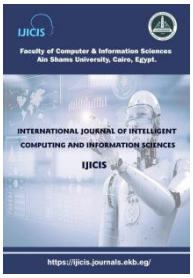
IJICIS, Vol.21, No.2, 65-81
DOI: 10.21608/ijicis.2021.61582.1058

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/



MACHINE LEARNING TECHNIQUES BASED ON FEATURE SELECTION FOR IMPROVING AUTISM DISEASE CLASSIFICATION

Basma Ramadan Gamal Elshoky*

Information technology section, Korean Egyptian faculty for Industry and Energy Technology, Beni Suef Technological University, Beni Suef, Egypt. Computer Science,Faculty of Science, Minia University, Minia, Egypt.
basma.r.gamal@gmail.com

Osman Ali Sadek Ibrahim

Computer Science, Faculty of Science, Minia University, Minia, Egypt.
osmaneg200@gmail.com

Abdelmgeid Amin Ali

Computer Science, Faculty of Science, Minia University, Minia, Egypt.
abdelmgeid@yahoo.com

Received 2021- 2-7; Revised 2021-4-27; Accepted 2021-4-30

Abstract: *Nowadays, Autism Spectrum Disorder (ASD) is one of the primary psychiatric disorders illness that rapidly increases. One of the main problems of medical diagnosis data and classification is the variance in symptoms between patients. Thus, finding the discriminative symptoms that distinguish the illness accurately is an important issue. This paper will explore various feature selection methods on four ASD datasets for extracting significant features for improving the ASD classification system. Datasets were created in 2017 and 2018 for child and adult gathered online. Several feature engineering techniques are applied to rank significant features. The correlation matrix method showed the association between features that enable us to select the highest significant features. Then each dataset split into 70% for training and 30% for test. Several machine learning classifiers are applied. After testing, the selected features achieve 100% accuracy, specificity, sensitivity, AUC, and f1 score with adaboost, linear discriminant analysis and logistic regression classifier on different size of data. I choose the adaboost model because it does the same performance with less time and less computational*

* Corresponding author: Basma Ramadan Gamal Elshoky
Information technology section, Korean Egyptian faculty for Industry and Energy Technology, Beni Suef Technological University, Beni Suef, Egypt.Computer Science,Faculty of Science, Minia University, Minia, Egypt.
E-mail address: basma.r.gamal@gmail.com

power in both dataset 2017 and 2018 for child and adult. Results were validated using cross-validation with 10 k-fold. The code applied in that paper in https://github.com/BasmaRG/ASD/ .

Keywords: *machine learning, AQ-10, logistic regression, correlation matrix, classification, autism spectrum disorder, Autism*

1. Introduction

ASD refers to a wide continuum of associated cognitive and neurobehavioral disorders and it affects a person's behaviour and performance. Autism affects verbal and non-verbal communication in social interaction. ASD has three features: 1) impairments in socialization, 2) impairments in verbal and nonverbal communication, and 3) restricted and repetitive patterns of behaviours [9]. A psychiatrist Leo Kanner [10] is the first one who describes a syndrome of "autistic disturbances" in 1943. He studied the case histories of 11 children who presented between the ages of 2 and 8 years. Then in 1988, Allen [11] describes it with the phrase autistic spectrum disorder. Early diagnosis of autism is essential for educational planning and treatment early. It is help provision for family education, supports, reduction stress, and the delivery of appropriate medical care to the child [12]. Autism symptoms can occur at any age. Thus, autism detection category can be split into four groups depending on age which are adult, adolescent, child, and toddler. There are many datasets available online such as functional magnetic resonance imaging (MRI), national survey of children's health (NSCH), and behaviour screening which are used to detecting ASD. ABIDE (Autism Brain Imaging Data Exchange) is a collaboration of 16 international imaging sites were collect several neuroimaging data from 539 individuals that suffering from ASD and 573 cases are typical controls. These 1112 instances are composed of structural and resting-state functional MRI data along with an extensive array of phenotypic information [33]. NSCH survey includes data about children from the age of 2 to 17 across every state in the United States of America and contains answers from primary caretakers of these children, data found at CDC website [34]. The behaviour screener is considered the most used in the world. A behaviour screener takes a few times and it doesn't need any equipment, and its data is easy to be understand. There are many behaviour screener methods that play an important role in detect ASD such as: 1) screening tool for autism in toddlers and young children (STAT), 2) childhood autism rating scale (CARS-2), and 3) autism spectrum quotient (AQ)[24].

2. Related work

This paper uses the autism spectrum quotient dataset called AQ-10 behaviour screening for adult and child. This dataset was used previously in [5], [6], [7], [8], [13], [23] However, this research has not provide the code for their work for research reproducibility. Thabtah and Peebles [13] proposed rules-machine learning to enhance classification performance. Thabtah [5] used the AQ-10 dataset for three group child, adolescent, and Adult. He used wrapping methods that integrate naïve bayes for select features for each group, applied two machine learning algorithms logistic regression (LR) and naive bayes (NB). LR outcome accuracy 92.80% for child, 91.34% for adolescent, and 95.73% for adult. NB outcome accuracy 97.94% for child, 97.23% for adolescent, and 99.85% for adult. Vaishali and Sasikala [6] were applied binay firefly feature selection wrapper in child dataset with NB, support vector machine (SVM), k-nearest neighbors (KNN), J48, and multilayer perceptron (MLP) algorithm using R

and WEKA. They compare algorithms performance before and after the feature selection process. After the selection feature process, the algorithms NB, KNN, and J48 were improved. However, the top accuracy they achieved after feature selection is 97.95%. Omar [23] collected three real data for the child, adolescent, and adult groups based on AQ-10 questions. He evaluated proposed techniques merging random forest-CART (Classification and Regression) that out-come accuracy 92.26%, 93.78%, and 97.10% accuracy in the child, adolescent, and adult. The performance of model in real dataset less than AQ-10 dataset. Akter [7] apply several features selection method and classification algorithms in AQ-10 datasets for adults, adolescent, child, and toddler. He obtained the best result for adult dataset

Table 1: summery of features that selected in previous studies with child and adult

using the adaboost algorithm. Z-score and Glm-boost for adolescents. He achieved 97.20%, 93.33%, 98.36, and 98.77 in child, adolescent, adult and toddler. The top accuracy he achieved is with selected features is 98.36% in adult. [8] Applied NB, KNN, SVM, LR and congenial neural network (CNN) in adults, adolescent, and child datasets. He hasn't to attention to make a selection features in basic ML algorithms so his result may be not accurate. Table 1 show you the summery of features in previous studies with child and adult. In [6], [8], [23] some general features are include such as {country, used the app before, gender, and more}. These features not have association with other features and will bad effect on the classification accuracy. Some questions also are not selected in [5], [7] may be effect on the performance of the classifier. Thus may be explain the high accuracy 99.85 result that the Thabtah [5] obtained in adult with LR and when he leave other features (question) the result in child with LR is 97.10%. Also the study [5] did not observe the tools, techniques and other configuration is used to achieve those accuracy that confirm the results are true. Although several approaches and tools have been developed to select features for analyze and detect the autism however, existing tools are not concentrated on the correlation between each variable and another on the datasets. The selecting features without strong relation between them will increase the training time and reduce classification accuracy. This paper will use that the criteria of feature selection that not used before in classification autism problem that will improve the accuracy of the classification system and reduce the time of learning.

