Predicting Autism Spectrum disorder by Machine learning from Healthcare Communities

Srividhya Ganesan[1], Dr. Raju[2]
[1]*Associate Professor, Computer Science and Engineering Department, AMC Engineering College, Bangalore, and PhD Scholar in SRET, SRIHER, Chennai,* srividhyaganesan@amceducation.in ;
[2]*Professor, SRIHER, Chennai,* raju.sret@gmail.com

ABSTRACT

Autism Spectrum Disorder is a syndromic disorder related to neurological elements resulting to the problems in communication, interacting socially, behavioral, and sensory. According to World Health Organization, the count of patients detected with Autism Spectrum Disorder is slowly increasing. Recent studies concentrate on data collection, Clinical analysis, and brain image laboratory analysis. They do not concentrate much on diagnosing Autism based on Artificial Intelligence and Machine learning.

Goal: This paper mainly intends to classify and categorize Autism data to give an understandable, rapid and simple means to help early intervention of Autism Spectrum Disorder.

***Methods*:** Three groups of Autism Spectrum Disorder datasets are taken for Child, adolescence, and adults. We applied k-Nearest Neighboring, Support vector Machine and Random Forest algorithm to classify the Autism Spectrum Disorder data. During our experimentations, the data was split at random into training sets and test sets. The sections of data were picked at random to assess the classification algorithms.

***Outcomes*:** The outcome and results were evaluated by average values. This is proven that Support Vector Machine and Random forest are efficient algorithms for Autism Spectrum Disorder classification. In specific, Random forest algorithm classified with 100% accuracy for all datasets.

***Conclusion*:** It is been observed that early intervention is possible absolutely. The accuracy of diagnosing Autism Spectrum Disorder will be higher if the data samples count is huge. The results show that Support Vector Machine and Random Forest algorithms gives good classification score compared to k-Nearest Neighboring algorithm w.r.t accuracy, F-measure, sensitivity, and Area Under Curve. We found that Random Forest algorithm is efficient and effective compared to Support Vector Machine and k-Nearest Neighbouring algorithm for data classification.

***Keywords*-** Machine learning, Autism, supervised learning, Support Vector machine, k-nearest neighboring, Random Forest

Article Received: 16th October, 2020; Article Revised: 30th December, 2020; Article Accepted: 08th January, 2021

1. INTRODUCTION

According to World health organization (WHO) statistical data, 63 percent of children are detected with Autism. Autism spectrum disorder (ASD) appears in children, adolescences, and adults. ASD is a neurodevelopment disorder with high health care expenses. Individuals with ASD have a problem in communication, social interaction, behavioral issues. It is exceedingly difficult for them to think and imagine, interact with other children, communication issues with other people.

Early intervention will improve the standard of life of individuals with Autism and plays a vital role in clinical diagnosis. The process can take long time to diagnosis ASD with expensive testing methods. In recent times, ASD cases are increasing rapidly across the globe. It is the motivation for scientist or doctors to invent more efficient screening methods.

Massive amount of data can be stored with an advancement in innovative technology. Data mining which is associated to machine learning plays an important role to take decisions based on the data collected. Machine learning is acquiring much importance on medical and biomedical field. Machine learning methodologies are mainly applied to help data interpretation in clinical decision-making and diagnostics. Hence, techniques of screening disorders along with the help of Machine learning are widely analyzed.

By the current scenario, there are many findings on Autism Spectrum Disorder. Few research targets on clinical trials, few focuses on exploring and analysing brain image data's, and few other focuses on narrow scope of every country. Hence it is vital to give a rapid, simple method to help early intervention of ASD which will also be helpful for descendants of ASD individual to find specialists and experts for the purpose of treatment.

In this paper, we aim on the classification of ASD datasets of Unique client identifier database (UCI). The datasets of ASD stand huge datasets associated to clinical diagnosis of different age groups. The information's were gathered from several nations by the reviews on mobile app called "ASD Test" which could be obtained in the URL "http://www.asdtests.com/". The app is open for iOS and Android systems. This was designed by Dr. Fayez from NMIT. But the data collected is inadequate and hence it is not reliable and consistent enough to make up clinical determinations directly. Three Autism datasets of UCI database are considered in this study. The datasets are classified by the age 4-11 years as Children, 12-17 years as Adolescence and greater than 18 years as Adult. The datasets include 20 characteristics which is used for further analysis, mainly for predicting ASD and improving the performance parameters such as efficiency, accuracy, and precision of ASD classification.

