15	autism	If any relative of the individual was diagnosed with autism	String (no, yes)
16	country_of_residence	Native country	String
17	used_app_before	If the screening test app has been used by the person before	String (no, yes)

18	score	Score out of 10 based on the screening test answers	Integer
19	relation	Who is answering the questions of screening test	String
20	Class_asd	ASD diagnosis of individual by the screening app	String (NO, YES)

B. Data Cleaning

The attributes like gender, jaundice, autism, used_app_before and Class_ASD are converted into binary 0/1 values to ease the implementation of classification algorithms.

The missing data in attributes like ethnicity, relation and age are dealt with by removing these records altogether from the datasets.

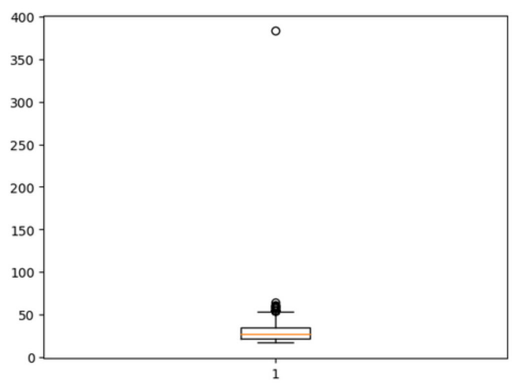

Fig. 1 Magnetization as a function of applied field. Note how the caption is centered in the column.

In the adult dataset, an outlier in age is discovered having value 383. This is not a feasible age and must be some typo. So, this is changed to 38.

C. Data Analysis

A score of more than 7 in Q & A for screening test automatically results in a positive ASD classification.

There is no significant relationship between a person who was born with jaundice and being diagnosed with ASD. There is no significant relationship between a person having a relative diagnosed with ASD and probability of the person lying in the autism spectrum himself. These facts prevail for all the three datasets.

1) Adult Autism Screening Dataset: The adult autism dataset has records with wide spanning, covering the young adult phase to the senile phase for both the genders. It is observed that ASD is more widely distributed in males in terms of age.

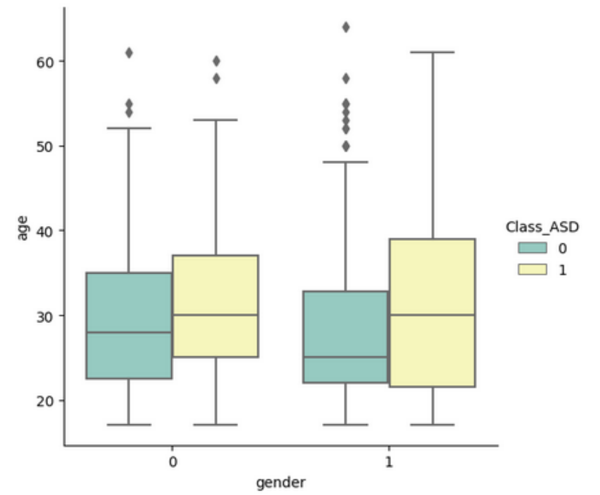

Fig. 2 Box plot of age against gender in adult autism dataset.

A score of more than 7 in Q & A for screening test automatically results in a positive ASD classification.

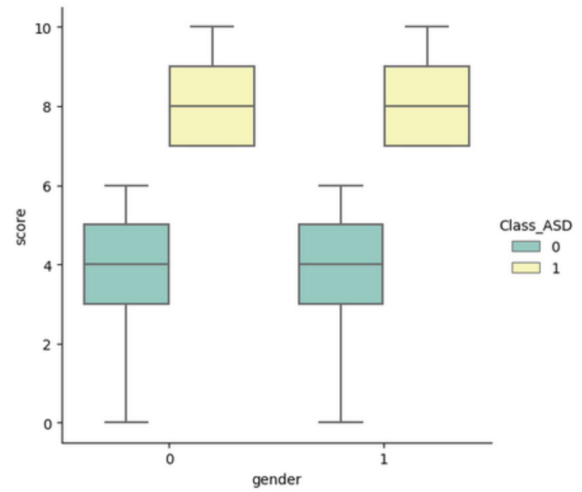

Fig. 3 Box plot of score against gender in adult autism dataset.