	Previous study [5]	Previous study [6]	Previous study[7]	Previous study [8]	Previous study [23]
No. of features in the dataset	Adult:12 Child: 4	Child :10	Adult:1 Child: 2	Adult:21 Child:21	Adult:16 Child: 16
Feature selection method	Decision tree algorithm	Binay firefly feature selection wrapper	CFSSE, GRAE, IGAE, and RFAE	-	Wrapping methods that integrates Naïve Bayes classifier
Features	Adult: Q1 to Q10, gender and used the app before Child: Q1, 4, 8 and Q10	Q1, 2, 3, 4, 5, 7, 8, 9, Q10 and relation	Adult: Q5 Child: Q9 and Q4	All features	Q1 to Q10, gender, ethnicity, jaundice, autism, country of res and result

3. Proposed approach

Figure 1 show a flowchart diagram of the proposed system based on features engineering. This study used python programming language version 3.7 with packages scikit-learn version 0.21.2, pandas 0.24.2 in windows 10 64-bit, Intel I5, 4096 MB RAM, and AMD Radeon HD card. The code also re-executed in Colab and Kaggle platform for ensure results.

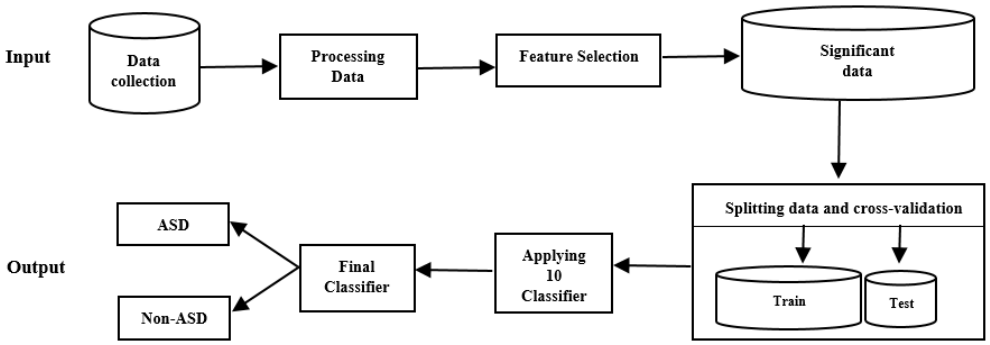


Figure. 1: Proposed system flowchart

Table 2: Summery of datasets

Dataset	Age	N0. of instance	ASD	Not ASD
Adult 2017	18 and more	704	189	515
Child 2017	4 -11	292	141	151
Adult 2018	18 and more	1118	358	760
Child 2018	4 -11	509	257	252

Table 3: Dataset Features

Feature name	Feature type	Feature description
10 Questions	Binary	The answer code of the question based on the screening
Age	Integer	Age in years
Gender/sex	String	Male or Female
Ethnicity	String	List of common ethnicities in text format
Jaundice	Boolean (yes or no)	Whether the case was born with jaundice
Austim/Family ASD	Boolean (yes or no)	Family member with PDD Boolean (yes or no) Whether any immediate family member has a PDD
Country/Residence	String	List of countries in text format
Used app before	Boolean (yes or no)	Whether the user has used a screening app
Result/Screening Score	Integer	The final score obtained based on the scoring algorithm of the screening method used.
Screening Type/Age description	Integer	The type of screening methods chosen based on age category (child, adolescent, adult)
Relation	String	Who is completing the test Parent, self, caregiver, medical staff, clinician, etc.
Class/ASD	Boolean (yes or no)	Have autism or not
Additional feature in 2018		
Language	String	Application Language

Why taken the screening	String	Why taken the screening

Table 4: Autism Spectrum Questions for adult

Question1	I often notice small sounds when others do not.
Question2	I usually concentrate more on the whole picture, rather than the small details.
Question3	I find it easy to do more than one thing at once.
Question4	If there is an interruption, I can switch back to what I was doing very quickly.
Question5	I find it easy to 'read between the lines' when someone is talking to me.
Question6	I know how to tell if someone listening to me is getting bored.
Question7	When I'm reading a story I find it difficult to work out the characters' intentions.
Question8	I like to collect information about categories of things (e.g. types of car, types of bird, types of train, types of plant etc).
Question9	I find it easy to work out what someone is thinking or feeling just by looking at their face.
Question10	I find it difficult to work out people's intentions.

Table 5: Autism Spectrum Questions for child

Question1	S/he often notices small sounds when others do not.
Question2	S/he usually concentrates more on the whole picture, rather than the small details.
Question3	In a social group, s/he can easily keep track of several different people's conversations.
Question4	S/he finds it easy to go back and forth between different activities.
Question5	S/he doesn't know how to keep a conversation going with his/her peers.
Question6	S/he is good at social chit-chat.
Question7	When s/he is read a story, s/he finds it difficult to work out the character's intentions or feelings.
Question8	When s/he was in preschool, s/he used to enjoy playing games involving pretending with other children.
Question9	S/he finds it easy to work out what someone is thinking or feeling just by looking at their face.
Question10	S/he finds it hard to make new friends.

Table 6: Feature types and

Type	binary	numeric	categorical	string
Feature name	A1	age and result	gender and class	country
Feature Value	0 or 1	continuous number such as 4, 5, …64	f (female) or m(male) and yes or no	Egypt

3. 1. Dataset

This paper gathered two versions of the Autism spectrum quotient (AQ) AQ-10 dataset for child and adult. AQ is a tool for screening autism created in 2001 by Baron-Cohen [1]. He made the tool with 50 items questionnaires and gave individuals score for in the range 0-50. 2012, Allison [5] reducing items tool to 10 questionnaires, the score will be in the range 0-10. The final score calculated by the application by summation all answers. Each answer to a question set of value 1 when the answer is either definitely or slightly Agree, otherwise 0 is set. The person will have ASD if result (>= 6). Data gathered online, first version 2017 through UCI machine learning repository [30] and second version 2018 through the Fadi Fayez website [31]. Fadi gathered these data-sets through a mobile application called ASD Tests [32] that he developed, based on the AQ-10 behavior screening tool. The summery of datasets presented in table 2, it shows the number of all instance, asd, non-asd cases, and the age for each category. Table 3 describes the features in each dataset. Tables 4 and 5 are describe questions for adult and child [2]. Table 6 describes the data types of features. Data types are four types' numeric, nominal (categorical), string, and binary.