Our Research focuses on Preprocessing and classification of ASD datasets and the datasets were gathered from surveys. The reviews or surveys contain queries on private information. The surveys also include screening queries associated to ASD. The data gathered will be converted into numerical data to enable processing. But some of the values might me missing and hence it is better to use an approach to fill up the missing data. For that, we need to implement classification methods. Depending upon the results after implementing classification methods, the medical doctor or Scientist diagnose ASD easily, precisely, and rapidly.

The remaining paper is organized as follows. Segment 2 represents Methods and Materials of the proposed system. Segment 3 illustrates the experimental results. The last segments represent the discussion and conclusion.

2. METHODS AND MATERIALS

In this Research, three different ASD datasets were used. They are cleaved into three distinct groups: Autism quotient (Autism screening Adult data set, Autism screening Children data set, Autism screening Adolescence Dataset) are shown in the Table 1. The datasets contain 20 attributes which is used for training activity. The attribute "Class" is mainly for storing results of ASD (i.e., ground truth). It is for estimating accuracy, Precision, F- measure scores, sensitivity, and Area under curve (AUC). Ground truth consists of bi digits (0,1) i.e YES or NO.

Name	All			After Exiting the Missing Data		
	Samples' Number	Attributes	Class	Samples' Number	Attributes	Class
Autism Screening Adult Data Set	704	20	2	609	20	2
Autistic Spectrum Disorder Screening Data for Children Data Set	292	20	2	249	20	2
Autistic Spectrum Disorder Screening Data for Adolescence Data Set	104	20	2	98	20	2

Table 1. Features and their descriptions.

The attributes of the datasets are illustrated in the Table 2. First Ten features are of personal information and the next ten features are of screening questions.

Table 3 describes the preprocessing data by Numeric transformation rule. Numeric transformation rule was applied for attributes such as Gender, Ethnicity, "Born with Jaundice", "Family member with Pervasive Developmental Disorders (PDD)", "Country of Residence", "Used the Screening app Before" and "Who is completing the test". The rule has not been applied for the traits of screening questions as the values were set to binary numbers 0 and 1. For "gender" attribute, numbers 1 and 0 for male and female are used correspondingly. For "ethnicity" attribute, numbers 1 to 12 are used for each value. For attributes "family member with PDD", "Born with Jaundice", and "used the screening application before", numbers 1 and 0 for the values "Yes" and "No" are used correspondingly. For "country of residence", numbers from 1 to 85 were set because only 85 countries data were collected. Lastly, for "who is completing the test" attribute, the numerals from 1 to 5 were set to every single value. Ten data records from AQ10 dataset Adult are shown in Table 4. This signifies before and after preprocessing by Numeric transformation rule.

Most important Machine learning algorithms were used to classify the data sets. They are K-Nearest Neighbouring (KNN) algorithm, Support Vector Machine (SVM) Algorithm and Random Forest (RF) algorithm.

K-Nearest Neighbouring Algorithm:

K- Nearest Neighbouring Algorithm (kNN) is an important machine learning algorithm for classification which is used in areas like data mining, pattern

recognition and further areas related to science. The position of the uncategorized data is decided by analysing the proximity of K of the categorized data. There are certain distances which could be applied to define the Nearest Neighbour in kNN algorithm, like…

Euclidean distance:

$$d(x,y) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2},$$

Manhattan distance:

$$d(x,y) = \sum_{i=1}^{k}|x_i - y_i|,$$

Minkowski distance:

$$d(x,y) = \left(\sum_{i=1}^{k}(x_i - y_i)^q\right)^{1/q}, q = 1,2,\dots$$

No.	Attribute Name	Data Type	Description
1	Question 1 Answer	Binary (0, 1)	I often notice small sounds when others do not
2	Question 2 Answer	Binary (0, 1)	I usually concentrate more on the whole picture, rather than the small details
3	Question 3 Answer	Binary (0, 1)	I find it easy to do more than one thing at once
4	Question 4 Answer	Binary (0, 1)	If there is an interruption, I can switch back to what I was doing very quickly
5	Question 5 Answer	Binary (0, 1)	I find it easy to read between the lines when someone is talking to me
6	Question 6 Answer	Binary (0, 1)	I know how to tell if someone listening to me is getting bored
7	Question 7 Answer	Binary (0, 1)	When I'm reading a story, I find it difficult to work out the character's intentions
8	Question 8 Answer	Binary (0, 1)	I like to collect information about categories of things (*e.g.* types of cars, types of bird, types of train, types of plant, *etc*)
9	Question 9 Answer	Binary (0, 1)	I find it easy to work out what someone is thinking or feeling just by looking at their face
10	Question 10 Answer	Binary (0, 1)	I find it difficult to work out people's intentions
11	Age	Number	Age in years
12	Gender	String	Male or Female
13	Ethnicity	String	List of common ethnicities in text format
14	Born with jaundice	Binary (no=0, yes=1)	Whether the case was born with jaundice
15	Family member with PDD	Binary (no=0, yes=1)	Whether any immediate family member has a PDD
16	Who is completing the test	String	Parent, self, caregiver, medical staff, clinician, *etc*.
17	Country of residence	String	List of countries in text format
18	Used the screening app before	Binary (no=0, yes=1)	Whether the user has used a screening app
19	Screening Method Type	Integer (0,1,2,3)	The type of screening methods chosen based on age category (0=toddler, 1=child, 2= adolescence, 3= adult)
20	Screening Score	Integer	The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner
21	Class Name	String	Ground truth

Table 2. Features description of the datasets

Gender		Ethnicity		Born with Jaundice		Family Member with PDD		Country of Residence		Used the Screening app Before		Who is Completing the Test	
String	Value	String	Value	String	Value	String	Value	String	Value	String	Value	String	Value
f	0	Asian	1	No	0	No	0	Afghanistan	1	No	0	Health care professional	1
m	1	Black	2	Yes	1	Yes	1	Albania	2	Yes	1	Others	2
-	-	Eastern	3	-	-	-	-	American Samoa	3	-	-	Parent	3
-	-	European	4	-	-	-	-	...	-	-	-	Relative	4
-	-	Hispanic	5	-	-	-	-	...	-	-	-	Self	5
-	-	Latino	6	-	-	-	-	...	-	-	-	?	6
-	-	Middle	7	-	-	-	-	...	-	-	-	-	-
-	-	Middle Eastern	8	-	-	-	-	...	-	-	-	-	-
-	-	Others	9	-	-	-	-	...	-	-	-	-	-
-	-	Pasifika	10	-	-	-	-	...	-	-	-	-	-
-	-	South Asian	11	-	-	-	-	...	-	-	-	-	-
-	-	Turkish	12	-	-	-	-	...	-	-	-	-	-

Table 3 The numeric transformation rule for pre-processing data.

Real Data	1	1	1	1	0	1	1	1	1	0	17	m	Asian	yes	yes	Bahamas	no	8	18	Health care professional	YES
	1	1	1	1	1	1	1	1	1	1	33	m	White-European	no	no	United States	no	10	18	Relative	YES
	0	1	0	1	1	1	1	0	0	1	18	f	Middle Eastern	no	no	Burundi	no	6	18	Parent	NO
	0	1	1	1	1	1	0	0	1	0	17	f	-	no	no	Bahamas	no	6	18	-	NO
	1	0	0	0	0	0	1	1	0	1	17	m	-	no	no	Austria	no	4	18	-	NO
	1	0	0	0	0	0	1	1	0	1	17	f	-	no	no	Argentina	no	4	18	-	NO
	1	1	0	1	1	0	0	1	0	1	18	m	Middle Eastern	no	yes	New Zealand	no	6	18	Parent	NO
	1	0	0	0	0	0	1	1	1	1	31	m	Middle Eastern	no	no	Jordan	no	5	18	Self	NO
	0	0	0	0	0	0	0	1	0	1	30	m	White-European	no	no	Ireland	no	2	18	Self	NO
	0	0	1	0	1	1	0	0	0	0	35	f	Middle Eastern	no	yes	United Arab Emirates	no	3	18	Self	NO
After pre-processing	1	1	1	1	0	1	1	1	1	0	17	1	1	1	1	12	0	8	18	1	1
	1	1	1	1	1	1	1	1	1	1	33	1	14	0	0	87	0	10	18	4	1
	0	1	0	1	1	1	1	0	0	1	18	0	8	0	0	20	0	6	18	3	0
	0	1	1	1	1	1	0	0	1	0	17	0	15	0	0	12	0	6	18	6	0
	1	0	0	0	0	0	1	1	0	1	17	1	15	0	0	10	0	4	18	6	0
	1	0	0	0	0	0	1	1	0	1	17	0	15	0	0	6	0	4	18	6	0
	1	1	0	1	1	0	0	1	0	1	18	1	8	0	1	60	0	6	18	3	0
	1	0	0	0	0	0	1	1	1	1	31	1	8	0	0	49	0	5	18	5	0
	0	0	0	0	0	0	0	1	0	1	30	1	14	0	0	45	0	2	18	5	0
	0	0	1	0	1	1	0	0	0	0	35	0	8	0	1	85	0	3	18	5	0

Table 4 An example about the data before/after pre-processing.

