A Comparative Study on Predicting Autism Spectrum Disorders (ASD) Using Gene Expression and Machine Learning

Hala Alshamlan
Information Technology Department
King Saud University, Saudi Arabia
halshamlan@ksu.edu.sa

Hissah AL-Nojaidi
Information Technology Department
King Saud University, Saudi Arabia
Hessah.n@hotmail.com

Maraheb AlSuliman
Information Technology Department
King Saud University, Saudi Arabia
Marahebalsuliman@gmail.com

Reham Alabduljabbar
Information Technology Department
King Saud University, Saudi Arabia
ralabduljabbar@ksu.edu.sa

Abstract
Objective: The aim of this study is to identify gene expressions that could have a high potential to be used in predicting autism spectrum disorder (ASD) by using the filter gene selection method with Bayes Networks classifier compared with the wrapper selection method with the SVM algorithm. Bayes Networks classifier was chosen after testing multiple classifiers on the dataset and it has the highest accuracy.
Method: The experimental data used in the analysis comprised an autism microarray dataset from the well-known public repository GEO (NCBI) [1]. We have downloaded it after its normalized as part of previous work [4]. The dataset consists of 146 observations (samples) and 9454 genes (features). We applied the Correlation-based (CFS) Attribute Selection filter after that we tested the different classifiers with 10-fold cross-validation to show the highest accuracy of them to be chosen in the proposed model.
Result The best accuracy founded when apply Bayes Networks classifier with 10-fold cross-validation on dataset filtered by the CFS feature selection method, which results in 91.7% accuracy. The proposed model has better accuracy when comparing to the filter gene selection methods proposed in the previous work.

Keywords— *Autism Spectrum Disorder (ASD); Filter gene selection; Bayes Networks; CFS; Machine Learning; Weka.*

I. INTRODUCTION

According to WHO [2], about 1 out of every 160 children has ASD. The environmental and genetic factors are probably the main causes of ASD according to the available scientific evidence from World Health Organization. With the help of the promising technology and the wide achievements of machine learning in different fields, there are many studies carried out of gene diagnostic prediction using machine learning. However, the major problem in the gene expression analysis of ASD is the difficulty in selection and identification of the genes that are most relevant to autism. This problem exists because the gene expression levels in autism disorder show considerable fluctuation among individuals and because the sequences of several of these genes are highly variable [3].

Furthermore, the large variance in the distribution of gene expression levels is associated with many types of symptomatic profiles of autism represented in the base. Therefore, the application of standard methods, which serve very well in recognition of other cases, for example different types of cancer, does not lead to the acceptable results for autism. These issues motivate us to propose a useful and effective ASD classification method.

However, one of the most important tasks in conducting gene expression analysis using machine-learning algorithms is the building of a classification model that recognizes the discriminative genes with the highest possible accuracy. However, not every classifier works effectively on all datasets. For each dataset, a unique classifier or a limited number of classifiers typically work best [4].

The main motivation of the study is to apply a transcriptomic approach to identify a gene expression signature with promising performance in the diagnostic prediction. We plan to use CFS feature selection method which are effective for gene selection when applied in microarray gene expression profiling, along with Bayes Networks classifier. We aim to develop the proposed algorithm using Weka tool for gene analysis and will conduct various combination permutation between filters and classifiers with different cross validation folds to demonstrate the best accuracy. Moreover, a comparation

Manuscript received November 5, 2020
Manuscript revised November 20, 2020
https://doi.org/10.22937/IJCSNS.2020.20.11.9

with previous work will be conducted and results will be discussed.

In this study, we will utilize 146 observations (samples) and 9454 genes (features) collected from the well-known public repository GEO (NCBI) [1] and normalized in previous work [4]. Furthermore, we will use Weka tool to analyze the gene selection and classification.

The study is organized as follows: section II presents brief description of the research background. After that, section III shows the state of art, and then the dataset description and preparation process described in section IV. The proposed model presented in section V. After that, the implementation and a detailed analysis of the results demonstrated in section VI. A research discussion is conducted in section II. Finally, a future work and conclusion of the paper is presented.