3. 2. Feature selection

The feature selection process is an important task for building accurate classification system. For this task, this paper applied a descriptive statistic to gain better understand variables/features in the dataset. Python packages pandas, seaborn, matplotlib and sklearn [4] helped us to exploratory, visualize, processing features. I pre-processed dataset before feature selection. There is a little missing data in columns, solved them by fill in missing by the median in the case of numeric value or drop in the case of a string value. Some data transformation did by transform category data such as yes and no to binary data 1 and 0. Two techniques filter and feature selection are executed on the dataset. We used univariate filter methods that have an advantage that select features instituted on a performance measure and faster than the wrapper approach. The filter method results are better because it is not dependent on the algorithm will use in the evaluation [21,22]. Then applied feature selection process. Three statistics methods CHI, Analysis of variance (ANOVA) and correlation matrix are applied to rank features for the feature selection process.

3. 3. Data splitting

The k-fold cross-validation techniques are used to split dataset to train, test, and validation the classification ASD model. Train-test splits dataset into a random train and test subsets. This method depends on the size of the dataset, I split each dataset to 70% for training and 30% for test. The cross-validation method (CV) split dataset randomly into K subsets or folds. The ideal value of k is 10. The method is repeated k times. Each time, one of the k subsets is used as the test set and the other k-1 subsets are put together to form a training set [29]. Each fold split to train and test that is make all dataset trained and tested.

3. 4. Evaluation ML

Ten supervised classification algorithm logistic regression (LR), linear discriminant analysis (LDA), decision tree classifier (CART), NB, KNN, SVM, adaboost (AB), gradient boosting (GBM), random forest (RF) and extra trees classifier (ET). Their performance execution was measured by time and classification accuracy.

A brief for popular supervised machine learning algorithm:

- **Decision Tree:**
  A decision tree (CART) is an algorithm based on classification and regression trees, developed by Breiman in 1984. The CART construct the model by recursively partitioning the data space and fitting a simple prediction model within each partition. The CART algorithm has advantages: it is nonparametric, flexible, can adjust in time, no assumptions, and computationally fast [14].
- **Discriminant Analysis:**
  Linear discriminant analysis (LDA) is a probability method used for dimensionality reduction and data classification which is proposed by R. Fischer in 1936 [15]. You can use the LDA algorithm for multi-classification problems (more than one class).

- **Boosting:**
  Boosting is a machine learning approach, combines many relatively weak and inaccurate rules for building a highly accurate prediction rule [26]. The primary concept of boosting is to add new models to the ensemble sequentially [16]. Gradient Boosting Machine (GBM) and Ada Boost (AB) is an ensemble boosting algorithm. The concept of GBM is to build the new base-learners to maximally correlate with the negative gradient of the loss function, associated with the whole ensemble. The concept of AB is based on interactively combining multiple less performing classifiers to generate a better-performing classifier. The basic rule of AD is to set the weights of classifiers and the training data sample in each iteration such that it ensures accurate predictions of unusual instances [17].
- **Logistic Regression:**
  Logistic regression (LR) is popular mathematical modeling, named for the function ' logistic function' used at the core of the method [25]. It is also called the sigmoid function. LR algorithm can use only for binary classification problems (only two classes).

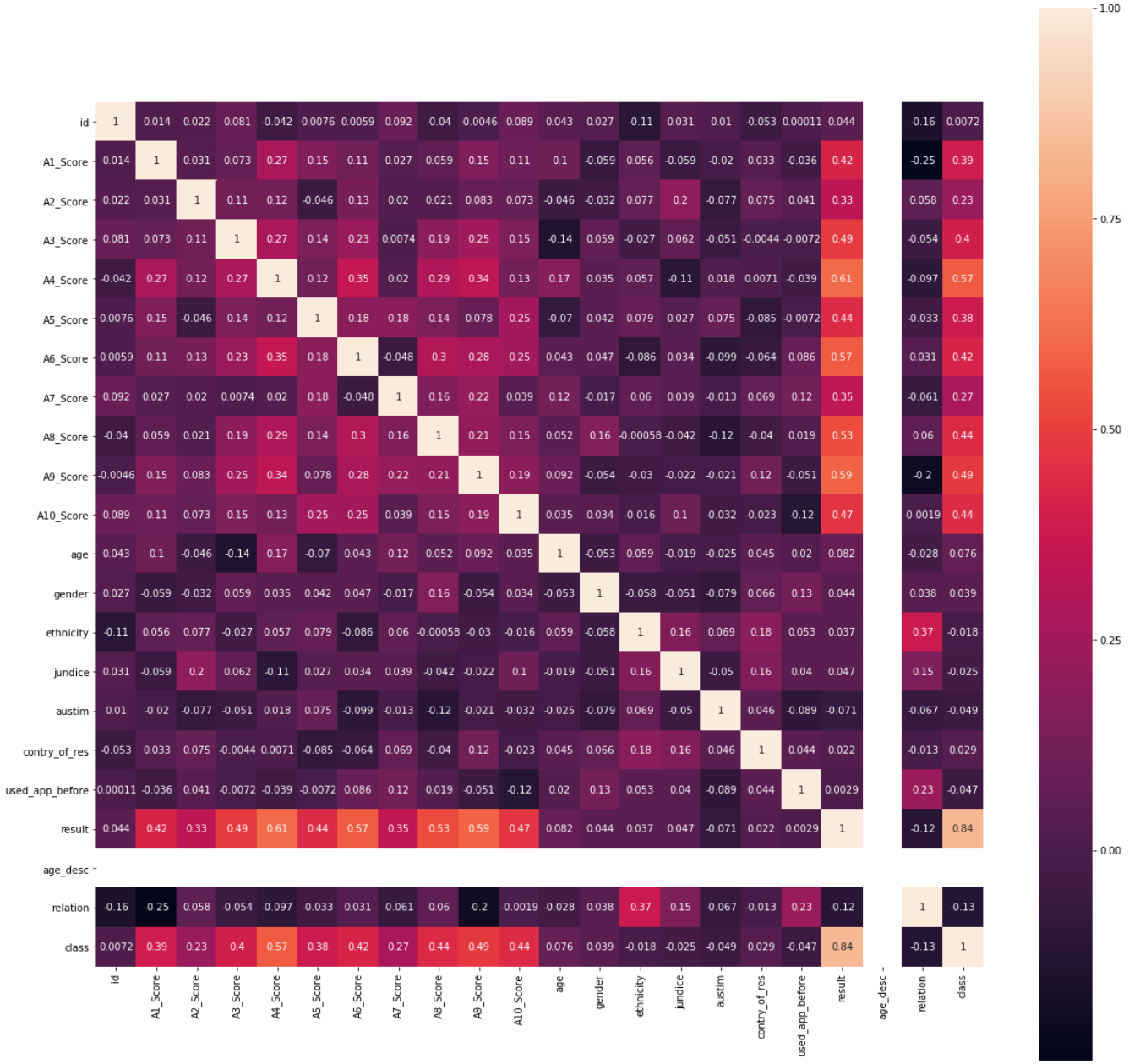


Figure. 2: The correction between features in Child 2017 dataset

- **Support Vector Machine:**
  A support vector machine (SVM) is a universal learning machine introduced by Smola and Vapnik (1997). SVM parameterized by set weights and support vectors to make the decision, also characterized by a kernel function [27].
- **Random Forest:**
  The random forests (RF) technique is an ensemble method that utilizes rankers based on bagging and sampling features [18]. Bagging refers to the procedure of combining multiple decision trees and calculating their average.