Support Vector Machine:

SVM is one of the important Machine learning (ML) algorithm for classification as well regression problems. SVM is a rapid classification algorithm with great precision and accuracy. The main objective of SVM is to identify a hyperplane in N dimensional space which classifies the samples, so that the hyperplane has maximal margin. In other words, the distance or the interval between the samples of both the classes has to be maximal.

The optimisation problem linked with SVM algorithm produces the resulting structure

$$\min_{w,b,\xi}\left\{\frac{1}{2}w^T w + C\sum_{i=1}^{l}\xi_i\right\}, \quad \text{constrains to}$$

$$y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0,$$

where $(x_i, y_i), i = 1, \ldots, l$ is an instance-label pair, $x_i \in \mathbb{R}^n, y \in \{1, -1\}^l, \phi$ is a mapping function, $C > 0$ is a penalty parameter. The function $K(x_i, x_j) \equiv \phi(x_i)^T\phi(x_j)$ is called the kernel function.

There are many structures and forms of Kernel function:

Linear function: $K(x_i, x_j) = x_i^T x_j,$

Polynomial function: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0,$

Radial basis function: $K(x_i, x_j) = \exp\left(-\gamma\|x_i - x_j\|^2\right), \gamma > 0,$

Sigmoid function: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r).$

Random Forests:

Random forest (RF) Algorithm classifies and organise the data samples depending on several classification trees. It is highly efficient on large databases. Every tree provides classification result.

RF is a group of decision trees of the structure:

$$h(x, \theta_k), \text{for } k = 1, .., K,$$

Where θ_k represents identical and independent distributed random attributes. Each of K trees provide the figure of class of sample X is attained by majority.

If the tree count is sufficient, the common error is

$$e \leq \frac{c(1 - s^2)}{s^2},$$

Where e is the common error, s is the metrical strength of trees and c stands for correlation amongst the trees.

Fig. (1). shows the flow chart of ASD detection method. From the flowchart, it is the seen clearly that the data is split into two sections after pre-processing the data. The first section is selecting α% differing in the scale of 50% till 90% and remaining section with (100- α) %. The first section refers to training data which is mainly used to train the classification model with RF, SVM and kNN algorithms. The remaining section refers to the testing data which is mainly used to test and evaluate the success ratio by certain performance parameters such as sensitivity, specificity, accuracy etc of classification algorithms. The "class name" is used to evaluate the results.

3. RESULTS

This research executes the prediction method of ASD on Jupyter-notebook. The numeral of nearest neighbouring is set to 3 and applied the Euclidean distance in kNN algorithm. The Radial Basis functions (RBF) kernel is a common kernel function which is used in SVM. The number of trees is 60 for Random forest algorithm.

The following parameters (sensitivity, area under curve, accuracy, F-measure) are applied to evaluate the performance of classification methods. They are specified as follows:

$$accuracy := \frac{TP + TN}{TP + FP + FN + TN},$$

$$sensitivity := \frac{TP}{TP + FN},$$

$$F - measure := \frac{2 * TP}{(2 * TP + FP + FN)},$$

$$AUC := \frac{(1 + \frac{TP}{TP + FN} - \frac{FP}{FP + TN})}{2},$$

Where TP is True Positive, FP is False Positive, TN is True Negative, FN is False Negative. Different α values

were used in our experimentations. Table 5 indicates the experiment names with different α values.

For each experimentation, the testing and training data were selected randomly 100 times. The outcome of sensitivity, area under curve, accuracy, F-measure were the average values of every equivalent score of all hundred instances.

Experiments	Training Data	Testing Data
Experiment 1 (α= 50)	50%	50%
Experiment 2 (α= 60)	60%	40%
Experiment 3 (α= 70)	70%	30%
Experiment 4 (α= 80)	80%	20%
Experiment 5 (α= 90)	90%	10%

Table 5: The experiments with different α values

3.1 ASD screening for AQ10 Adult Dataset

Table 6 indicates the results of classification for the case of complete data. As seen from table 6, SVM and RF methods achieved the score of 100% but kNN algorithm classifies with least scores.

