A Clustering Approach for Autism based Autistic Trait Classification

Said Baadel[1,2 *], Fadi Thabtah,[3] Joan Lu[1]

1. Faculty of Engineering and Computing Science, University of Huddersfield, Huddersfield, UK.

2. Faculty of Communication, Arts and Sciences, Canadian University Dubai, Dubai, UAE

3. Dept of Digital Technologies, Manukau Institute of Technology, Manukau, New Zealand

** s.baadel@gmail.com*

A Clustering Approach for Autism based Autistic Trait Classification

Machine learning (ML) techniques can be utilized by physicians, clinicians, as well as other users, to discover Autism Spectrum Disorder (ASD) symptoms based on historical cases and controls to enhance autism screening efficiency and accuracy. The aim of this study is to improve the performance of detecting ASD traits by reducing data dimensionality and eliminating redundancy in the autism dataset. To achieve this, a new semi-supervised ML framework approach called Clustering-based Autistic Trait Classification (CATC) is proposed that uses a clustering technique and validation of the classifiers is done by classification techniques. The proposed method identifies potential autism cases based on their similarity traits as opposed to a scoring function used by many ASD screening tools. Empirical results on different datasets involving children, adolescents, and adults were verified and compared to other common machine learning classification techniques. The results showed that CATC offers classifiers with higher predictive accuracy, sensitivity, and specificity rates than those of other intelligent classification approaches such as Artificial Neural Network (ANN), Random Forest, and Random Trees, and Rule Induction. These classifiers are useful as they are exploited by diagnosticians and other stakeholders involved in ASD screening.

Keywords: Autism Diagnosis; Classification; Clustering; Machine Learning; OMCOKE; Predictive Models

1: Introduction

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that contributes to the delay of social and communication behaviors of individuals.[8,10.] Typically, ASD diagnosis is done by clinicians in a clinical set up using observable behavioral indicators in a process referred to as clinical judgment (CJ). [37, 45.] The official diagnosis process of ASD involves multiple examinations, which in turn cause the waiting time for patients to be lengthy[40]. For instance, the waiting time for an ASD diagnosis in the UK averages over 3 years[16]. Therefore, it is vital that the administration time needed for both screening and diagnosis be reduced to cater for the growing number of ASD patients.[25, 27]

Autism screening is a fundamental step that addresses whether individuals exhibit

potential autistic traits related to communication, social or repeated behaviour. [1] This step is crucial as the individual and the concerned family become aware of the possibility of ASD traits early and hence can search for the needed formal assessments. There are many ASD screening tools developed by researchers such as Autism Spectrum Quotient (AQ) and Childhood Autism Rating Scale (CARS). [6, 7, 24, 38] Most of these screening methods have been developed using existing clinical autism diagnosis methods and are represented as questionnaires in which each question is associated with a few possible answers in a multiple-choice fashion. The questionnaires used contain measurable indicators (variables/questions) that address communication, behavior and social skills, of individuals. For example, the Child Behavior Checklist (CBCL) screening method contains more than 100 questions,[2] and the AQ method contains 50 questions [7]. These make the process of screening lengthy besides inaccessible as most existing screening methods normally do not exist in simply accessible platforms such as mobile. [40, 41]

Most of the existing autism screening methods utilize scoring functions that compute a final score based on the answers given by users undergoing the screening (caregivers, parents, medical staff, teachers or even the adult patients). To be specific, the screening methods take the answers given in the questionnaire as an input for the scoring function, which in turn processes the input and computes a final score to reflect whether the individual is associated with ASD traits. For instance, in AQ method, a cut-off score of larger than 32 is an indication of autistic traits.[4, 7] Therefore, the final decision of having ASD traits lay solely on the score calculated by the function. This function in most cases just sums up the behavioural indicators' answers and does not attempt to seek for correlations among these indicators and the target class (ASD traits).

To address these shortcomings, there is a need for intelligent methods that can replace the scoring function and improve the efficiency of the screening. Since ASD screening involves

forecasting whether individuals have the possibility of ASD traits based on a predefined characterized variable then this issue be a predictive analysis problem in ML. The screening of ASD traits can be considered a classification problem in which historical data that have been already classified with and without ASD traits is utilized as an input to construct a classification system. This system is then used to guess whether a new individual exhibits any autistic traits. ML can be utilized for ASD screening to improve the classification of the screening and to reduce the process of the screening time. More importantly, ML may provide models that can contain useful information about ASD traits to the diagnosticians especially the correlation among behavioral indicators and how they relate to ASD screening. ML techniques use artificial intelligence and statistics to create intelligent models by discovering hidden patterns in data, so users can improve decisions. [41]

There have been recent attempts to adopt ML techniques in autism screening and diagnosis, i.e. [1, 9, 11, 15, 25, 37, 40]. These studies focused primarily on improving time, accuracy, and reducing the dimensionality of the dataset by pinpointing influential autistic symptoms. Thabtah et al.,[41] proposed a new feature selection method called Variable Analysis (Va) to determine the most influential features related to ASD based on datasets related to adults, adolescents, and children. The authors were able to minimize the number of features to 5-7 based on predictive analysis and filter methods. Abbas et al., [1] used Random Forest to improve the diagnosis process of autism and Levy et al., [25] compared 17 different classification-based ML algorithms to seek improvements on the diagnosis performance of autism for children.

In this paper, we propose a new semi-supervised learning method called Clustering based Autistic Trait Classification (CATC), to improve the accuracy of the autism screening problem. The utilization of clustering and classification together as a semi-supervised learning is rare in autism screening research. Unlike existing methods that primarily focused on the classification phase of cases and controls, we intend to utilize clustering with classification to

validate instances in the training dataset prior to constructing the classification systems. CATC integrates unsupervised learning in the pre-processing phase with supervised learning in the classifiers construction phase. By integrating clustering with classification, there is a potential for improving the resulting classification systems by detecting ASD traits more accurately. With the CATC technique, the predictive model performance can be enhanced by answering the following research questions;

(1) Can the algorithm identify only the relevant features during the pre-processing phase of the dataset and clustering them in the training phase for the classification algorithm?

(2) Can data dimensionality be reduced by eliminating features redundancy?

Our proposed semi-supervised technique and nature of the algorithm can;

a) wrap those traits that may appear in multiple clusters and identify them as stronger or more significant features for the classification algorithms. By clustering the data first, we will identify relevant features that can be used in the ASD learning phase. Clustering can wrap those traits that may appear in multiple clusters and identify them as stronger or more significant features for the classification algorithms.

b) considers the hard cases to be classified (cases that exhibit few autistic symptoms). These cases may exhibit some autistic traits but may not be qualified to be on the spectrum. These cases often cause large false positives and false negatives, which deteriorate the performance of the classification algorithm. Thus, by having clustering at the pre-processing phase will enhance the predictability of the classification algorithm and improve the classifier accuracy, sensitivity, specificity, and error rates among others.

The rest of the paper is structured as follows: section 2 reviews the literature around machine learning in ASD research. Section 3 discusses some of the evaluation measures used in

the predictive models for ASD dataset. Section 4 outlines the experimental preparations and settings including a description of the datasets used and the pre-processing of the data. Section 5 provides the results and analysis and a comprehensive comparison of different ML techniques including CATC. Lastly, we provide a conclusion in section 6.