The White European ethnicity accounts for nearly one-third of the data. The ethnic groups who score more in screening test are White Europeans, Asians and Blacks.

2) Adolescent Autism Screening Dataset: The individuals ranging from ages 12 to 16 lie in this age group. The number of instances classified to have ASD is twice more.

3) Child Autism Screening Dataset: The children in the dataset are 4 to 11 years of age. The median is 6. The number of instances classified to have ASD is slightly more. A score of more than 7 in Q & A for screening test automatically results in a positive ASD classification. Black ethnic group seems to have higher scores in screening test.

D. Data Preprocessing

An autism screening score of more than 7 automatically classifies the patient to be lying in autism spectrum. So, the score variable is redundant and can be removed as the correlation between Class_ASD and score is 0.83, which is a really high value. It can lead to multicollinearity.

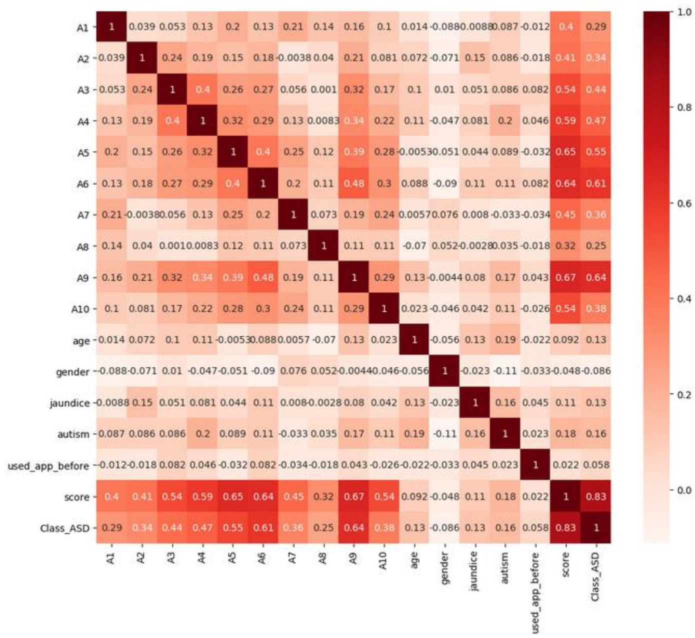

Fig. 4 Correlation matrix of the adult autism screening dataset.

The ethnicity attribute has 10 categories. Similarly, relation attribute has 5 categories. In order to perform classification, these two attributes can be turned to integer representation of categories or one-hot encoding can be used. In one-hot encoding each category of the categorical attributes are converted to binary values to facilitate machine learning algorithms on the dataset.

E. Feature Importance

Lasso regression feature importance model is used for feature importance determination. After shrinkage, the 'especially influential' factors are the ones with highest coefficients. Although, if the coefficient is really high, there is a chance of multicollinearity. The variables with coefficient zero are eliminated. It is clear that for all age group, the Q and A of screening test is the major deciding factor.