Tests	Methods	Accuracy	Sensitivity	F-measure	AUC
Experiment 1	kNN	94.75 ± 1.10	93.69 ± 1.57	93.36 ± 1.38	0.94±0.02
	RF	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00
	SVM	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00
Experiment 2	kNN	95.10 ± 1.17	94.12 ± 1.69	93.81 ± 1.48	0.93±0.01
	RF	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00
	SVM	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00
Experiment 3	kNN	95.35 ± 1.32	94.55 ± 1.92	94.15 ± 1.68	0.95±0.01
	RF	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00
	SVM	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00
Experiment 4	kNN	95.93 ± 1.77	95.40 ± 2.18	94.90 ± 2.19	0.94±0.02
	RF	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00
	SVM	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00
Experiment 5	kNN	95.70 ± 2.63	95.15 ± 3.22	94.64 ± 3.25	0.96±0.03
	RF	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00
	SVM	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00

Table 6 Classification results for the AQ-10-Adult dataset for the case of complete data.

3.2 ASD screening for AQ10 Adolescence Dataset

Table 7 indicates the results of classification for the case of complete data. As seen from Table 7, RF method attained 100% score for all experiments. SVM algorithm achieved 100% score except Experiment 4 and 5. But kNN algorithms classifies with lowest score.

Tests	Methods	Accuracy	Sensitivity	F-measure	AUC
Experiment 1	kNN	88.20 ± 2.34	88.41 ± 2.29	88.18 ± 2.36	0.87±0.02
	RF	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00
	SVM	99.88 ± 0.61	99.88 ± 0.59	99.88 ± 0.61	0.97±0.03
Experiment 2	kNN	88.22 ± 2.97	88.40 ± 2.93	88.19 ± 2.99	0.88±0.05
	RF	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00
	SVM	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00
Experiment 3	kNN	88.13 ± 3.12	88.31 ± 3.08	88.10 ± 3.14	0.88±0.06
	RF	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00
	SVM	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00
Experiment 4	kNN	89.43 ± 3.55	89.61 ±3.50	89.41 ± 3.57	0.89±0.05
	RF	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00
	SVM	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00
Experiment 5	kNN	88.17 ± 5.03	88.31 ± 4.98	88.13 ± 5.05	0.88±0.07
	RF	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00
	SVM	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00

Table 7 Classification results for the AQ-10-Adolescence dataset for the case of complete data.

3.3. ASD screening for AQ-10 child Dataset

Table 8 indicates the results of classification for the case of complete data. As seen from Table 8, RF method attained more than 99% score for all cases. SVM algorithm achieved 90% to 94% score for all cases whereas kNN algorithm classifies lowest score compared to SVM and RF methods.

Tests	Methods	Accuracy	Sensitivity	F-measure	AUC
Experiment 1	kNN	85.78 ± 3.33	82.89 ± 4.22	84.11 ± 4.02	0.85±0.02
	RF	99.92 ± 0.39	99.91 ± 0.46	99.92 ± 0.41	0.99±0.01
	SVM	89.16 ± 4.77	88.81 ± 5.23	88.60 ± 5.12	0.89±0.03
Experiment 2	kNN	86.83 ± 4.51	83.96 ± 5.44	85.22 ± 5.41	0.88±0.02
	RF	99.98 ± 0.24	99.97 ± 0.31	99.97 ± 0.26	0.99±0.01
	SVM	90.98 ± 4.64	90.68 ± 4.96	90.51 ± 4.93	0.93±0.02
Experiment 3	kNN	85.77 ± 4.70	82.79 ± 5.50	84.02 ± 5.53	0.85±0.02
	RF	99.94 ± 0.45	99.93 ± 0.49	99.93 ± 0.48	0.99±0.01
	SVM	91.77 ± 5.37	91.60 ± 5.37	91.37 ± 5.59	0.91±0.02
Experiment 4	kNN	87.29 ± 6.56	84.51 ± 8.06	85.53 ± 7.90	0.87±0.03
	RF	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00
	SVM	92.14 ± 5.04	91.56 ± 5.73	91.57 ± 5.47	0.92±0.01
Experiment 5	kNN	86.80 ± 10.04	84.29 ± 11.81	84.83 ± 12.09	0.86±0.03
	RF	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	1.00±0.00
	SVM	92.30 ± 7.90	92.00 ± 8.49	91.80 ± 8.51	0.92±0.01

Table 8 Classification results for the AQ-10-Child dataset for the case of complete data.