2: Literature Review

Crane, et. al.[17], highlighted some of challenges for a timely and adequate ASD diagnosis including the inadequate of the tools used to aid screening of ASD. Early and accurate diagnosis is beneficial as it can lead to early intervention and educational advice to parents with ASD children on what approaches to pursue going forward [23]. Early detection can assist parents with ASD children cope with parental stress that accompany ASD diagnosis which tend to lead to diminished behavioural interventions[19]. Thus, it is imperative that technology-enabled tools with predictive capabilities be employed to enhance ASD screening. For these reasons, many researchers use ML classification algorithms to predict ASD screening and diagnosis, i.e. in [9, 11, 12, 15, 21, 30, 31, 32, 39, 40, 44, 46, 47].

Thabtah et al., [41] improved the efficiency of the screening process by reducing the number of items in the self-assessment screening tool called AQ-10, [3]. The authors proposed a new feature selection ranking method called variable analysis (VA) that would derive small yet effective autistic traits. The authors used different datasets of adult, adolescent, and child in their study and compared their algorithm performance measures with other classification tools RIPPER and C4.5 [18, 35]. The results analysis showed that VA selected influential features for the three datasets (6, 8, and 8 respectively) without compromising on the specificity, sensitivity, and prediction accuracies measurements.

Similarly, Thabtah and Peebles[42] devised a new machine learning method called Rules-Machine Learning that detected autistic traits and offered users knowledge based rules that could be used by specialists to understand the logic behind such classification.

In a separate experiment, Thabtah, et al. [43] proposed a new ML framework that uses information gain (IG) and chi-square (CHI) testing to pinpoint a few influential features in the screening phase of ASD. The authors then use logistic regression algorithms to predict the accuracy on the results of their model. The results of their study indicated some of the strong (mainly communication and social behaviour) features that were commonly identified by IG and CHI to be influential in the ASD screening. The study, however, was only conducted on adult and adolescent datasets and did not focus on children and toddlers.

Abbas et al., [1] conducted a clinical study of 162 at-risk children that had received a clinical diagnosis. They collected their dataset by splitting their screening process into two parts. The first part is answered by the parent about the child based on the Autism Diagnostic Interview-Revised [ADI-R] [28] that have 93 multi-part questions. The second part is a video screener used by parents based on the Autism Diagnostic Observation Schedule [ADOS] [27]. The authors applied their datasets to Random Forests classifiers. They later combined the questionnaire and video screeners using regularized logistic regression. They then compared their results with some of the non-machine learning screening tools such as the modified checklist for autism in toddlers (MCHAT) and CCBL. Their results suggest that combining the video and questionnaire into a single assessment boosted the sensitivity and specificity rates and overall performance of the study sample.

A study by Levi et al., [25] utilized 2 ADOS modules; one for children with phrased speech (Module 2) and the other for children with verbal fluency (Module 3) to build sparse models that were used to train about 17 classifiers from 5 different classifier families (linear regressions, nearest neighbor models, general linear models, support vector machines, and tree-based classifiers) for autism screening and diagnosis. The module 2 dataset consisted of 1389 cases where 1319 were considered as ASD and only 70 as No-ASD. Module 3 dataset had 3143 cases with 2870 considered as ASD and 273 No-ASD. The study was applied on. The authors aimed

at showing reduced subsets of features with their best parameters that can be used in the classifiers to predict ASD and No-ASD cases. They concluded that SVM and logistic regression performed best with ROC of 93% and 92% respectively and logistic regression and Lasso performed best on module 3 with a ROC of 93%.

In their study of how some frequency-specific brain indices can be used in the early detection of ASD, Chen, et al., [15] used a limited data set from the Autism Brain Imaging Data Exchange database (ABIDE) of 240 with 112 with ASD and 128 with No-ASD. They conducted the experiment by looking at the brain functional connectivity as the frequency bands which were considered as the feature attributes of the dataset. The researchers used the support vector machine algorithm and could predict the ASD diagnosis with a classification accuracy of 79%.

Duda, et al., [21] did an experimental comparison of six classification algorithms on a real dataset consisting of 2900 cases with 65 features. The authors first pre-processed the data by removing any instances with more than four missing values. They applied logistic regression models, Random forests, support vector machine, C4.5 among other classification algorithms. They concluded that function based algorithms such as regression models performed better with high classification accuracy compared to the decision tree based algorithms such as Random Forest.

Others such as Pratap, et al., [32] and Pratap & Kanimozhiselvi [33] use multiple supervised and unsupervised machine learning algorithms such as Naïve Bayes, self-organization feature map (SOM), learning vector quantization (LVQ), artificial neural network (ANN), K-means and fuzzy c-means to test how machine learning methods can be used in the assessment of autism diagnosis. These two studies used a limited dataset of only 100 cases of children and were able to show that using unsupervised learning such as clustering improved the accuracy of the ASD based on the childhood autism rating scale (CARS) diagnostic tool. This study is limited in size

does not measure the improvement of other key classification metrics and have yet to be verified in other works.

Al-Diaba [14] used fuzzy rule-based technique to extract rules that can be used to predict autistic traits. The model learns the IF-THEN rules based on the different variables and applying the fuzzy unordered rule induction algorithm derive features that can be used to predict ASD for children in the screening phase. The authors compared their algorithm to that of JRip, RIDOR, and PRISM i.e. all rule based classification algorithms since they generate If-Then rules. This study is also limited in size as it only considered data for children without testing with other datasets.

A more recent review by Thabtah, [39] analysed some of the cons associated with ASD classification studies conducted earlier. The authors instigated that earlier studies had pitfalls in their datasets that were limited in size and had several missing values and imbalances. The author also pointed out that while the studies showed promising results, none were embedded in a screening tool.

Allison, et al., [3] study was aimed at reducing the AQ and Q-CHAT method screening tests by determining the highest ranked items based on DI measure scores. The authors were able to prove that only ten items can be used for screening first level ASD traits. These ten items were adopted in a later study by Thabtah, et al., [40] to build Adult, Adolescent, and Child datasets based on the AQ screening tool. These new datasets are used in the experiments in our paper.

Our study considers key classification evaluation measures. We evaluate and compare the results to highlight the significance of integrating clustering algorithms and specifically Multi-Cluster Overlapping K-Means Extension (OMCOKE) [5] in the pre-processing phase of the screening data. Clustering of the dataset adds the following value to the classification process;

(a) Reduces data dimensionality by eliminating redundancy.

(b) Identifies relevant and strong features that were only used in the supervised learning models. These features may have been cases that exhibited some autistic traits but not qualified to be on the spectrum hence causing large false positives and false negatives.

The following section discusses the proposed clustering based autistic trait classification technique.

3: The Proposed Clustering based Autistic Trait Classification (CATC)

In this section, we discuss the proposed CATC method based on the architecture shown in Figure 1 below. Three data sets (adult, adolescent, and child) are collected via a mobile screening app called ASDTest [37, 38]. The data is then cleaned for our experimentations and is ran through an unsupervised machine learning clustering algorithm. The result of this process is used as our initial model that is loaded to a classifier for the predictive phase. The performance of the classifier is then tested and evaluated for better accuracy, sensitivity, and specificity rates. Further details for each of the steps are outlined in the subsections that follow.