- **Naive Bayes:**
  Naive Bayes (NB) is a simple learning algorithm based on the Bayes rule. It is using the information in-sample data to calculating the posterior probability P(y | x) (where y is the class 'y and 'x' is an object) [19]. You can use the NB algorithm for binary (two-class) and multi-class classification.
- **Extra Trees:**
  Extra Trees (ET) used the classical top-down procedure to build an ensemble of decision trees. It splits nodes by choosing cut-points fully at random and uses the whole learning sample to grow the trees [20].
- **K-Neighbors:**
  K-Neighbors (KNN) used the K-closest samples from the training set to predict a new sample. The K-closest training set samples are determined via the distance metric like Euclidean and Minkowski [28].

This paper applied several evaluation metrics of binary classifier systems to represent the performance of different classification models and compare their performance based on these metrics. Metrics are classification accuracy, classification/error rate, specificity, sensitivity, area under the curve, and f1 score represented by confusion metrics. Table 7 describe the confusion matrix for a binary classification problem (which has only two classes - positive and negative). The confusion metrics is used to summarize the performance of a binary classification tasks represented by calculating the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) values calculated as follow [3].

Classification accuracy: calculated by the following formula:

$$\boldsymbol{ACC} = \frac{\boldsymbol{TP} + \boldsymbol{TN}}{\boldsymbol{FP} + \boldsymbol{FN} + \boldsymbol{TP} + \boldsymbol{TN}}$$

Sensitivity: is synonymous to recall and the true positive rate which calculated by the following formula:

$$\boldsymbol{SEN} = \frac{\boldsymbol{TP}}{\boldsymbol{FN} + \boldsymbol{TP}}$$

Specificity: is synonymous to the true negative rate which calculated by the following formula:

$$\boldsymbol{SPC} = \frac{\boldsymbol{TN}}{\boldsymbol{TN} + \boldsymbol{FP}}$$

F1 score: can be interpreted as a weighted average of the precision and recall, where an f1 score reaches its best value at 1 and worst score at 0. Summarizes both precision and recall which calculated by the following formula:

$$\boldsymbol{F1} = \mathbf{2} * \frac{\boldsymbol{precision} * \boldsymbol{recall}}{\boldsymbol{precision} + \boldsymbol{recall}}$$

Precision: calculated by the following formula:

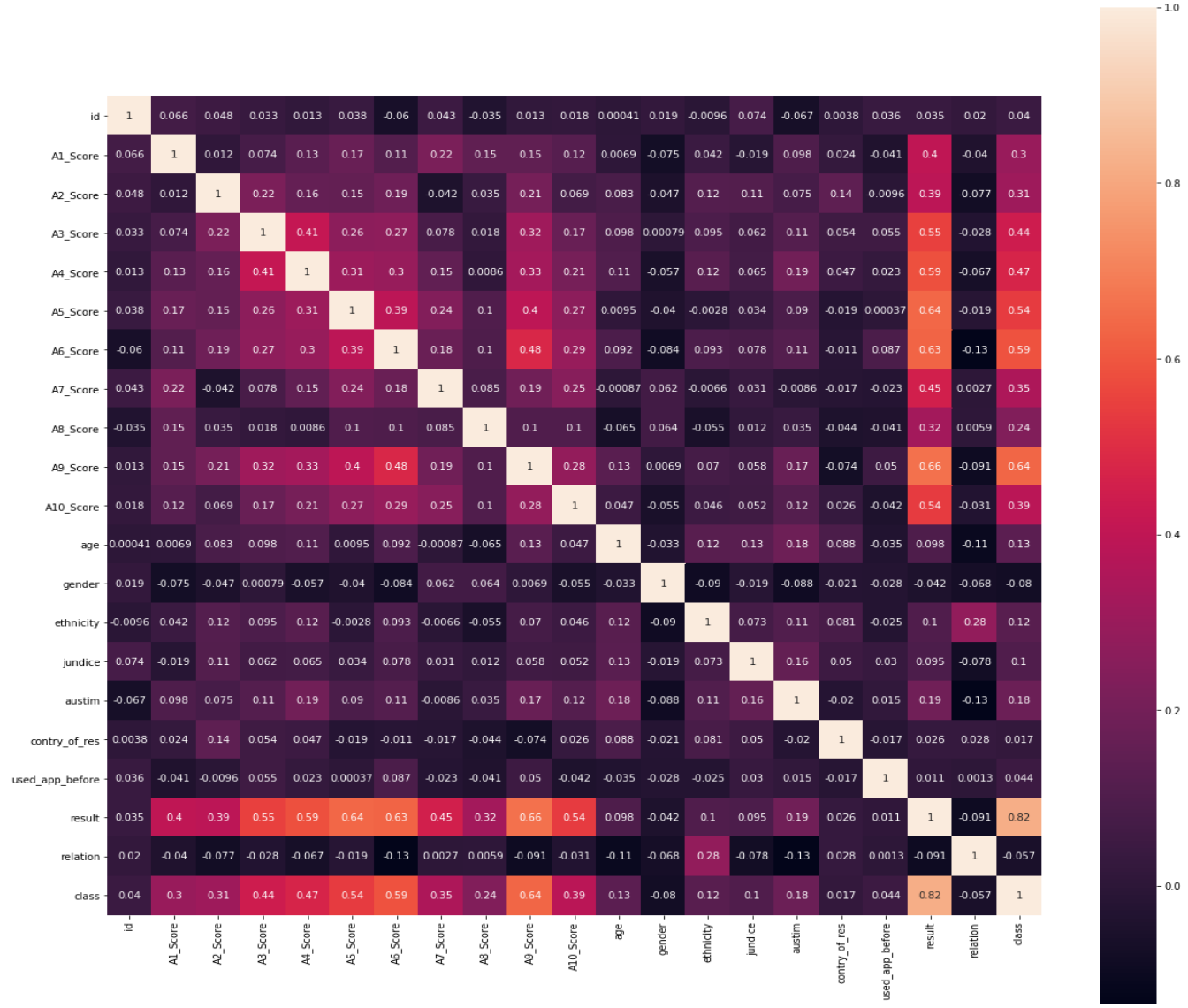


Figure. 3: The correction between features in adult 2017 dataset

$$PRE = TP * \frac{TP}{FP}$$

4. Experiment Results

The results examined the feature selection using statistic methods CHI, ANOVA, and correlation matrix (denoted as FS1, FS2, and FS3). Figure 2, 3, 4, and 5 illustrate association/relationship between variables/features in each dataset. An instance of row data {4, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 5, m, white, no, no, Russia, no, 8, 4-11 years, russian, parent, YES} after remove un significant feature will be {4, 0, 1, 1, 1, 1, 1, 0, 1, 1 ,1}. The rank correlation measures the linear association between two variables class and each variable. The association degrees that i followed are:

correlation > 0.75 Very strong association
0.75< corr* >0.5 Moderate positive association
0.5<corr* > 0.25 week positive association
0.25<corr* > 0.0 Negligible positive association
corr* <= 0 No association