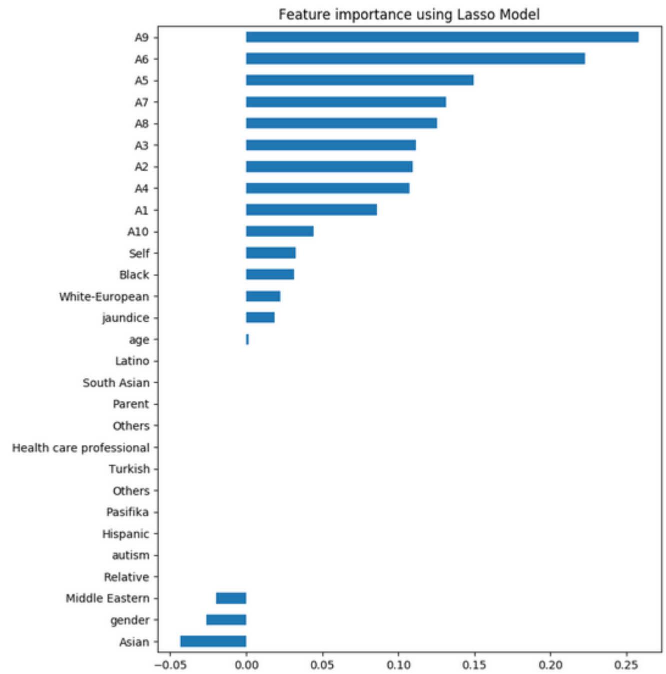

Fig. 5 Lasso model on adult autism screening dataset.

As per the model on the adult ASD screening dataset, the answer A9, corresponding to the question, 'I find it easy to work out what someone is thinking or feeling just by looking at their face', is the most important attribute in ASD diagnosis. The Lasso model picks up 18 variables and eliminates 11 variables.
The Lasso model picks up 14 variables while eliminating the other 16 variables in adolescent dataset. The question to the answer A5: 'S/he frequently finds that s/he doesn't know how to keep a conversation going', is the most important feature.

For the child dataset, the response A4, 'S/he finds it easy to go back and forth between different activities' is of utmost importance.

IV. PERFORMANCE METRICS

1) True positive (TP): The number of records that were actually positive and were classified positive

2) False Negative (FN): The number of records that were positive but were classified negative

3) True Negative (TN): The number of records that were actually negative and were classified negative

4) False Positive (FP): The number of records that were positive but were classified negative

5) F1 Score: F1 Score is a better measure than accuracy because of the uneven distribution of classes among the instances. The records with Class_ASD as 0 cover more than half of the instances.

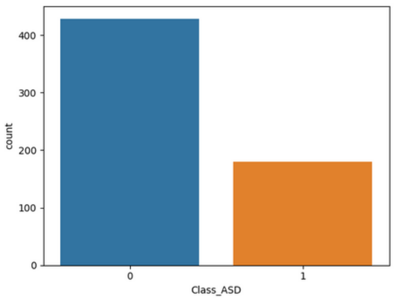

Fig. 6 Class distribution in adult autism screening dataset.

6) Accuracy Score: Accuracy score is the fraction of predictions the classification model predicted correctly.

$$Accuracy\ Score = \frac{TP + TN}{TP + FN + TN + FP}$$

7) ROC (Receiver Operating Characteristic) curve and AUC (Area Under the Curve): ROC is a probability curve plotted with the True Positive Rate (y-axis) against the False Positive Rate (x-axis). AUC is the measure of how much the model is capable of distinguishing between the classes. Higher the AUC value, better is the model at correctly predicting whether the individual has ASD. When AUC is 0.5, it is the worst case. Because then the model could not distinguish between the positive and negative classes reliably. At AUC zero, the model is just reciprocating the classes.

8) Sensitivity/ Recall: Sensitivity is the fraction of individuals who will be correctly predicted as positive class. High sensitivity means more number of individuals have been correctly predicted to have ASD.

$$Sensitivity = \frac{TP}{TP + FN}$$

9) Specificity: Sensitivity is the fraction of individuals who will be correctly predicted as negative class. High specificity means more number of individuals have been correctly predicted not to have ASD.

$$Specificity = \frac{TN}{TN + FP}$$

V. RESULTS AND EVALUATION

A. Model Building on Adult autism screening dataset

Five classification models are used: Decision Tree, Random Forest, Logistic Regression, Support Vector Classifier, Artificial Neural Network (ANN). The decision tree model is an overfitted model, with train dataset accuracy of 1 and test dataset accuracy of 0.8798. It is also the least optimal model as per the ROC curve. The AUC is the least for this model, indicating the separability between the two classification classes is poor. The ROC curve is also the least optimal. The F1 score is the least at 78.85%.