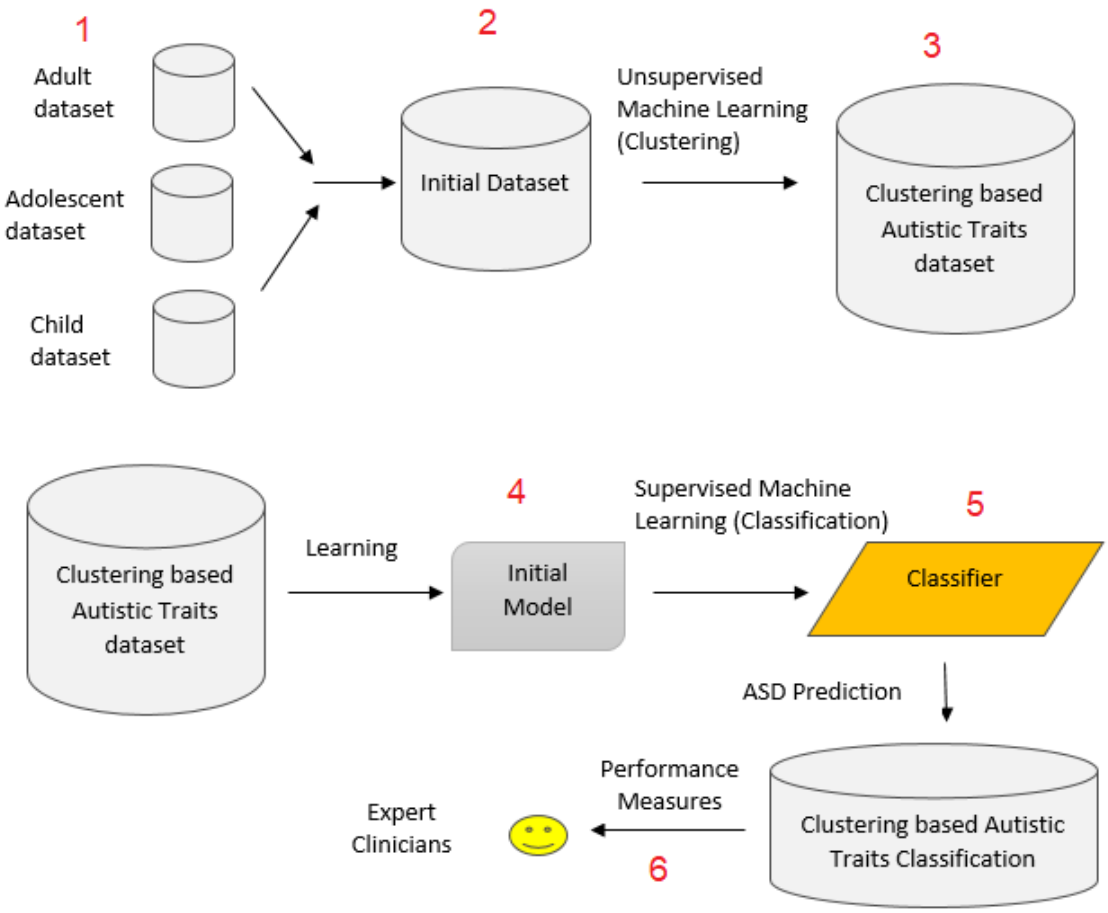

Figure 1. CATC-based methodology

3.1: Data Collection

Initially, data is collected using a mobile screening tool called ASDTests [37, 38]. This tool contains questionnaires based on the Q-CHAT 10, AQ-10 Child, AQ-10 Adolescent, and AQ-10 Adult screening methods by Allison, et. al., [3]. The child, adolescent and adult datasets that have been collected contain instances for individuals between 4-11 years old, 12-16 years old and above 16 years respectively. These datasets have been disseminated recently at the University of California Irvine data repository [26] by Thabtah et al., [40].

During the screening process using the ASDTests mobile application, a user answers the screening questions and a value is calculated based on the answers they enter with a score between 0 and 10. The attribute Class (attribute number 23 in table 2 below) is assigned a YES or a NO based on the score of the answers entered. A score of 6 and above based on [3] indicates

that the individual has some ASD traits and the class label is labeled as YES. Otherwise, the class is given a value of NO.

The size of the datasets varies between the three groups. The adult dataset has the highest number of instances followed by the child and adolescent. Table 1 and Figure 2 below summarizes the dataset based on the number of instances and the history of the users with regards to having a family member previously diagnosed with ASD.

Table 1. Statistics of used Datasets

Dataset	**Instances**	**Family History of ASD**	
		Yes	No
Adolescent	248	44	204
Adult	1118	183	935
Child	509	86	423

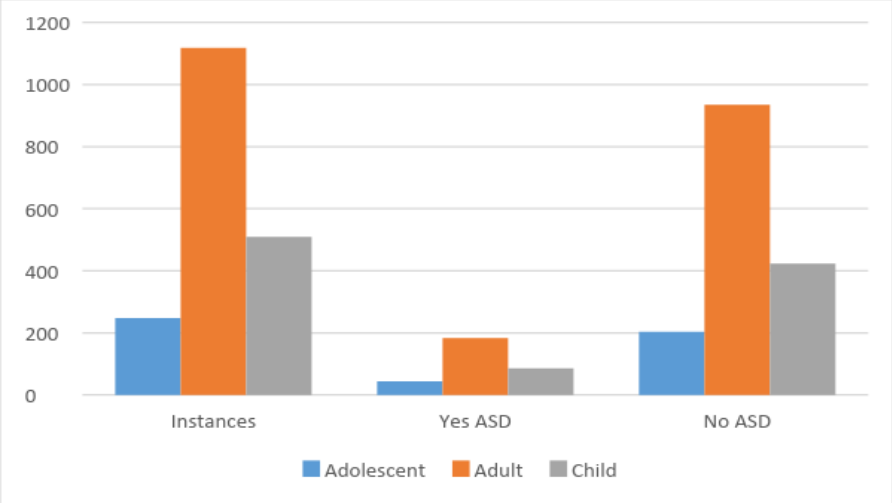

Figure 2. Statistics of used Datasets

Ethical Considerations

The data is published and made public [25] by its prospective author Thabtah et al., [40]. The authors of the datasets had obtained ethical approval from the University of Huddersfield, Huddersfield, UK.

3.2: The initial Dataset and Data Transformation

The initial datasets are of multivariable nature with categorical, continuous and binary attributes that contain a total of 23 features (see Table 2).

The ASDTest mobile application assigns a "1" if the respondent to any of the questions is "slightly agree" or "definitely agree", otherwise a zero "0" is allocated for questions 1, 5, 8, and 10 in the AQ-10 Adolescent, questions 1, 5, 7, and 10 in the AQ-10 Child, and questions 1, 7, 8, and 10 in the AQ-10 Adult. A "slightly disagree" or "definitely disagree" had a score of "1" on all remaining questions.

We modified the dataset to include only 18 attributes by removing features marked 16-22 in Table 2 below in the three datasets. The said features are general questions regarding the user and the app. We deem these features to have no direct significance and hence have been discarded a priori to the learning phase. The "Screening Score" (Feature #19 in Table 2) has been removed to avoid any possibility of model overfitting since this feature indicates whether individuals have autistic traits based on the scoring function in the AQ-Child 10, AQ-Adult 10 and AQ-Adolescent 10 methods. The screening features (A1 to A10 in Table 2) have been transformed by mapping its original values in the screening method to Boolean values 1/0 for the sake of simplicity.