II. BACKGROUND

1.1 ASD

Autism spectrum disorders (ASDs) are devastating neurodevelopmental disorders characterized by deficits in social communication and interaction across multiple contexts as well as restricted, repetitive patterns of interests and behavior. The Centers for Disease Control recently presented that the prevalence of ASD has risen to approximately 1 in 68, and most children are not diagnosed with ASD until after 4 years of age in the United States[2].

1.2 Machine Learning

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention [5]. We describe below briefly five Machine learning classifiers: SMO, K-NN, Random Forest, J48 and BayesNet.

1.2.1 SMO

Sequential Minimal Optimization (SMO) uses heuristics to separate the training of problems into smaller problems which can be analytically solved. This largely depends on the assumptions behind the heuristics (working set selection) whether or not it works well. It usually speeds up quite a bit of training.

1.2.2 Random Forest

The random forest is an algorithm for classification consisti ng of many trees of decisions. While constructing that indivi dual tree, it uses bagging and features randomness to try to c reate an uncorrelated tree forest whose prediction by commi ttee is more accurate than any individual tree.

1.2.3 KNN

k Nearest Neighbor (KNN) is a distance-based classifier and it has been extensively studied and discussed with respect to classification. Generally, in this algorithm "distance" is used to classify a new sample based on the labels of its neighbors selected from the training set.

1.2.4 J48

J48 is an algorithm for the top-down classification of a decision tree. The algorithm takes into account all possible tests that can divide the set of data and selects a test that gives the best IG. A check with results as many as the number of distinct values of the attribute is considered for each discrete attribute. Binary tests involving each distinct value of the attributes are considered for each attribute that continues. To gather competently the entropy gain from all these binary tests, the training data set of the node in considerations sorted by the values of the continuous attribute and the entropy gains of the binary cut based on each separate value shall be determined in one scan of the sorted data.

1.2.5 BayesNet

A Bayesian Network (BN) Represents a JPD over a set of ra ndom variables V. Bayesian Network defines random variab les and conditional dependencies by using a directed graphi cal model. For example, for a person with a particular diseas e, we can use a BN.

1.3 Gene Expression

The process by which the information encoded in a gene is converted into an observable phenotype (most commonly production of a protein). With the increasing development of machine learning algorithms, we have noticed their use in many researches that help to predict and detect autism.

1.4 Gene Selection Method

Gene selection is a process of selecting the most and small subset of informative genes for genetic disease that are most predictive to its related class.
The gene selection methods can be classified into three categories: Filter Gene Selection Methods, Wrapper Gene Selection Methods and Hybrid Gene Selection Methods.

1.4.1 Filter Gene Selection Methods

The selection of features (gene) is a frequently used preprocessing technology in microarray gene expression data analysis for successful autism classification tasks. Widespread approaches to gene selection focus mainly on filter methods. For high-dimensional data, filter methods are generally considered to be very effective and efficient.
Correlation-based feature selection (CFS) is a heuristic algorithm that tests the correlation between attributes and rewards those subsets of features in which each feature is highly correlated with class and uncorrelated with other subset features.

1.5 Weka Software

Weka is a Java-based open-source platform with many machine learning algorithms. Weka is a series of algorithms for machine learning for data mining tasks. We can either apply the algorithms directly to a dataset or call them from our own Java code. Weka provides tools for pre-processing, classification, regression, clustering, rules of association and visualization of results. It is also ideal for the development of new machine learning schemes.

It can be used to identify the different secret trends in your dataset and find many of the most important factors [6].