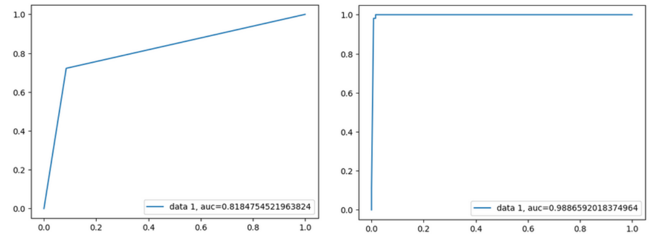

Fig. 7 ROC graph for Decision Tree (left) and ANN (right) for adult autism screening dataset.

Random Forest, Logistic Regression and Artificial Neural Network give near optimal ROC curve with a high AUC. The F1 Score is the highest for the ANN model at 98.15%.

Algorithm	F1 Score	Accuracy	AUC	Sensitivity	Specificity
Decision Tree	0.7885	0.8798	0.8447	0.7593	0.9302
Random Forest	0.8544	0.9180	0.9812	0.8148	0.9690
Random Forest (hyperparameter)	0.9524	0.9727	0.9977	0.9259	0.9922
Logistic Regression	0.9725	0.9836	0.9959	0.9814	0.9845
Support Vector	0.8627	0.9235	0.9886	0.8148	0.9690
Artificial Neural Network	0.9815	0.9891	0.9887	0.9815	0.9922

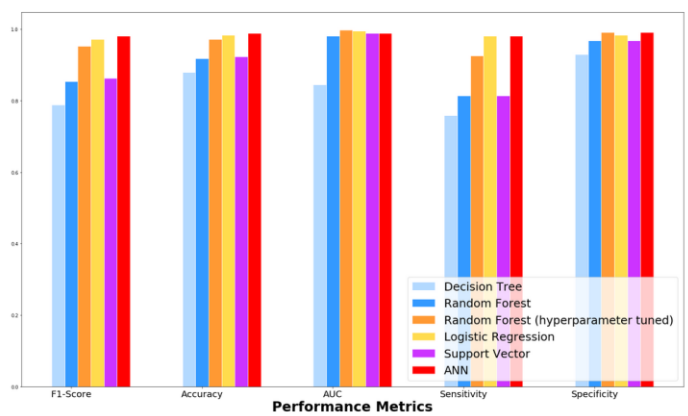

Fig. 8 Comparison graph for adult autism screening dataset between all the algorithms in terms of F1 Score, Accuracy, AUC, Sensitivity and Specificity.

ANN, Logistic Regression and Random forest (hyperparameter tuned) give exceptional F1 scores. While Decision tree gives the worst.

All in all, ANN performs the best and is the optimal model.

B. Model Building on Adolescent autism screening dataset

Since this is a small dataset, building machine learning models is not advisable. Overfitting is observed in the decision tree and the random forest models. The logistic regression model performs the best with the best ROC curve and the highest AUC.

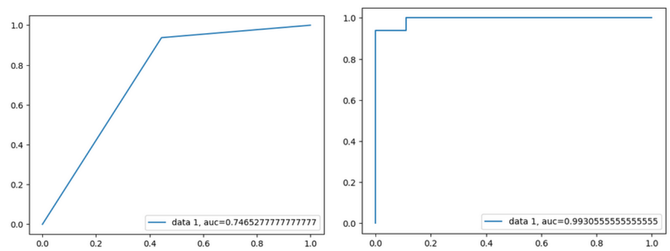

Fig. 9 ROC graph for Decision Tree (left) and Logistic Regression (right) for adolescent autism screening dataset.