Table 2. Feature Attributes

#	**Feature**	**Type**
1	A1	Binary (0, 1)
2	A2	Binary (0, 1)
3	A3	Binary (0, 1)
4	A4	Binary (0, 1)
5	A5	Binary (0, 1)
6	A6	Binary (0, 1)
7	A7	Binary (0, 1)
8	A8	Binary (0, 1)
9	A9	Binary (0, 1)
10	A10	Binary (0, 1)
11	Age	Integer
12	Gender	String
13	Ethnicity	String
14	Born with jaundice	Boolean (yes or no)
15	Family member with PDD	Boolean (yes or no)

16	Country of residence	String
17	Used the screening app before	Boolean (yes or no)
18	Why_are_you_taken_the_screening	String
19	Screening Score	Integer
20	Screening Method Type	Integer (0,1,2,3)
21	Language	String
22	Who is completing the test	String
23	Class	String

The AQ-10 screening questionnaire is used by the University of Cambridge autism research center as a referral guide. A sample of the adult questionnaire is provided in Table 3 below.

Table 3. AQ-10 Adult Questionnaire

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

3.3: Clustering Phase

The datasets are pre-processed by applying an unsupervised machine learning clustering method. We employ the OMCOKE algorithm which groups all items into two clusters. The process is summarized in Figure 3.

Input: Dataset with N number of attributes

Output: Reduced dataset with N+1 number of attributes

Given a test dataset, the pre-processing works as follow:

1. Apply Unsupervised ML Clustering (OMCOKE) filter to the dataset
2. A new "Cluster" attribute is appended on the dataset with each instance on the dataset assigned to either cluster1 or cluster2
3. **Repeat**
4. rename cluster1 as YES and cluster2 as NO
5. compare attribute "Cluster" with attribute "Class"
6. If Cluster Matches Class
7. Keep instance
8. Else
9. Prune instance
10. End if
11. **Until** all instances in the dataset are exhausted

Figure 3. Clustering phase Pseudocode in CATC method

The OMCOKE clustering technique assigns instances to either cluster1 or cluster2 based on their attribute similarities. The OMCOKE algorithm is based on K-means where initial k clusters are selected at random and data points are assigned to each cluster using distance to the centroids. The centroids are recomputed and the process is repeated until there is no movement or change in the assignment of data points to their closest centroid. The OMOCKE

algorithm takes into consideration outlier or noise data in the dataset and separates these points to an outlier cluster built on the fly. Algorithm 1 below summarizes the OMCOKE clustering.

Input: Number of clusters k, a set of a data vector
Output: membership assignment

Given a dataset, the algorithm works as follow:

1. Select *k* random points as initial cluster centers (centroid C_k)
2. **Repeat** For each x_i C_k
3. Centroid C_k dist(centroid C_k) // Re-compute and update the centroids
4. If (dist (x_i, centroid C_k) ≤ *averdist*)
5. Cluster ← x_i //Assign each data point to its closest centroid based on Euclidean distance
6. Else
7. If (dist (x_i, centroid C_k) ≥ *maxdist* * *maxdistThreshold*)
// data point greater than calculated threshold
8. Outlier_Cluster ← x_i // data point assigned to outlier cluster
9. Else
10. Cluster ← x_i // otherwise assigned to closest centroid
11. End if
12. End if
13. **Until** //convergence criteria are met and no change on each cluster

Algorithm 1 Unsupervised ML technique based on OMCOKE

3.4: Clustering Phase

The datasets contain a Boolean attribute named “Class” that has a value of YES/NO based on a score. This attribute Class is used to assess whether the user has been screened to have ASD or not and is used in the supervised learning algorithm for their predictions. At the end of step 2 in the CATC pre-processing phase above, we create a new attributed "Cluster" that

is appended at the end of the dataset file. Each item is assigned to either cluster1 or cluster2 based on their attribute similarities. These assignments are then compared to the attribute Class to see if they match. Where there is a match we keep that instance, otherwise we discard it and remove it from the dataset.

The new reduced clustering based autistic dataset only has instances that the clustering algorithm deems to have been accurately labeled during the unsupervised screening process. Key features of applying CATC process includes:

(1) Grouping the data items into two clusters based on their strong attributes. The clustering algorithm has assisted in identifying relevant and strong features that were only used in the supervised learning models.

(2) Reduce data dimensionality by eliminating redundancy. By clustering the significant features and comparing them to the class score, we toss out any insignificant or redundant items.

We adopt the clustering based autistic traits dataset which has been efficiently streamlined and enhanced to be used in the learning phase in the machine learning process. For example, assume the following simple dataset represented in figure 4 below as our original data.

Figure 4. Sample dataset

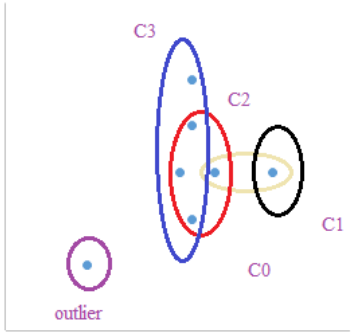

Figure 5. Clustered dataset

The clustering algorithm groups the data based on its strong attributes which are identified in C2 (red) and C3 (blue) clusters and the data anomaly is labeled as an outlier and discarded for use in the supervised learning model as indicated in the example on figure 5.

3.5: Classification Phase

Finally, we adopt any classification algorithm for our predictive phase. Classification algorithms are generally divided into a two-step process where the dataset is divided into training data and testing data. A model is developed in the training phase by analyzing the attributes of the training data. Class labels are built based on the rule techniques that are applied in the training dataset. This training data is further employed in the testing phase where the classifier is used to examine the accuracy of the models derived.