III. LITRATURE REVIEW

A review of the literature that many studies have been carried out in the predicting gene expression of Autism Spectrum Disorders "ASD" using machine learning methods. Those studies were conducted providing different methods in order to give more accuracy and appropriate solutions to find out the most affected genes in the ASD. The studies presented and categorized based on its novelty as follows:

In recent research Islam et al., (2019) the authors proposed an effective prediction model based on ML technique. Moreover, they developed a mobile application for predicting ASD for people in any age. They used two types of datasets, the AQ-10 dataset and real dataset in terms of the accuracy, the proposed model can predict autism with 92.26%, 93.78%, and 97.10% accuracy in case of child, adolescent and adult persons, respectively. Moreover, the proposed model can predict autism traits for different age groups, which is missed in many other existing approaches. The results showed marginal performance in terms of accuracy (77% to 85%) for real dataset. The research concluded that the Random Forest-CART showed better performance than the Decision Tree-CART algorithm, while the proposed (merging Random Forest-CART and Random Forest-ID3) algorithm provide better performance comparing to both the Random Forest-CART and Decision Tree-CART algorithm. Finally, a user-friendly mobile application has been developed for end users based on the proposed prediction model so that any individual can use the application to predict the autism traits easily [7].

While the objective of this study Ying Lin et al.(2018) was to employ a machine learning-based approach to predict ASD risk genes using human brain spatiotemporal gene expression signatures, gene-level constraint metrics, and other gene variation features. they compared the prediction accuracy of four machine learning algorithms using five-fold CV. The random forest model achieved the best prediction accuracy for autism risk genes with an AUC 88% [8].

However, Dong Hoon Oh et al (2017) tried to use a transcriptomic approach to identify a signature of gene expression with a promising performance in young adults with ASD diagnostic prediction. They used the Gene Expression Omnibus database released microarray data (GSE26415), which included 21 young adults with ASD and 21-year-old and sex-matched controls. Nevertheless, in young adults with ASD diagnostic prediction, Dong Hoon Oh et al (2017) tried to use a transcriptomic approach to classify a gene expression signature with promising quality. They used the microarray data released by the Gene Expression Omnibus registry (GSE26415), which included 21 young adults with ASD and 21-year-old and sex-matched controls. Using the R language limma package (adjusted p-value < 0.05), nineteen differentially expressed samples from a training data set (n=26, 13 ASD cases, and 13 controls) were identified and further analyzed using machine learning algorithms in a test data set (n=16, 8 ASD cases, and 8 controls). Examination of the hierarchical cluster revealed very well-discriminated subjects with command ASD. The validation of 19-DE samples with a test data set resulted in an overall class prediction accuracy of 93.8% as well as a sensitivity and specificity of 100% and 87.5% respectively, based on the support vector machine and K-nearest neighbor review [9].

Hameed et al., (2017) The purpose of this research is to improve the accuracy of gene classification for the spectrum of autism disorder by using filters and classifiers to find the most efficient and reliable approach. The GEO public database dataset consists of 146 samples and 54,613 genes. The samples were split into two classes, a monitoring class with 69 samples and a class of autism with 77 samples. Second, they use the different filters to exclude very similar genes by adding a medium and median ratio criterion. The data set was then divided into two parts; 85% of the data used in the model training and validation (testing) process; The other section, consisting of 15% of the data, was classified as an uninvolved subset to be used as a new real-world dataset. They then used statistical filters that are: the two-sample t-test (TT), the group correlation of features

(COR) and the Wilcoxon rank sum test (WRS). And the last step is the choice of genes by using a GBPSO-SVM wrapper-based algorithm along with the filters used. The advantage of using this algorithm is because GBPSO starts with a random number of selected genes and searches in each iteration for the appropriate subset of genes. Using 10-fold cross-validation, the SVM classifier is used to test the output of each candidate sub-set. The GBPSO algorithm contributes to the choice of an optimal sub-set of genes offering the highest accuracy of classification. The results showed that the most selective genes found in the first and last steps of selection included the existence of a repetitive gene (CAPS2), which was assigned as the gene most strongly correlated with risk of ASD. The combined gene subset selected by the GBPSO-SVM algorithm has been able to increase the accuracy of the classification [4].