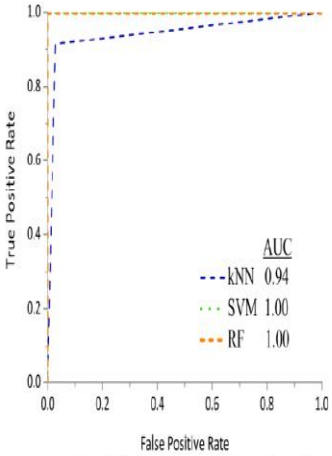
4. DISCUSSION

As seen from experiments, it is clear that AQ10 - Adult, Adolescence and child were categorized with a precision and accuracy of 100%. The data were split into training dataset and testing dataset with different α values differing from 40% to 90%. For each α case, the testing and training data were chosen at random. The results were shown as average scores of hundred subgroups of data. These scores were standard w.r.t performance parameters like accuracy, precision, sensitivity, AUC and F-measure.

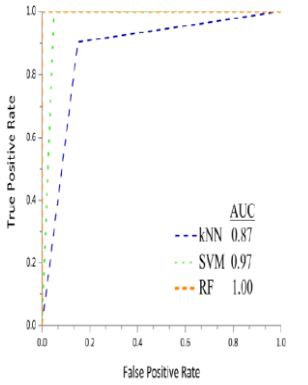
From Table 6(AQ10-Adult dataset), the SVM and RF algorithms were classified with a score of 100% for entire test cases. From Table 7(AQ10-Adolescence dataset), the RF algorithm were classified with a score of 100% and SVM algorithm with α>=60% also were classified with a score of 100%. From Table 8(AQ10-child dataset), the RF algorithm were classified with score more than 99% of all cases. With α>=80%, RF algorithm achieved 100% score and SVM algorithm achieved with the score of more than 90% in all cases.

For AQ10-child dataset, only 102 data samples were there. The number of data samples are very less. If it is of more data samples, then the scores would be higher and accurate for all the methods. From all above three datasets, kNN algorithm gave less scores compared to RF and SVM algorithm. RF algorithm attained best classification outcome and attained 100% accuracy, F-measure scores and sensitivity in most of the test cases.

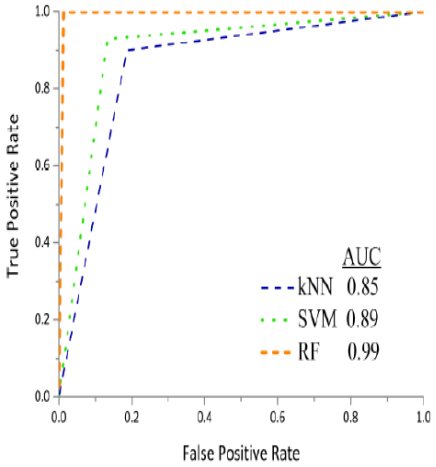
As a conclusion, we shall affirm that RF algorithm is a good classification algorithm for clinical decision support system. In other words, we shall predict ASD with high accuracy using RF classification algorithm.



AQ-10-Adult for the case of complete data



AQ-10-Adolescence for the case of complete data



AQ-10-Child for the case of complete data

CONCLUSION

In this paper, we applied various methods to predict Autism using machine learning algorithms viz SVM, KNN and RF. The tests were carried out on 3 ASD datasets- AQ10 Adult, AQ10 Adolescence and AQ10 child of Unique client identifier database. The obtained results show that SVM and RF algorithms gives good

classification score compared to kNN algorithm w.r.t accuracy, F-measure, sensitivity, and AUC. We noticed that RF algorithm is more efficient than kNN and SVM algorithm for data classification.

It is been observed that early intervention is possible absolutely. The accuracy of diagnosing ASD will be higher if the data samples count is large. The precision mainly relies on completion level of data collected. The accuracy of early intervention will be high if the data is complete.

At last, we could predict ASD fast, accurate and easily by using RF algorithm. Hence effective treatment for Autism can result in improved quality in the lives of the patients with ASD together with their families.

Data availability

The information endorsing the findings of the paper is accessible in UCI ML Depository at https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult.