Table 7: Confusion Matrix

	Predicted Positives	Predicted Negative
Actual Positives	TP	FN
Actual Negative	FP	TN

TABLE 8: Feature rank

Dataset	Child2017			Child2018			Adult2017			Adult2018		
Feature selection method	FS1	FS2	FS3	FS1	FS2	FS3	FS1	FS2	FS3	FS1	FS2	FS3
A1_Score	16.57	53.14	0.39	10.48	3.667	0.37	13.36	68.22	0.3	8.286	3.244	0.29
A2_Score	7.134	16.05	0.23	6.534	1.635	0.21	37.32	75.37	0.31	11.00	2.164	0.31
A3_Score	11.73	53.78	0.4	2.507	1.134	0.4	74.31	169.5	0.44	19.60	4.084	0.44
A4_Score	42.33	138.4	0.57	4.157	1.016	0.58	78.40	198.9	0.47	14.40	3.178	0.47
A5_Score	10.82	48.90	0.38	2.978	1.170	0.41	101.7	284.4	0.54	24.06	5.213	0.57
A6_Score	14.62	61.13	0.42	5.198	1.976	0.46	176.6	378.9	0.59	24.74	3.756	0.62
A7_Score	8.63	23.52	0.27	3.693	1.027	0.33	50.63	98.91	0.35	7.685	1.384	0.38
A8_Score	28.25	68.99	0.44	9.154	2.039	0.43	13.89	41.83	0.24	4.345	1.237	0.26
A9_Score	34.98	89.75	0.49	16.38	3.669	0.45	192.2	475.7	0.64	19.11	3.080	0.6
A10_Score	15.48	69.60	0.44	6.354	2.435	0.4	44.67	122.8	0.39	13.43	3.485	0.4
age	1.431	1.650	0.075	9.535	1.137	0.088	22.73	2.489	0.059	74.77	2.068	0.076
Gender/Sex	0.126	0.436	0.039	11.16	4.590	0.024	2.177	4.564	-0.08	4.303	0.920	-0.069
ethnicity	0.202	0.091	-0.018	28.58	2.193	0.033	26.79	11.10	0.12	140.6	9.389	0.18
jaundice	0.133	0.182	-0.025	10.50	1.559	-0.001	6.626	7.402	0.1	22.31	2.505	0.082
autism/Family ASD	0.578	0.692	-0.049	8.830	1.182	-0.015	19.29	22.81	0.18	27.42	3.345	0.15
country_of_res	2.461	0.243	-0.029	979.5	6.226	0.049	1.695	0.200	0.017	1246	10.70	0.046
Relation	4.047	4.736	-0.13	2.269	1.705	0.035	0.353	2.282	-0.057	1.081	1.311	0.002
result/Score	170.1	672.4	0.84	12.34	1.615	0.83	608.8	1456	0.82	87.60	7.320	0.83

In Figure 2 the best association degrees of features with other features in the rang of 0.54 to 0.23 where A4_Score feature with value 0.54 is a Moderate positive association and 0.23 is a Negligible positive association. Table 8 represented the rank of the child and adult dataset 2017 and 2018 features. After analysis, I selected features using the correlation matrix methods for child and adult dataset in four datasets. The significant features are A1_Score, A2_Score, A3_Score, A4_Score, A5_Score, A6_Score, A7_Score, A8_Score, A9_Score, and A10_Score based on the association between features. The experimental results of testing the model classifiers are in figure 6. Table 9 compares the performance of LR, NB, KNN, and SVM classifier with [5], [8] and [6] in dataset AQ-10 2017. The LR achieves higher accuracy than [5] and [8] in the child and adult. The NB improves the accuracy of [5]and [8] only in the child. The accuracy of SVM is almost similar to [8] in the child while it improved to 99.29 in adult. KNN is also improved only in adult.