Kou et al. (2012) used supervised techniques of machine learning to identify known and predict new genes associated with these diagnoses. Computational approaches were used, including two classifications based on networks and one classifier based on attributes. Lastly, 10 SVM classifiers were used using positive gene sets (i.e. genes associated with ASD or ID) and negative sets that were generated randomly with 200 genes in each set. The SVM performed better than both classifiers depending on the network. For truth information, two non-overlapping lists containing 114 known uncommon, high-risk ASD-related genes and 223 ID-related genes were created. The SVM classifiers are able to discriminate between ASD-related genes, ID-related genes and other genes with 80 to 98 percent accuracy. Classifier sensitivity ranged from 76% to 89%, specificity ranged between 89% and 96%, and AUC ranged from 94% to 97%. In addition, the ASD classifiers had better performance than the ID classifiers [10].

Finally, the research conducted in Kong et al., (2012) proposes an algorithm that helps to distinguish between 170 situations of ASD and 115 age / sex-matched controls and to evaluate the usefulness of gene expression profiling as a tool to help diagnose ASD. By using a cross-validation strategy, specifically on male samples, they developed a model of 55-gene prediction. Which with the testing group achieved 68 percent identification accuracy (area below the receiver operating characteristic curve (AUC): 0.70[95 percent confidence interval [CI]: 0.62–0.77]). On the other hand, the same does not work well with female specimens (AUC 0.51, 95 percent CI0.36–0.67). Research shows that blood expression analysis can be used to diagnose ASD. They indicated that the most accurate prediction model could be developed by measuring the coefficient of variation of AUCs with 100 external cross-validation tests. The review of numerical forecasting in the thesis was carried out using the packages of the caret and RWeka R library. Also, 5 additional prediction methods were tested; Logistic regression, Native Bayes, k-Nearest Neighbors, Random Forest, and Vector Machine Support using 55 genes with 5 fold LGOCV strategy to suggest the most possible approach [11].

1. LR Dataset Specifications

In this section we will describe the dataset in detail for the state of art studies.

Table 1: Gene Selection Algorithms for LR

Paper Ref.	Datasets	Number of classes	Number of samples	Number of genes
[4]	GEO(NCBI)	2	146	54,613
[9]	ASD microarray Dataset (GSE26415)	1	42	19
[8]	ASD and ID Game List	2	--	223
[10]	ASD microarray Dataset (GSE26415)	2	60000	1,288
[7]	AQ-10 database [12]	3	1,100	--
[7]	AQ-10 database [12]	3	1,100	--

2. LR Gene Selection Algorithms Specifications

In this section we will describe the applied algorithms and the corresponding accuracy for each study.

Table 2: Gene Selection Algorithms for LR

Page Ref.	Algorithm	Highest Accuracy
[4]	Filter Gene Selection t-test(TT) + SVM COR + SVM WRS + SVM Wrapper Gene Selection GBPSO – SVM	 86.3% 81.8% 83.8% 92.1%
[9]	SVM KNN LDA	93.8% 93.8% 68.8%
[10]	SVM	80% - 98%
[8]	The Random Forest Model	88%
[7]	Decision Tree-CART Random Forest-CART Random Forest-CART + Random Forest-ID3	89.2% 96.91% 97.10%
[11]	AUC	68%

IV. DATASET

In this section, the dataset attributes described in detail, also we will present the steps of preparing the dataset to be applicable to use in weka tool.

A. Dataset Description

We applied classification, and feature selection methods to the ASD dataset in this work. The dataset is publicly available in microarray repository GEO (NCBI) [1]. We have downloaded it after its normalized as part of previous work [4] where they initially filtered the genes in the original dataset resulting of reduced number of genes from 54,613 to 9454 genes.

The used dataset consists of 146 samples and 9454 genes (features). The samples are divided into two classes, a control class containing 69 samples and an autism class containing 77 samples. The autistic patients have been diagnosed by medical professionals (developmental pediatricians and psychologists) according to the DSM-IV criteria, and the diagnosis was confirmed on the basis of the ADOS and ADI-R criteria [1].