Algorithm	F1 Score	Accuracy	AUC	Sensitivity	Specificity
Decision Tree	0.8571	0.8000	0.7465	0.9375	0.4444
Random Forest	0.9000	0.8800	0.9444	0.8750	0.8889
Logistic Regression	0.9412	0.9200	0.9931	1.0000	0.7778
Support Vector	0.8649	0.8000	0.9861	1.0000	0.4444
Artificial Neural Network	0.8125	0.7600	0. 9861	0.9815	0.8125

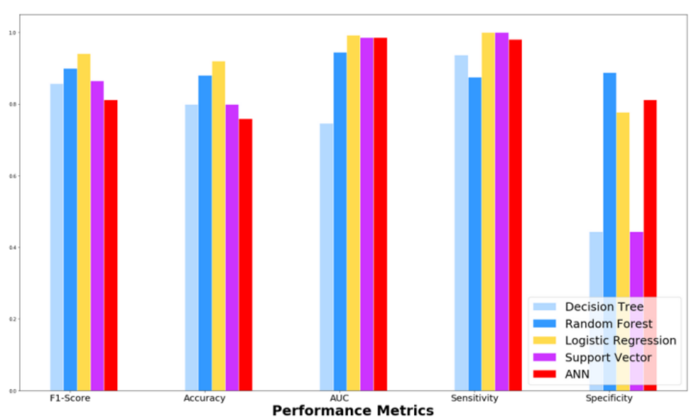

Fig. 10 Comparison graph for adolescent autism screening dataset between all the algorithms in terms of F1 Score, Accuracy, AUC, Sensitivity and Specificity.

Random forest model gives high specificity but fails to achieve the optimal results in all other performance measures.

Decision tree and random forest models do not perform well. They are overfitted models too.

Logistic Regression is the best model when all the metrics are taken into account.

C. Model Building on Child autism screening dataset

Both the decision tree and the random forest models are overfitted and are hence, poor models. The random forest model after hyperparameter tuning is a fair model. SVM is the most optimal model, followed by ANN.

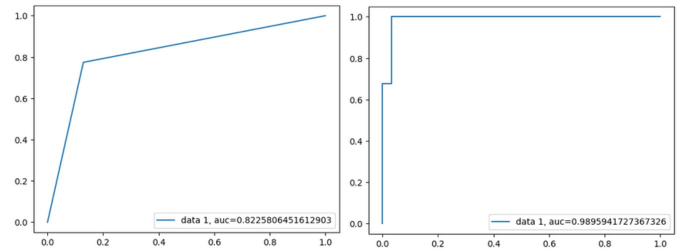

Fig. 11 ROC graph for Decision Tree (left) and Support Vector (right) for child autism screening dataset.

Algorithm	F1 Score	Accuracy	AUC	Sensitivity	Specificity
Decision Tree	0.8136	0.8226	0.8225	0.7742	0.8710
Random Forest	0.8197	0.8226	0.9433	0.8065	0.8387
Random Forest (hyperparameter)	0.9062	0.9032	0.9977	0.9355	0.8701
Logistic Regression	0.9062	0.9032	0.9865	0.9354	0.8710
Support Vector	0.9688	0.9677	0.9896	1.0000	0.9355
Artificial Neural Network	0.9508	0.9516	0.9896	0.9677	0.9355

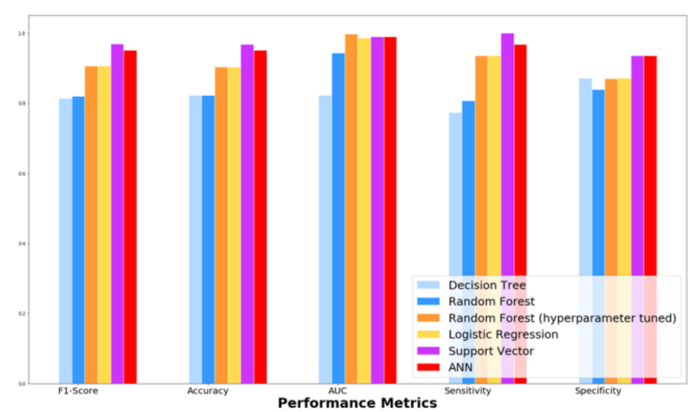

Fig. 12 Comparison graph for child autism screening dataset between all the algorithms in terms of F1 Score, Accuracy, AUC, Sensitivity and Specificity.