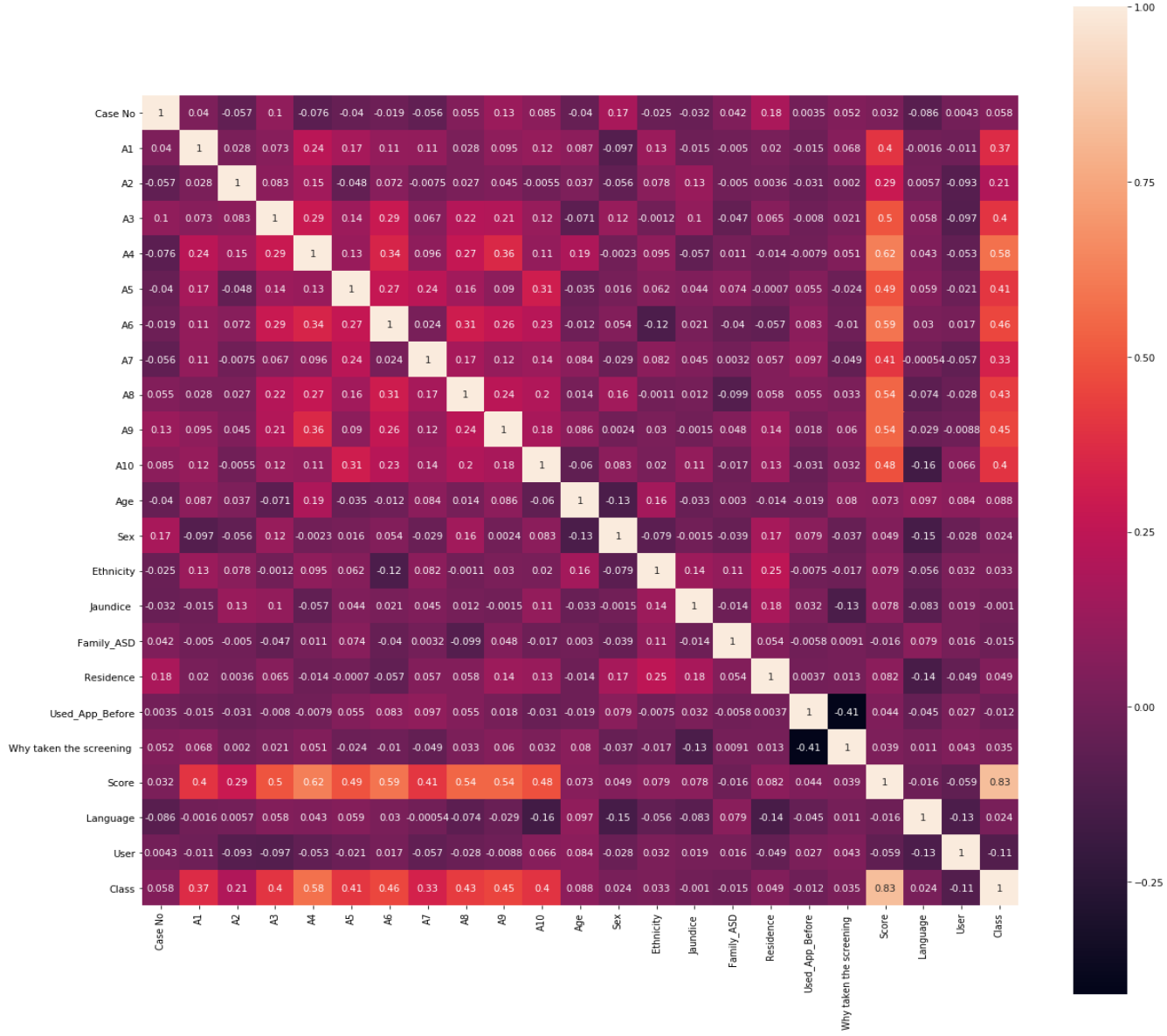


Figure. 4: The correction between features in Child 2018 dataset

The result in figure 6 showed that the classifier AB and LR in (a,b,c,d) achieved 100% accuracy, specificity, sensitivity, auc and f1 score for adult and child in the dataset 2017 and 2018. LDA in (c) achieved 100% accuracy, specificity, sensitivity, auc and f1 score for child dataset 2017. The cross-validation results are in Figure 7 ensures the results of AB, LR, and LDA obtained using the train-test split technique figure 6. The LR classifier will be the main classifier for building an ASD classification system in adult and child. I choose the model that takes less computational power because it does the same performance with less time and less computational power. And because adaboost is an ensemble model it is more complex. Table 10 compares the performance of our proposed model with [5],[6], [23], [7], and [8] in dataset 2017. Our model achieves higher performance rather than all previous models.

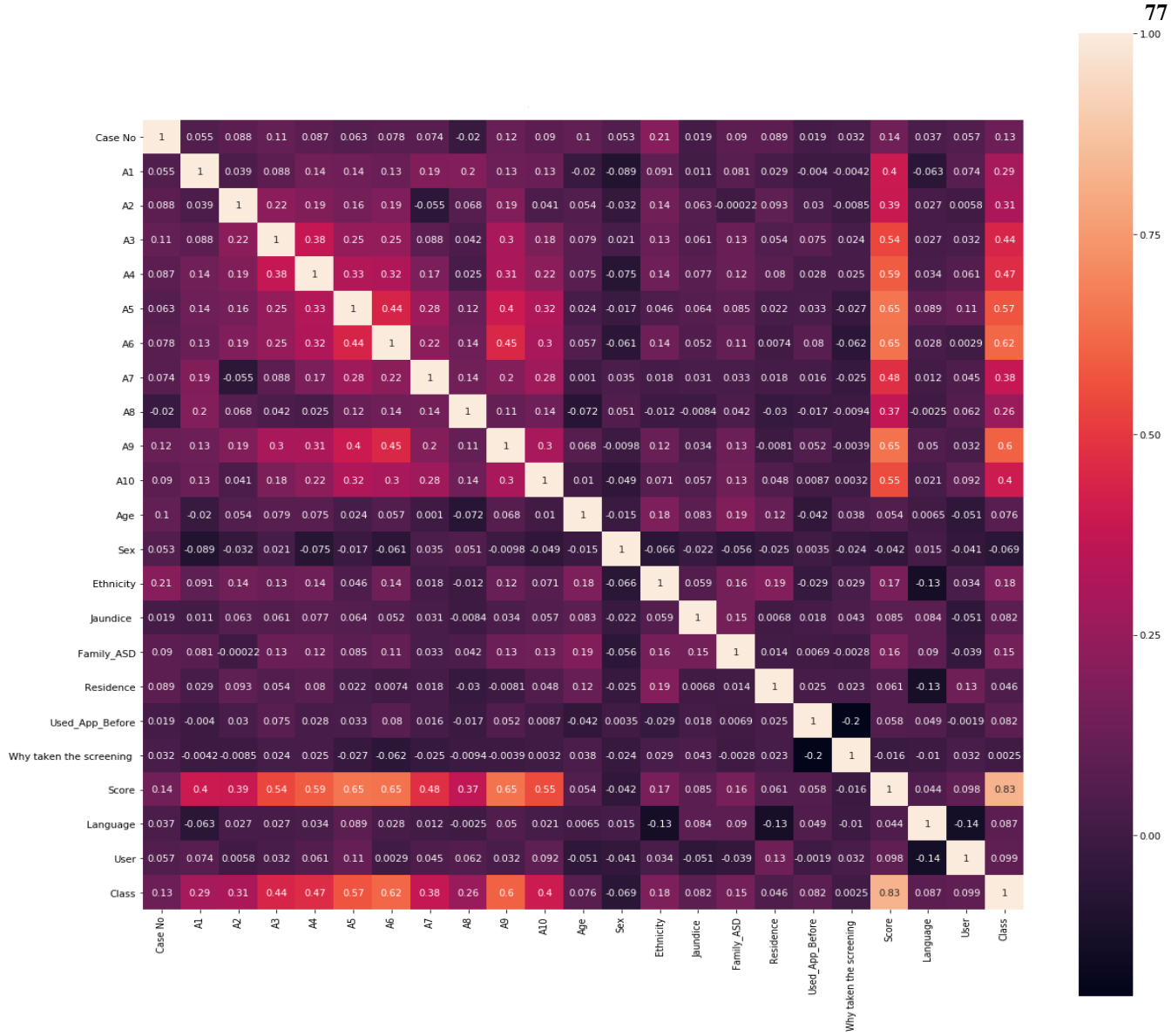


Figure. 5: The correction between features in adult 2018 dataset

Table 11 compares the performance of our model in the dataset 2017 with 2018 (denoted as V1 and V2). The performance of our model is the same in AQ-10 datasets 2017 and 2018. The performance measurement classification accuracy, specificity, sensitivity, area under the curve and f1 score are the same 100% for adult and child. The time performance is the same 0.078 seconds in the child while time adult 2017 is 0.062 seconds and 2018 is 0.127 second.

5. Conclusion and feature work

Because of the increasing of people with ASD every day, Researchers in the field of AI tried to make a prediction system to classify ASD early. Since the performance of these systems needs to improve this study did this. This proposed study applied feature engineering as a machine learning technique in the AQ-10 dataset 2017 & 2018 for adult and child for improving ASD classification system. They steps

are filter, select features, splitting dataset, classification algorithms, measure time execution, and performance. This paper using confusion metrics for calculating classification accuracy, classification/error rate, specificity, sensitivity, area under the curve, and f1 score. The outcome of the proopsed approach prove that the feature engineering improved accuracy comparing with [5], [6], [7], [8], [23] study in dataset 2017. This also approached successful with another different size of ASD screening dataset comparing with the AQ-10 dataset 2018. The performance of our proposed model is better than other studies. When comparing version 2017 and 2018 is same performance in adult and child. The paper is the first study that achieve 100% for child and adult 2017. The first study also is using ASD screening version 2018 for adult and child. Also it is provide a public code for reusability. However, I offered an efficient approach for classification ASD but the limitation of this paper is applied only ASD classification on numeric dataset with machine learning techniques. In the future, I will apply ASD classification on another types of dataset with new techniques such as computer vision and deep learning.