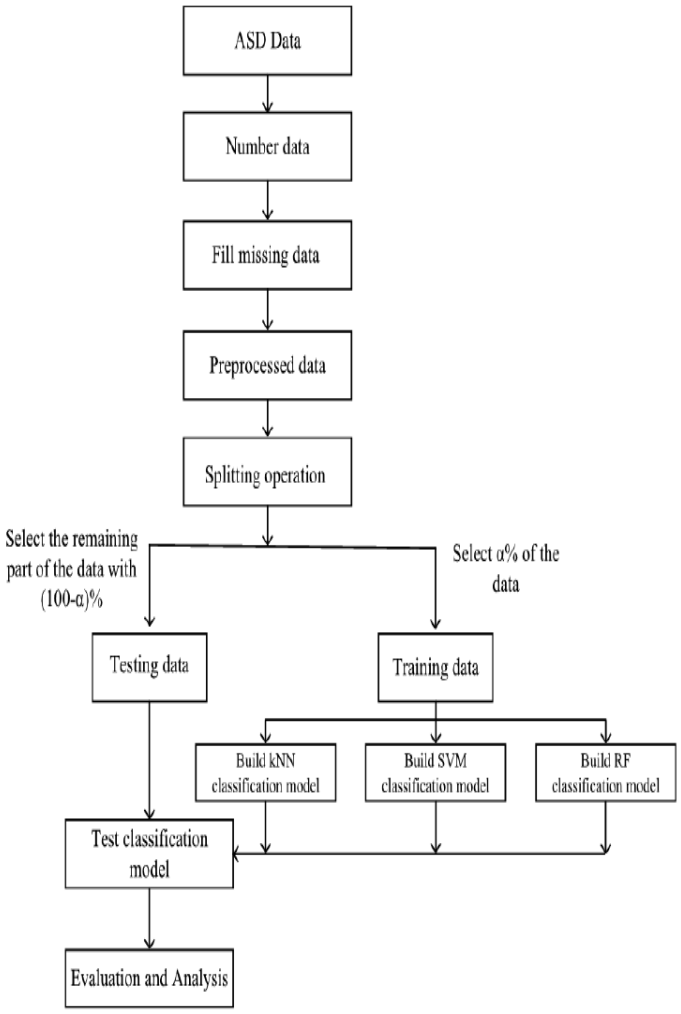


Fig. (1). The flowchart of the ASD detection, $\alpha \in \{50\%, 60\%, 70\%, 80\%, 90\%\}$.

CONFLICT OF INTEREST

The authors affirm no Conflict of Interest, commercial or otherwise.

ACKNOWLEDGEMENTS

We express gratitude to many people who created and making the UCI ML Repository a success. We could share the datasets under a free license. We would also like to thank all the Professors from Sri Ramachandra Institute of Higher Education and Research, Chennai for their continuous encouragement, guidance, and support.