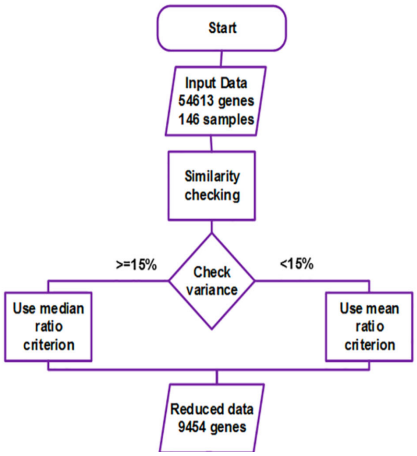

Figure 1 : Genes normalization process applied in the previous work [4]

Table 3: Dataset Details

Datasets	Number of classes	Number of samples	Number of genes
GEO (NCBI)	**2**	**146**	**9,454**

B. Data Pre-processing & Cleaning

The dataset consists of ID_REF column which indicate the genes. An addition Identifier column to identify the common name of each gene in the dataset. The samples were shown as rows.

In order to implement the dataset in WEKA, we have applied some changes in the dataset as follows: First, we excluded the identifier column as it might affect the filtration process negatively in Weka. Secondly, we have

added the class column and identify each sample whether its "Control" or "Autism" as per the sample information provided in the GEO site. After that, we transposed Excel columns to rows where samples in rows and features in columns. At the end, we transported the dataset file format from CSV to Ariff format.

Along with these dataset modifications, we validated the accuracy of attributes with weka pre filtration process. So, we validated that for each sample the minimum, maximum, mean and standard deviation measures are calculated correctly as shown in following figure.

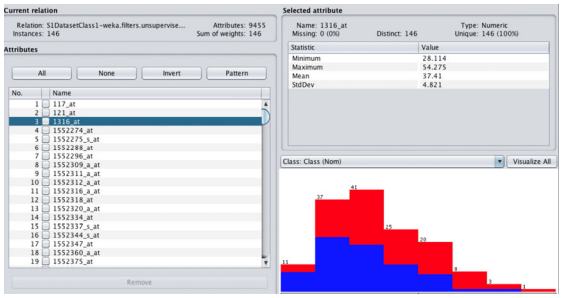

Figure 2: 1316_at gene statistics in Weka tool after data preparation process

V. PROPOSED MODEL

After various experiments and applying multiple combination of filters and classifiers, we proposed the model which shows the best accuracy conducted in the study.

The dataset first filtered by CFS feature selection, then will apply Bayes Networks classifier with 10-fold cross validation on dataset filtered by CFS feature selection method. The following figure shows the proposed model in this work.

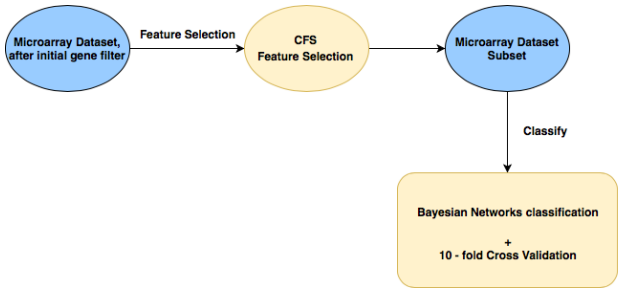

Figure 3 : The proposed Model for ASD dataset

VI. IMPLEMENTATION AND RESULTS

In this section, we will demonstrate the steps for gene expression implementation from different classifiers in weka, then will discuss the results of each experiment.

A. Implementation Stages

The experimental procedure of the current work was implemented in Weka tool through three basic steps; these are briefly described below.

First; the classification methods (Support Vector Machine "SMO", K-Nearest Neighbors "KNN", Random Forest, C4.5 decision tree "J48" and Bayesian Networks) that shows high accuracy in previous works were first applied to all datasets without performing any feature selection filters. Results of 10-fold cross validation have been shown in Table 4.