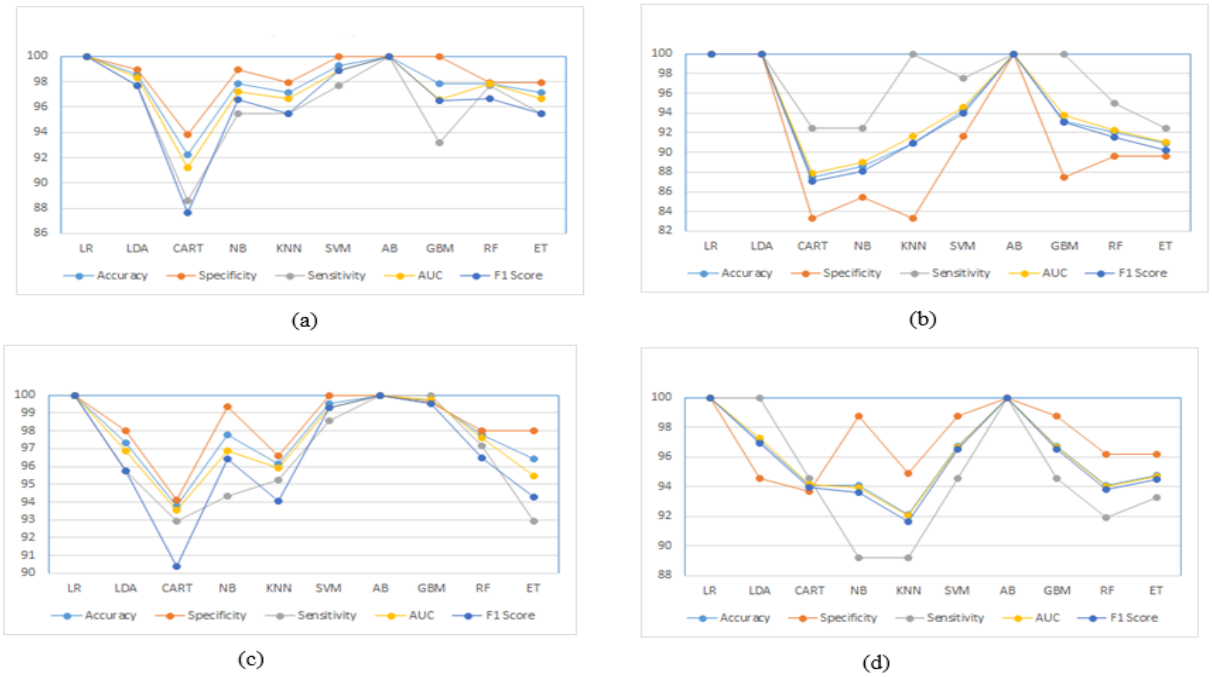


Figure. 6: The performance of ml algorithms in a) adult 2017, b) child 2017, c) adult 2018, and d) child 2018 using train test split

Table 9: Comparison ML algorithms with previous studies in child and adult dataset 2017

Dataset	Study	LR	NB	SVM	KNN
Child	[5]	97.94	92.80	-	-
	[6]	-	95.5	97.95	93.84
	[8]	98.30	94.91	98.30	88.13
	current	100.0	88.69	98.28	91.78
Adult	[5]	99.85	95.73	-	-
	[8]	96.69	96.220	98.11	95.75
	current	100.0	97.87	99.29	97.16

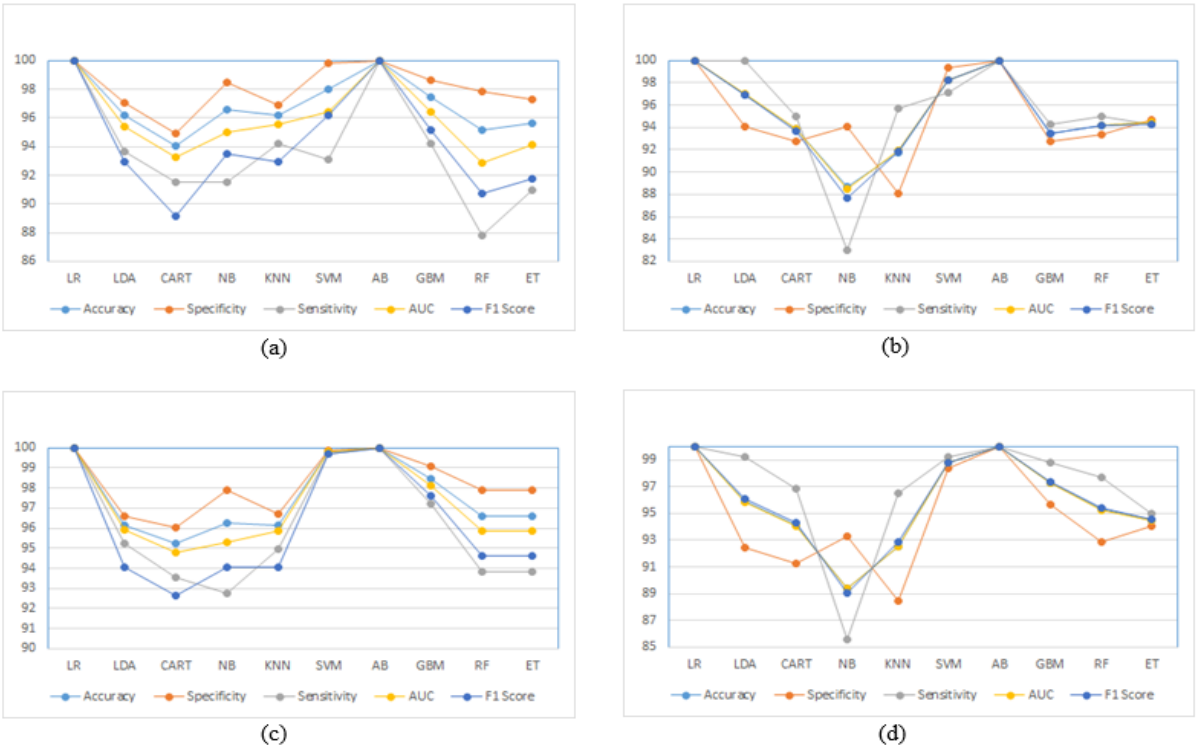


Figure. 7: The performance of ml algorithms in a) adult 2017, b) child 2017, c) adult 2018, and d) child 2018 using cross-validation