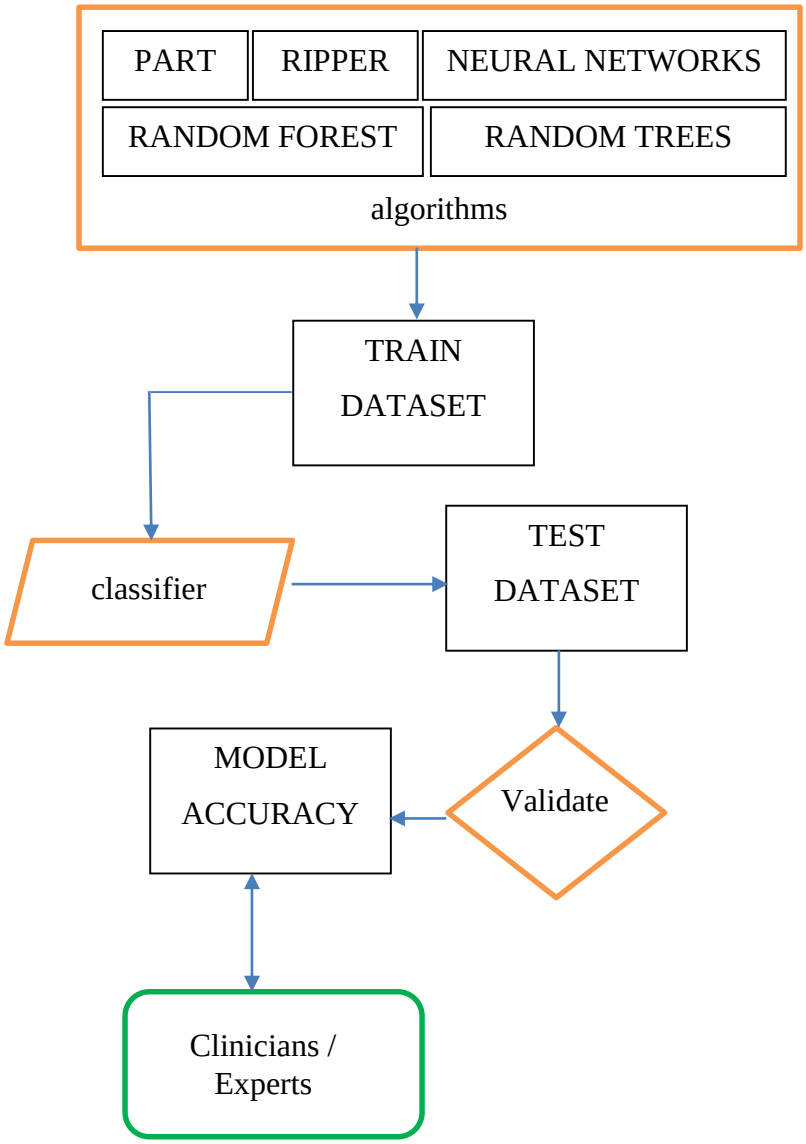

Figure 6. Classification Phase

Different classification algorithms (RIPPER[17], PART[21], Random Forest[13], Random Trees[19], and Artificial Neural Networks[45]) have been utilized to measure the performance of the proposed framework (CATC).

All empirical tests were conducted on WEKA version 3.8.2. WEKA stands for Waikato Environment for Knowledge Analysis and is an open source tool based on the Java platform that contains implementations for different ML methods including filtering, classification, clustering, evaluation, and visualisation among others. To build the CATC framework, we employed the OMCOKE algorithm built by Baadel, et al. [5] embedded in WEKA.

We then validate and evaluate the test dataset for better accuracy, sensitivity and specificity rates in order to test the performance of the clustering phase (See section 4.1 for further details).

4: Experimental Settings and Empirical Results

4.1: Experimental Settings

Our experiments are conducted on real-life ASD screening datasets to measure the effectiveness of the enhanced screening data used to identify and predict diagnosis. The three datasets of adult, adolescence, and child have a wide diversity in their ethnicity, language, and age group and are all in the application domain of the study, hence making it suitable for use as benchmarks.

We describe some common predictive model evaluation criteria such as accuracy, sensitivity, specificity, one-error, harmonic mean a.k.a. F1, and other related measures such as false positive (FP), false negative (FN), true positive (TP), and true negative (TN).

All experiments have been run on an Intel Core i7 computer with a 3.4 GHz processor

and 8.0 GB RAM running on a 64-bit Windows 10 Operating System. We utilized a number of evaluation measures to show the benefits and negatives of the proposed algorithm when compared with other classification algorithms in ML. Precisely, the below measures have been used to evaluate the CATC integration in supervised learning:

$$Sensitivity(\%) = \frac{|TP|}{|TP + FN|} \quad (1)$$

$$Specificity(\%) = \frac{|TN|}{|FP + TN|} \quad (2)$$

$$Accuracy(\%) = \frac{|TP + TN|}{|TP + TN + FP + FN|} \quad (3)$$

$$One\_error(\%) = 1 - Accuracy \quad (4)$$

$$F1 = 2x\frac{|\mathrm{Pr}ecisionx\,\mathrm{Re}call|}{|\mathrm{Pr}ecision + \mathrm{Re}call|} \quad (5)$$

Evaluation of a supervised learning model is a very critical process to assess the performance of the models derived. For ML predictive models, a matrix called the error table, or the confusion matrix, has been adopted. The confusion matrix is typically used to evaluate the performance of predictive models with respect to the different measures that are primarily related to the performance of the models. This is summarized in the confusion matrix in table 4 below.

Table 4. Confusion matrix

	Predicted Class	
Actual Class	**ASD**	**No ASD**
ASD	TP	FN
No ASD	FP	TN

Equations 1 through 5 (above) are used as evaluation metrics in determining the ML prediction performance. The Sensitivity ratio (equation 4.1) is a measure of all cases that have been identified correctly to have ASD in the overall test cases i.e. the true positive rate, whereas the Specificity ratio (equation 4.2) is a measure of all cases that have been identified correctly as a No ASD in the overall test cases i.e. the true negative rate. The Accuracy ratio measures the overall classification prediction that has been correctly identified as ASD and No ASD in all test cases (i.e. the confidence level of the classification), whereas the One error is the opposite of Accuracy and denotes the number of misclassified instances on the test dataset.

We conduct the experimentation twice for each dataset.

Different classification algorithms have been utilized to measure the true performance of the proposed framework (CATC). Particularly, we adopted RIPPER[17], PART[21], Random Forest[13], Random Trees[19], and Artificial Neural Network [ANN][45] algorithms to process the considered autism datasets with and without clustering. Thus two type of experiments have been conducted per dataset as follows:

Experiment (1): We first load the original datasets (adult, adolescent, child) without any clustering. Then we run the classification algorithms (RIPPER, PART, Random Forest, Random Trees, and Artificial Neural Network [ANN]) using their default settings and record their output results.

Experiment (2): CATC processed dataset in which the clustering is applied and all default settings of OMCOKE are maintained except the number of k clusters is changed from the default 3 to k = 2. Once this data has been pre-processed, then it is run using the classification algorithms above.

A tenfold cross-validation testing method on all the classifiers has been deployed in all experiments. This means that the dataset is partitioned into ten subsets where nine data subsets are used for the training phase and one subset for the prediction phase. The process is then

repeated 10 times. This will reduce overfitting and ensure a fair evaluation of the derived classifiers.

4.2: Results and Analysis

The experiments were conducted for the three datasets i.e. adult, adolescent, and child. The tables and the figures below show side by side comparison of the machine learning classifiers performance with/out CATC integration. The column CATC is marked as "No" when CATC was not applied to the dataset, otherwise a "Yes" is indicated. Table 5 compares the overall classification prediction i.e. accuracy rate for ML classifiers, noting a significant improvement when CATC is applied before the classification procedure.

Table 5: Accuracy Rates of the Classifiers

Dataset	**Classifier**	**CATC Clustering**	**Adult**	**Adolescent**	**Child**
Accuracy	RIPPER	No	0.942	0.807	0.878
		Yes	**0.969**	**0.944**	**0.936**
	PART	No	0.962	0.879	0.916
		Yes	**0.970**	**0.917**	**0.971**
	Random Forest	No	0.972	0.911	0.951
		Yes	**0.975**	**0.963**	**0.975**
	Random Tree	No	0.924	0.863	0.874
		Yes	**0.979**	**0.946**	**0.961**
	ANN	No	0.970	0.992	0.970
		Yes	**0.980**	**0.978**	**0.980**

Table 5 shows the accuracy rate of the models derived by the ML methods on the adult, adolescent, and child datasets. In all cases, the accuracy of the classifier has been improved by the ML method when CATC was applied prior training phase. In particular, RIPPER has seen an increase in the accuracy rate by 2.7%, 13.7% and 5.8% for the adult, adolescent, and child

datasets respectively when CATC was applied on these datasets. In addition, PART predictive accuracy has improved by 0.8%, 3.8% and 5.5% on the three datasets respectively when CATC was applied. Similarly, Random Forest and Random Tree classifiers when integrated with CATC have improved (0.3%, 5.2%, and 2.4%) and (5.4%, 8.3%, and 8.7%) respectively. No significant change is noted in the ANN method. The significant improvement in the accuracy can be attributed to the fact that having clustering in the pre-processing phase of the dataset was able to reduce data dimensionality by eliminating redundancy in the dataset. This shows overall better accuracy and lower error rates for all datasets including those that have large numbers of instances, i.e. adult dataset, and those with a lower number of instances, i.e. the adolescent dataset.