Table 4 : Percentage accuracy of 10-fold cross validation of classification methods for all genes

Classification methods	SMO	Random Forest	KNN	J48	BayesNet
Accuracy %	77.397%	70.547%	64.383%	64.383%	61.643%

In our dataset SMO and Random Forest performed better than other classification methods. The accuracy of KNN and J48 was the same 64.3836 %, the best accuracy 77.3% was for SMO. Minimum accuracy calculated was 61.64% for Bayes Network.

And then, we applied Correlation based (CFS) Attribute Selection filter that minimized the number of genes to 134 genes. CFS is a supervised filter method, we decided to choose it as it approved its efficiency in different gene datasets. As well as it considered one of the faster and less computationally expensive filter methods.

After that; we tested the classifiers SMO, Random Forest, KNN, J48 and Bayes Network with a supervised attribute filter, almost the accuracy performance classifiers were improved after applying feature selection method to the dataset. The best accuracy in Table 3 was for Bayes Network with 91.78 % after that Random Forest with 85 % then SMO with 80.8% and the last methods were KNN with 76% and the lowest accuracy was 74.65% for J48 classifier.

Moreover, to eliminate overfitting problem we have apply above classifiers with cross validation in different folds such as 2, 5, 10..etc. the highest accuracy results with 10-fold cross validation.

Table 5 : Percentage accuracy after attribute selection filter applied to the classification methods.

Classification methods	SMO	Random Forest	KNN	J48	BayesNet
Accuracy %	80.821%	85.616%	76.0274%	74.6575%	91.7808%

B. Results Analysis

The effect of feature selection is apparently appearing. Hence, pairwise combinations of the feature selection and classification methods were examined for our dataset as it is shown in figure 4 and figure 5 which show the percentage accuracy of 10-fold cross validation of classification methods for dataset before and after applying CFS feature selection method.

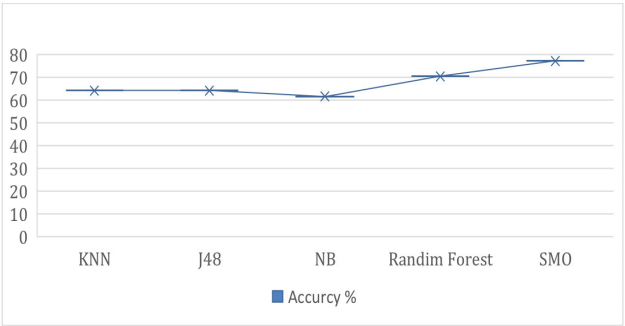

Figure 4 : Percentage accuracy of 10-fold cross validation of classification methods for all genes

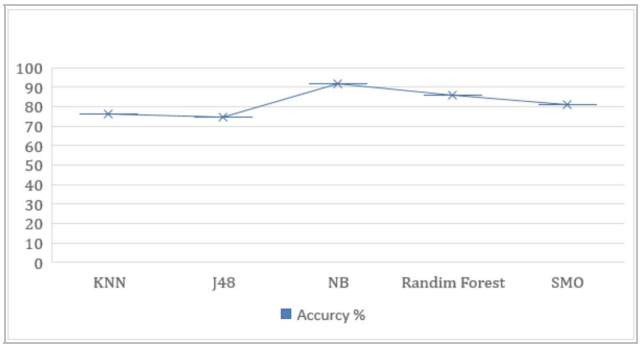

Figure 5 : Percentage accuracy after CFS feature selection filter applied to the classification methods

VII. DISCUSSION

In this section we will briefly conduct an analysis and comparison between our proposed model and the previous work model.

As shown in previous section, the best accuracy founded is from applying Bayes Networks classifier with 10-fold cross validation on dataset filtered by CFS feature selection method, which results of 91.7% accuracy.

We have compared our experiment results with the previous work in which they conduct study in the same dataset. Their study conducts both filter gene selection and wrapper gene selection methods, were they used Python for implementation. They have applied three filters (TT, COR and WRS) along with the SVM method where the highest accuracy shown is from TT and SVM which results of 86.3% accuracy.
Whilst our model results of 91.7% accuracy. This demonstrated that our proposed model has better accuracy when comparing to their filter gene selection methods. In the other hand, their wrapper gene selection proposed model (GBPSO-SVM) has higher accuracy 92.1% comparing to our model.
Yet, Filter methods are much faster than wrapper methods and therefore are better suited to high dimensional data sets.