Decision tree and random forest models perform poorly when compared to other models in terms of all the performance metrics. Support Vector Classifier performs the best.

VI. DISCUSSIONS

Image processing on MRIs, evaluation of gene expression, the risk factors involving ASD and study on various biomarkers which may relate to the disorder have been studied previously. The ABIDE (Autism Brain Imaging Data Exchange) dataset used by Heinsfield[1] aims to identify the areas of brain which may define if an individual has ASD. Hassan et al[2] used decision tree to identify the genetic and environmental risk factors that may contribute to ASD. Similarly, Choudhery et al[4] used gene expression data for ASD patients to observe gene expression changes which may predict changes in brain regions. This paper focuses on the data collected from a self-screening application for ASD, created by Thabtah et al[3]. The test can easily be taken by the individuals if there is any probability of them being diagnosed with ASD. Machine learning models are proposed and compared which give a really good accuracy. The datasets, however, especially the child and adolescent datasets, are really small in size and are not suitable for building machine learning models. The adult autism screening dataset itself is biased towards the class 0 for Class_ASD. More data is required for reliable ASD prediction.

VII. CONCLUSION

An individual with Autism Spectrum Disorder needs early treatment and a progressive learning curve. The sooner ASD is diagnosed, the better are the results in long term. Often times ASD does not even get diagnosed until adulthood. This paper, based on the three autism screening datasets contributed by Thabtah, adopted various machine learning algorithms to find the optimal models for each of the datasets. Data analysis is done to figure out the relation between the attributes.

Decision Tree results in an overfitted model in all the datasets. ANN performs the best on the adult autism dataset. Logistic Regression gives the optimal result for adolescent autism dataset. Support Vector is the best for child autism screening dataset.

Although, the data we have so far is insufficient. The datasets are really small to derive a suitable model for prediction. Certain conclusions can still be derived through data analysis of the datasets. A proper diagnosis method for the disorder is crucial.

REFERENCES

[1] Anibal Sólon Heinsfeld, Alexandre Rosa Franco, R. Cameron Craddock, Augusto Buchweitz and Felipe Meneguzzi, "Identification of autism spectrum disorder using deep learning and the ABIDE dataset," *NeuroImage: Clinical*, vol. 17, pp. 16-23, 2018.

[2] Mariam M. Hassan and Hoda M. O. Mokhtar, "Investigating autism etiology and heterogeneity by decision tree algorithm," *Informatics in Medicine Unlocked*, vol. 16, 100215, 2019.

[3] Fadi Thabtah, Firuz Kamalov and Khairan Rajab, "A new computational intelligence approach to detect autistic features for autism screening," *International Journal of Medical Informatics*, vol. 117, pp. 112-124, sep 2018.

[4] Sanjeevani Choudhery, Chuan Huang and Daifeng Wang, "T253. A Machine Learning Approach to Predict the Changes of Brain Functional Connectivity in Autism Spectrum Disorder From the Gene Expression Data," *Biological Psychiatry*, vol. 83, issue 9, supplement, pp. S227-S228, 1 May 2018.

[5] Alex M. Pagnozzi, Eugenia Conti, Sara Calderoni, Jurgen Fripp and Stephen E. Rose, "A systematic review of structural MRI biomarkers in autism spectrum disorder: A machine learning perspective," *International Journal of Developmental Neuroscience*, vol. 71, pp. 68-82, dec 2018.