The accuracy rate alone may not be the best measure of performance because even with a 95% accuracy rate we might simply be predicting majority class correctly. Our focus, however, should be the other 5% minority class who might have been screened and diagnosed with autism. Thus, a good predictor of the model performance would be the true positive rate (sensitivity) and the true negative rate (specificity).

Figure 7 below shows the specificity and sensitivity results of the three datasets by the classifiers with and without CATC. The figure clearly reveals that when CATC was utilized prior to learning the sensitivity and specificity rates of the ML have improved on all datasets. For example, in RIPPER algorithm case, there is a modest sensitivity rate improvement on the adult and child datasets (2.9% and 2.8% respectively) and a 6.8% on the adolescent dataset, when CATC was applied. In addition, PART classifier's sensitivity rate went up by 0.9%, 6.9% and 7.5% on the adult, adolescent, and child respectively, when CATC was integrated. Similarly, when CATC was applied, Random Forest and Random Trees classifiers observed (1.8%, 11.1%, and 5.2%) and (5.3%, 8.1%, and 14.5%) increase in the sensitivity rates respectively. There is only a minuscule change in the ANN classifier.

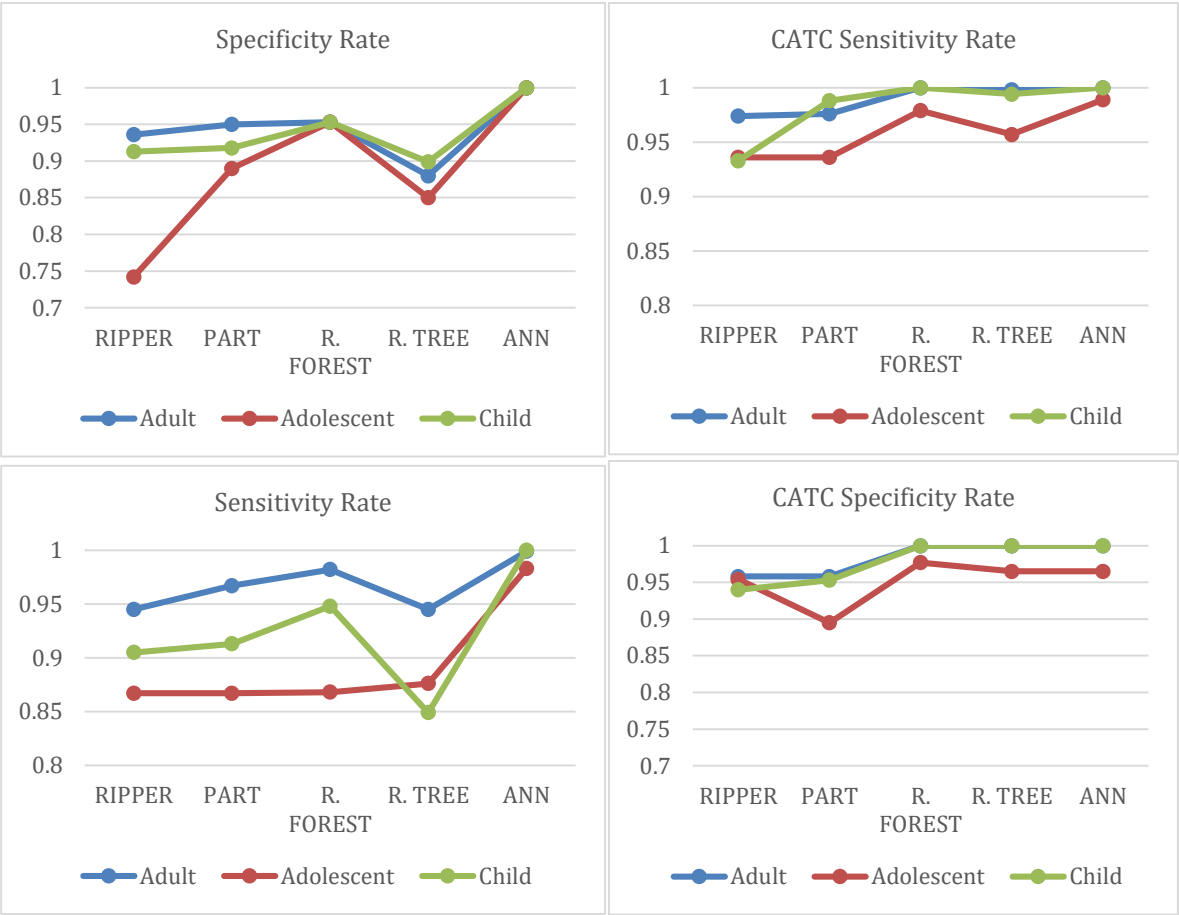

Figure 7. Sensitivity & Specificity Rates of the Classifiers

Here we note that clustering the datasets identified and grouped similar cases that would have otherwise been difficult to be classified correctly. Cases that may have exhibited some autistic traits but not qualified to be on the spectrum due to overlapping features of No ASD with ASD cases. These cases tend to confuse the learning algorithm in the classification process hence causing large false positives and false negatives. By considering the overlaps and clustering them based on the similarity of their attributes, this issue is resolved and the performance of the classification algorithms is dramatically improved as shown in Figure 7 and Table 4.

The specificity rates as shown in Figure 7 has seen an improvement of 2.2%, 0.8%, 4.7% and 12% for the adult dataset on the classifiers RIPPER, PART, Random Forest, and Random Tree respectively when CATC was applied. On the adolescent dataset, when CATC was applied, the percentage increment of the RIPPER, PART, Random Forest, and Random Tree classifiers

are 21.2%, 0.5%, 2.4%, and 11.5% respectively. Similarly, the performance of the classifiers went up by 2.7%, 3.5%, 4.7%, and 10.1% respectively on the child dataset.

To further understand the sensitivity and specificity rates, we investigated the confusion matrix results produced by the classifiers. Of all the three datasets, the adult dataset had the overall highest number of incorrectly classified instances by the RIPPER, PART, Random Forest, and Random Tree classifiers, whereas, the adolescent dataset had the least. Random Tree had the highest number of incorrectly classified instances (85) followed by RIPPER (65), PART (43), and Random Forest (31) in the adult dataset. Specifically, Random Tree predicted 43 instances with ASD traits that shouldn't have been classified resulting in the lowest specificity rate among the classifiers. On the other hand, Random Forest had the lowest number of false negatives with only 17 instances. CATC improved the classifiers by reducing the number of incorrectly classified instances to 21, 20, 0 and 1 for RIPPER, PART, Random Forest, and Random Tree respectively. In the adolescent dataset, CATC significantly reduced the incorrectly classified instances by the classifiers RIPPER, PART, Random Forest, and Random Tree from 48 to 10, 30 to 15, 22 to 4, and 34 to 7 respectively. Thus, CATC classifiers showed improvement in both sensitivity and specificity rates across the board compared with all classifiers.