Table 6: Comparation table of the proposed models

Paper	Gene Selection Type	Algorithm	# Genes	# Samples	Highest Accuracy
Previous work [4]	Filter Gene Selection	t-test (TT) + SVM COR + SVM WRS + SVM	9454	146	86.3% 81.8% 83.8%
	Wrapper Gene Selection	GBPSO - SVM			92.1 %
Our Proposed Model	Filter Gene Selection	CFS + Bayes Networks classifier with 10-fold	9454	146	91.7%

VIII. CONCLUSION AND FUTURE WORK

In this paper a brief review of different machine learning approaches employed in the selection and identification of the genes that are most relevant to autism are presented.

However, one of the most important tasks in conducting gene expression analysis using machine-learning algorithms is the building of a classification model that recognizes the discriminative genes with the highest possible accuracy. Depending on the results of our experiments in using different classifiers, we have proposed the CFS attribute selection with Bayes Networks classifier with 10-fold cross-validation that presented the highest accuracy, which results in 91.7% accuracy. The proposed model has better accuracy when comparing to the filter gene selection methods proposed in the previous work. for future work, we suggest using a hybrid feature selection method which is effective for gene selection when applied in microarray gene expression profiling, also we suggest using Python which may results better accuracy and had wide methods than weka.

References

[1] "GEO DataSet Browser." [Online]. Available: https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4431. [Accessed: 05-Oct-2019].

[2] "Autism spectrum disorders." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders. [Accessed: 05-Oct-2019].

[3] K. K. Hyde et al., "Applications of Supervised Machine Learning in Autism Spectrum Disorder Research: a Review," Rev J Autism Dev Disord, vol. 6, no. 2, pp. 128–146, Jun. 2019.

[4] S. S. Hameed, R. Hassan, and F. F. Muhammad, "Selection and classification of gene expression in autism disorder: Use of a combination of statistical filters and a GBPSO-SVM algorithm," PLoS ONE, vol. 12, no. 11, p. e0187371, 2017.

[5] P. B. Brazdil, Ed., Machine Learning: ECML-93: European Conference on Machine Learning, Vienna, Austria, April 5-7, 1993. Proceedings. Berlin Heidelberg: Springer-Verlag, 1993.

[6] Introduction to data mining with case studies. 2015.

[7] K. S. Omar, P. Mondal, N. S. Khan, M. R. K. Rizvi and M. N. Islam, "A Machine Learning Approach to Predict Autism Spectrum Disorder," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 2019, pp. 1-6.

[8] Y. Lin, A. M. Rajadhyaksha, J. B. Potash, and S. Han, "A machine learning approach to predicting autism risk genes: Validation of known genes and discovery of new candidates," 2018.

[9] Oh, Dong Hoon et al. "Predicting Autism Spectrum Disorder Using Blood-based Gene Expression Signatures and Machine Learning." Clinical psychopharmacology and neuroscience : the official scientific journal of the Korean College of Neuropsychopharmacology vol. 15,1 (2017): 47-52. doi:10.9758/cpn.2017.15.1.47

[10] Y. Kou, C. Betancur, H. Xu, J. D. Buxbaum, and A. Ma'ayan, "Network- and attribute-based classifiers can prioritize genes and pathwaysfor autism spectrum disorders and intellectual disability," Am J Med Genet C Semin Med Genet, vol. 160C, no. 2, pp. 130–142, May 2012.

[11] "Characteristics and Predictive Value of Blood Transcriptome Signature in Males with Autism Spectrum Disorders." [Online]. Available:

[12] "UCI Machine Learning Repository." [Online]. Available: https://archive.ics.uci.edu/ml/index.php. [Accessed: 05-Oct-2019].