Table 10: Comparison our proposed model with previous studies

Dataset	Study	Accuracy	Specificity	Sensitivity	AUC	F1 Score	Time
Child	[5]	97.80	97.35	98.00			
	[23]	92.26	88.52	96.52			
	[7]	97.20	98.46	98.40	99.98		
	[8]	98.30	100.0	0.967			
	[6]	97.95					
	Proposed approach	100.0	100.0	100.0	100.0	100.0	0.078
Adult	[5]	99.85	99.70	99.90			
	[23]	97.10	97.11	97.07			
	[7]	98.36	96.11	99.30		98.61	
	[8]	99.53	0.9939	100.0			
	Proposed approach	100.0	100.0	100.0	100.0	100.0	0.062

TABLE 11: Comparison our proposed model on two version data set (2017 & 2018)

Dataset	Version	Accuracy	Specificity	Sensitivity	AUC	F1 Score	Time
Child	V1	100.00	100.00	100.00	100.00	100.00	0.078
	V2	100.00	100.00	100.00	100.00	100.00	0.078
Adult	V1	100.00	100.00	100.00	100	100.00	0.062
	V2	100.00	100.00	100.00	100	100.00	0.127

References

1. Baron-Cohen, Simon and Wheelwright, Sally and Skinner, Richard and Martin, Joanne and Clubley, Emma, The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians, Autism and developmental disorders. 31 (1) (2001) 5-17.
2. Allison, Carrie and Auyeung, Bonnie and Baron-Cohen, Simon, Toward brief "red flags" for autism screening: the short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls, The American Academy of Child & Adolescent Psychiatry. 51(2) (2012) 202-212.
3. Raschka, Sebastian, An overview of general performance metrics of binary classifier systems, arXiv :1410.5330, 2014.
4. Pedregosa, Fabian and Varoquaux, Gael and Gramfort, Alexandre and Michel, Vincent and Thirion, Bertrand and Grisel, Olivier and Blondel, Mathieu and Prettenhofer, Peter and Weiss, Ron and Dubourg, Vincent and others, Scikit-learn: Machine learning in Python, machine learning research. 12 (2011) 2825-2830.
5. Thabtah, Fadi, An accessible and efficient autism screening method for behavioural data and predictive analyses, Health informatics. 25 (4) (2018) 1739-1755.
6. Vaishali, R and Sasikala, R, A machine learning based approach to classify Autism with optimum behaviour sets, International Journal of Engineering & Technology. 7 (2018) 18.
7. Akter, Tania and Satu, Md Shahriare and Khan, Md Imran and Ali, Mohammad Hanif and Uddin, Shahadat and Li{\'o], Pietro and Quinn, Julian MW and Moni, Mohammad Ali, Machine Learning-Based Models for Early Stage Detection of Autism Spectrum Disorders, IEEE Access. 7 (2019) 166509-166527.
8. Raj, Suman, and Sarfaraz Masood, Analysis and Detection of Autism Spectrum Disorder Using Machine Learning Techniques, Procedia Computer Science. 167 (2020) 994-1004.
9. American Psychiatric Association [APA], 1994.
10. Kanner, Leo and others, Autistic disturbances of affective contact, Nervous child. 2 (3) (1943) 217-250.
11. Allen, Doris A, Austistic spectrum disorders: clinical presentation in preschool children, Journal of child neurology. 3 (1) (1988) 48-56.
12. Cox, A and Klein, K and Charman, T and Baird, G and Baron-Cohen, S and Swettenham, J and Drew, A and Wheelwright, S and Nightengale, N,The early diagnosis of autism spectrum disorders: use of the autism diagnostic interview--revised at 20 months and 42 months of age, J Child Psychol Psychiatry. 40 (1999) 705-718.
13. Thabtah, Fadi and Peebles, David, A new machine learning model based on induction of rules for autism detection, Health informatics journal. 26(1) (2020) 264-286.
14. Timofeev, R., Classification and regression trees (cart) theory and applications. Humboldt University, Berlin. (2004) 1-40.
15. Balakrishnama, S., Ganapathiraju, A., Linear discriminant analysis-abrief tutorial, Institute for Signal and information Processing. 18 (1998) (1998) 1-8.

16. Natekin, A., Knoll, A., Gradient boosting machines, a tutorial, Frontiers in neurorobotics. Frontiers. 7 (2013) 21.
17. Freund, Y., Schapire, R.E., A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences. 55 (1) (1997) 119-139.
18. Geurts, P., Ernst, D., Wehenkel, L., Extremely randomized trees, Machine learning. Springer. 63 (1) (2006) 3-42.
19. Webb, G.I., Keogh, E., Miikkulainen, R., Naive bayes. Encyclopedia of machine learning. 15 (2010) 713-714.
20. Breiman, Leo, Random forests. Machine learning. 45(1) (2001) 5–32.
21. Jovic, Alan and Brkic, Karla and Bogunovic, Nikola, A review of feature selection methods with applications, In: the IEEE international convention on information and communication technology, electronics and microelectronics (MIPRO), 2015, p.1200-1205.
22. ánchez-Maroño, Noelia, Amparo Alonso-Betanzos, and María Tombilla-Sanromán, Filter methods for feature selection--a comparative study, In: the springer International Conference on Intelligent Data Engineering and Automated Learning, 2007, p.178-187.
23. Omar, Kazi Shahrukh and Mondal, Prodipta and Khan, Nabila Shahnaz and Rizvi, Md Rezaul Karim and Islam, Md Nazrul, A machine learning approach to predict autism spectrum disorder, In: the IEEE International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, p.1-6.
24. Thabtah, Fadi, Autism spectrum disorder screening: machine learning adaptation and DSM-5 fulfillment, 2017, p.1-6.
25. Kleinbaum, David G and Dietz, K and Gail, M and Klein, Mitchel and Klein, Mitchell, Logistic regression, Springer. (2002).
26. Schapire, R.E., Explaining adaboost, Empirical inference. Springer. (2013) p.37-52.
27. Burges, C.J., et al., Simplified support vector decision rules, ICML, Citeseer. 96 (1996) p.71-77.
28. Kuhn, M., & Johnson, K., Nonlinear regression models. In Applied Predictive Modeling.Springer. (2013)p.141-171.
29. Schneider, Jeff, Cross Validation, https://www.cs.cmu.edu/~schneide/tut5/node42.html (Accessed on 02/22/2020).
30. Dua, Dheeru and Graff, Casey, [UCI] Machine Learning Repository, https://archive.ics.uci.edu, University of California, Irvine, School of Information and Computer Sciences, 2017.
31. Thabtah, Fadi, Autism datasets. http://fadifayez.com/autism-datasets/ (Accessed July 2019).
32. Thabtah, Fadi, Asd tests Mobile application. http://asdtests.com/, 2017.
33. ABIDE Preprocessed, http://preprocessed-connectomes-project.org/abide/ (Accessed on 2019).
34. SLAITS - National Survey of Childrens Health, https://www.cdc.gov/nchs/slaits/nsch.htm (Accessed on 2019).