With respect to the imbalance in the adult dataset due to the class variable, we included the F1 metric also known as the harmonic mean that not only takes into consideration the precision but also the sensitivity (equation 3.5 above). The F1 measure for the classifiers is shown to have increased by 12.3%, 0.9%, 2.8%, and 7.4% for RIPPER, PART, Random Forest, and Random Tree respectively when CATC was utilized (See Figure 8).

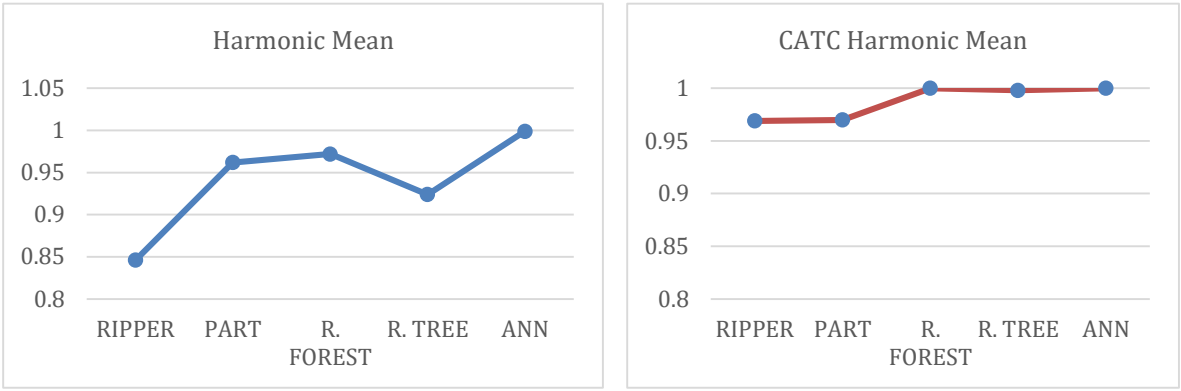

Figure 8. Harmonic mean on the classifiers

The Receiver Operating Characteristic (ROC) is an evaluation measure that contrasts the true positives and false positives of the machine learning model. The measure contrasts how the number of correctly classified true positives with the number of incorrectly classified negative values. Figure 9 summarizes the ROC values of the classifiers and an improvement across the board in the ROC Area rates when CATC is applied. While the improvement rates were decent in the adult dataset, it was slightly significant in the smaller dataset such as the Adolescent dataset. The ROC Area rate was up by 12.3%, 1.0%, 1.6%, and 9.8% on RIPPER, PART, Random Forest, Random Tree classifiers respectively, when CATC was utilized prior to learning.

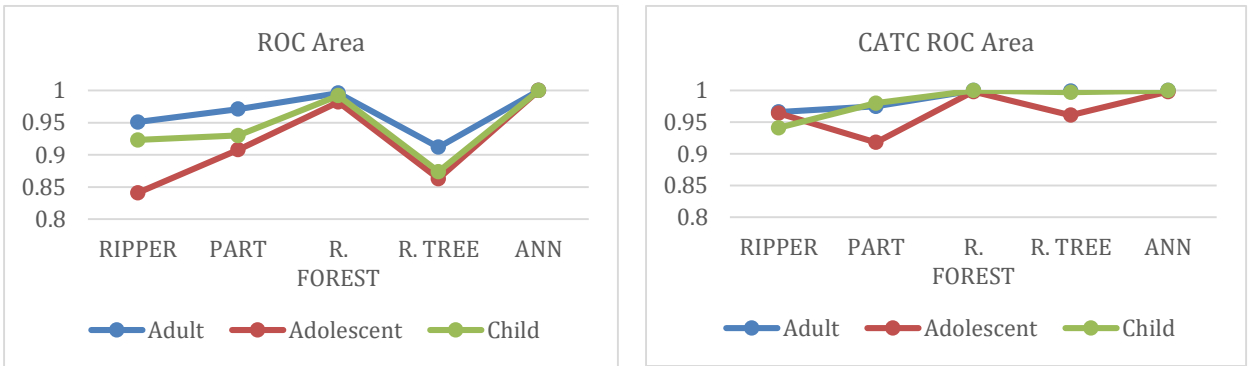

Figure 9. ROC Area of the classifiers

We also note that the number of rules generated while running the three datasets on RIPPER and PART decrease when CATC is applied as shown in figure 10 (below).

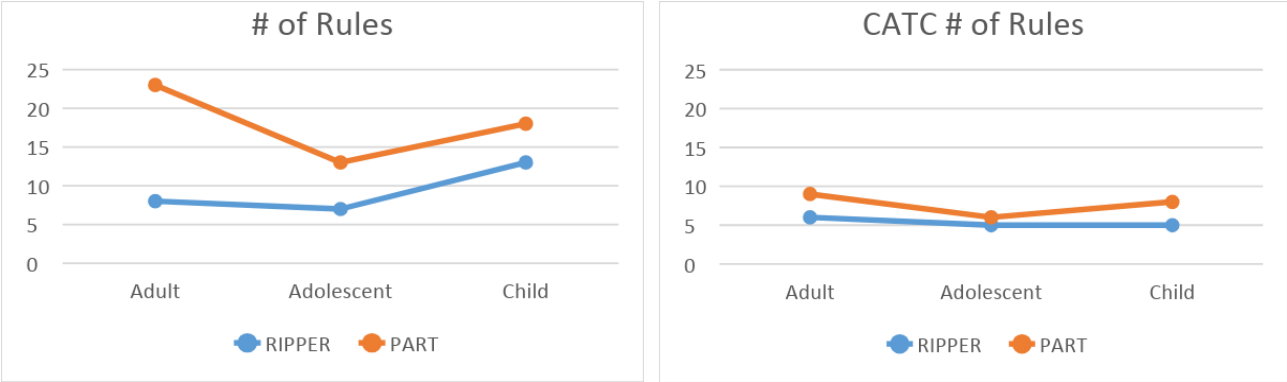

Figure 10. # of Rules Generated in PART and RIPPER classifiers

This can be attributed to the fact that redundant rules have been removed in the building of the classifier due to the pre-processing of the dataset and clustering them based on their strong attributes. Thus, the pre-processing with clustering algorithm have assisted in identifying relevant and strong features that were only used in the supervised learning models. This is useful for diagnosticians as fewer rules could mean a reduced amount of time needed in the screening of autism patients.

5: Conclusion

The utilization of clustering and classification together as a semi-supervised learning is rare in autism screening research. In this paper, we proposed a method that utilizes both clustering and classification in autism screening, a first that we are aware of. We used a screening app available on both Android and iOS mobile users and accessible easily online by the users. By introducing clustering a priori to classification we were able to add value to existing research in four folds;

(1) We were able to reduce data dimensionality by eliminating redundancy in the dataset.

(2) Cases that may have exhibited some autistic traits but not qualified to be on the spectrum due to overlapping features which caused large false positives and false negatives were resolved.

(3) Our method did not rely on the scoring function feature popularly used in other research to determine autistic traits in the screening phase but rather used unsupervised ML clustering algorithm to identify features based on their similarity measures.

(4) Clustering the data before application in the learning phase streamlined the data based on only strong features resulting in reduced number of rules generated by the classifiers.

There were a few limitations in this study. The data used was limited to what was collected using the mobile app. The datasets were limited in size and the adult dataset was slightly imbalanced. The study could have benefited with larger balanced datasets. Also, instances related to toddlers are rare and hard to obtain and were not included in this study.