REFERENCES

[1] World Health Organization. Autism spectrum disorders [Internet]. 2017 [cited 2018 Dec 4]. Available from: http://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders
[2] Hlavatá P, Kašpárek T, Linhartová P, *et al.* Autism, impulsivity and inhibition a review of the literature. Basal Ganglia 2018; 14: 44-53. [http://dx.doi.org/10.1016/j.baga.2018.10.002]
[3] McDermott JH, Study DDD, Clayton-Smith J, Briggs TA. The TBR1-related autistic-spectrum-disorder phenotype and its clinical spectrum. Eur J Med Genet 2018; 61(5): 253-6. [http://dx.doi.org/10.1016/j.ejmg.2017.12.009] [PMID: 29288087]
[4] Saresella M, Piancone F, Marventano I, *et al.* Multiple inflammasome complexes are activated in autistic spectrum disorders. Brain Behav Immun 2016; 57: 125-33. [http://dx.doi.org/10.1016/j.bbi.2016.03.009] [PMID: 26979869]
[5] Heinsfeld AS, Franco AR, Craddock RC, Buchweitz A, Meneguzzi F. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. Neuroimage Clin 2017; 17: 16-23. [http://dx.doi.org/10.1016/j.nicl.2017.08.017] [PMID: 29034163]
[6] Thabtah F, Kamalov F, Rajab K. A new computational intelligence approach to detect autistic features for autism screening. Int J Med Inform 2018; 117: 112-24. [http://dx.doi.org/10.1016/j.ijmedinf.2018.06.009] [PMID: 30032959]
[7] Azer SA, Bokhari RA, AlSaleh GS, *et al.* Experience of parents of children with autism on YouTube: are there educationally useful videos?. Informatics Heal Soc Care 2018; 43(3): 219-33.
8] Franz L, Adewumi K, Chambers N, Viljoen M, Baumgartner JN, de Vries PJ. Providing early detection and early intervention for
autism spectrum disorder in South Africa: stakeholder perspectives from the Western Cape province. J Child Adolesc Ment Health
2018; 30(3): 149-65. [http://dx.doi.org/10.2989/17280583.2018.1525386] [PMID: 30403918]
[9] Pagnozzi AM, Conti E, Calderoni S, Fripp J, Rose SE. A systematic review of structural MRI biomarkers in autism spectrum disorder: a machine learning perspective. Int J Dev Neurosci 2018; 71: 68-82. [http://dx.doi.org/10.1016/j.ijdevneu.2018.08.010] [PMID: 30172895]
[10] Casanova R, Barnard RT, Gaussoin SA, *et al.* Using high dimensional machine learning methods to estimate an anatomical
risk factor for Alzheimer's disease across imaging databases. Neuroimage 2018; 183: 401-11. [http://dx.doi.org/10.1016/j.neuroimage.2018.08.040] [PMID: 30130645]
[11] Parisi L. RaviChandran N, Manaog ML. Feature-driven machine learning to improve early diagnosis of Parkinson's disease. Expert Syst Appl 2018; 110: 182-90. [http://dx.doi.org/10.1016/j.eswa.2018.06.003]
[12] Khamparia A, Saini G, Pandey B, *et al.* KDSAE: Chronic kidney disease classification with multimedia data learning using deep stacked autoencoder network. Multimed. Tools Appl 2019.
[13] Priyadarshini R, Barik RK, Panigrahi C, *et al.* An Investigation Into the Efficacy of Deep Learning Tools for Big Data Analysis in Health Care. Int J Grid High Perform Comput 2018; 10(3): 1-13. [http://dx.doi.org/10.4018/IJGHPC.2018070101]
[14] Yonekura A, Kawanaka H, Surya Prasath VB, *et al.*
[15] Escudero J, Ifeachor E, Zajicek JP, Green C, Shearer J, Pearson S. Machine learning-based method for personalized and cost-effective detection of Alzheimer's disease. IEEE Trans Biomed Eng 2013; 60(1): 164-8. [http://dx.doi.org/10.1109/TBME.2012.2212278] [PMID: 22893371]
[16] Liu S, Liu S, Cai W, *et al.* Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. IEEE Trans Biomed Eng 2015; 62(4): 1132-40. [http://dx.doi.org/10.1109/TBME.2014.2372011] [PMID: 25423647]
[17] Abdar M, Zomorodi-Moghadam M, Zhou X, *et al.* A new nested ensemble technique for automated diagnosis of breast cancer. Pattern Recognit Lett [Internet] 2018. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0167865518308766 [http://dx.doi.org/10.1016/j.patrec.2018.11.004]
[18] Adem K, Kiliçarslan S, Cömert O. Classification and diagnosis of cervical cancer with softmax classification with stacked autoencoder. Expert Syst Appl 2019; 115: 557-64. [Internet]. [http://dx.doi.org/10.1016/j.eswa.2018.08.050]
[19] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Stat 1992; 43(3): 175-85.
[20] Dua D, Graff C. UCI Machine Learning Repository. Available at: http://archive.ics.uci.edu/mlIrvine.
[21] Thabtah F. Autism Spectrum Disorder screening: Machine learning adaptation and DSM-5 fulfillment. ICMHI '17 Proceedings of the 1st International Conference on Medical and Health Informatics 2017; Taichung City, Taiwan May 20-22.
[22] Thabtah F. ASDTests. A mobile app for ASD screening. [Internet]. 2017 [cited 2018 Dec 20]. Available from: ww.asdtests.com
[23] Thabtah F. Machine learning in autistic spectrum disorder behavioral research: A review and ways forward. Informatics Heal. Soc. Care 2019; 44(3): 278-97.

[24] Cortes C, Vapnik V. Support-Vector Networks. Mach Learn 1995; 20: 273-97. [http://dx.doi.org/10.1007/BF00994018]

[25] Breiman L. Random forests. Mach Learn 2001; 45: 5-32. [http://dx.doi.org/10.1023/A:1010933404324]