In conclusion, the paper shows employing CATC in the screening phase significantly improved the performance of the classifiers in all measures and especially the accuracy and sensitivity rates. CATC improved the classifiers by reducing the number of incorrectly classified instances and improved on the sensitivity and specificity rates on the classifiers. We also saw a significant reduction on the rules generated by PART. These contributions provide important implications in social science, thus making a substantial positive difference in the prediction of the ASD diagnosis class. The proposed model is useful since they are exploited by diagnosticians and other stakeholders involved in ASD screening besides highlighting the most influential features. The methods used in our study can easily be adopted and applied to other clinical science application domain such as screening for dementia. Our future work will be to build a mobile screening app that will embed our clustering algorithm to assist clinicians in the diagnosis process of ASD in a clinical setting by considering wider options of diagnosis methods. Deep learning applications have a structure of algorithms inspired by the biological neural networks of the human brains in the form of artificial neural networks. Some of the ASD

traits involving facial recognition and behaviour are extremely difficult to detect and the study would benefit from data analytics deep learning algorithms.

Declaration of Conflict of Interest

The authors report no conflicts of interest

References

1. Abbas, H., Garberson, F., Glover, E., & Wall, D. P. (2018). Machine learning approach for early detection of autism by combining questionnaire and home video screening. *Journal of the American Medical Informatics Association. 25*(8) Pp. 1000-1007.

2. Achenbach, T. M., Rescorla, L. A. (2001). Manual for the ASEBA school-age forms & profiles. Burlington, VT: University of Vermont, Research Centre for Children, Youth, & Families.

3. Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief "Red Flags" for autism screening: The short autism spectrum quotient and the short quantitative checklist for autism in toddlers in 1,000 cases and 3,000 controls. *Journal of the American Academy of Child Adolescent Psychiatry, 51*(2), pp. 202–212.

4. Auyeung, B., Baron-Cohen, S., Wheelwright, S., & Allison, C. (2008). The autism spectrum quotient: Children's version (AQ-Child). *Journal of Autism Development Disorder, 38* (7), Pp. 1230–1240

5. Baadel, S., Thabtah, F., & Lu, J. (2016). Overlapping clustering algorithms: A review. In *Computing Conference* (SAI), London, UK. IEEE.

6. Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism spectrum quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism Development Disorder, 31*, pp. 5-17.

7. Baron-Cohen, S., Hoekstra, R. A., Knickmeyer, R., & Wheelwright, S. (2006). The autism-spectrum quotient (AQ)—Adolescent Version. *Journal of Autism and Development Disorder. 36*(3), Pp. 343 – 350.

8. Belmonte, M. K., Allen, G., Beckel-Mitchener, A., Boulanger, L. M., Carper, R. A., & Webb, S. J. (2004). Autism and abnormal development of brain connectivity. *Journal of Neuroscience, 24*, pp. 9228–9231.

9. Bekerom, B. (2017). *Using machine learning for detection of autism spectrum disorder.* 26th Twente Student Conference on IT. Enschede, The Netherlands.

10. Bolton P, Macdonald H, Pickles A, Rios P, Goode S, Crowson M, Bailey A, Rutter M. (1994). A case-control family history study of autism. Journal of Psychology & Psychiatry. 35(35) pp. 877–900. doi:10.1111/jcpp.1994.35.

11. Bone, D., Bishop, S., Black, M., … & Goodwin, M., (2016). Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion. *Journal of Child Psychology and Psychiatry, 57*(8), Pp. 927–37

12. Bone, D., Goodwin, M. S., Black, M. P., Lee, C., Audhkhasi, K., & Narayanan, S. (2014). Applying machine learning to facilitate autism diagnostics: Pitfalls and promises. *Journal of Autism and Developmental Disorders 45*(5), pp. 1–16.

13. Breiman, L. (2001). Random forests. *Machine Learning Journal, 45*(1), pp. 5-32.

14. Al-Diaba, M. (2018). Fuzzy Data Mining for Autism Classification of Children. International Journal of Advanced Computer Science and Application, 9(7). Pp. 11-17.

15. Chen, H., Duan, X. Liu, F., Lu, F., …& Ma, X (2016). Multivariate classification of autism spectrum disorder using frequency-specific resting-state functional connectivity - A multi-centre study. *Progress in Neuro-Psychopharmacology and Biological Psychiatry, 64*, pp. 1-9.

16. Crane, L., Chester, J., Goddard, L., Henry, L., Hill, E. (2016). Experiences of autism diagnosis: A survey of over 1000 parents in the United Kingdom. *Autism: The International Journal of Research and Practice, 20*(2). Pp. 153-162.

17. Crane, L., Betty, R., Adeyinka, H., et. al. (2018). Autism Diagnosis in the United Kingdom: Perspectives of Autistic Adults, Parents and Professionals. Journal of Autism Dev Disorder, 48(11). Pp. 3761-3772

18. Cohen, W. (1995). *Fast effective rule induction*. In proceedings of the Twelfth International Conference on Machine Learning. Tahoe City, California. Morgan Kaufmann.

19. Crowell, J. A., Keluskar, J., Gorecki, A. (2019). Parenting behavior and the development of children with autism spectrum disorder. Comprehensive Psychiatry Journal. 90 (2019). Pp. 21-29.

20. Cutler, A., Zhao, G. (2001). PERT-perfect random tree ensembles. *Computing Science and Statistics, 33,* pp 490-497.

21. Duda, M., Ma, R., Haber, N., & Wall, D. P. (2016). Use of machine learning for behavioural distinction of autism and ADHD. *Translation Psychiatry, 9*(6), pp. 1-5.

22. Frank, E., & Witten, I. (1998). *Generating accurate rule sets without global optimisation*. Proceedings of the Fifteenth International Conference on Machine Learning, pg. 144–151. Madison, Wisconsin.

23. Jacobs, D., Steyaert, J., Dierickx, K., & Hens, K. (2018). Implications of an Autism Spectrum Disorder Diagnosis: An Interview Study of How Physicians Experience the Diagnosis in a Young Child. Journal of Clinical Medicine, 7(10), 348. Pp. 1-16.

24. Krug, D., Arick, J., & Almond, P. (2008). Autism screening instrument for educational planning (3rd Edition). ProEd. Austin, TX.

25. Levy, S., Duda, M., Haber, N., & Wall, D. (2017). Sparsifying machine learning models identify stable subsets of predictive features for behavioural detection of autism. Molecular Autism, 8 (65). PMC5735531. doi: 10.1186/s13229-017-0180-6

26. Lichman, M. (2013). UCI machine learning repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

27. Lord, C., & Jones, R. M. (2012). Annual research review: Re-thinking the classification of autism spectrum disorders. *Journal of Child Psychology and Psychiatry, 53*(5), pp. 490–509.

28. Lord, C., Risi, S., Lambrecht, L., Cook, E. H. Jr, Lambrecht, B. L, DiLavore, P. C, Pickles, A., … & Rutter, M. (2000). The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism Development Disorder, 30*(30), Pp. 205–223.

29. Lord, C., Rutter, M., Le Couteur, A. (1994). Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism Development Disorder, 24*(24). Pp. 659–685.

30. Mythili, M., & Shanavas, M. R. (2014) A study on Autism spectrum disorders using classification techniques. *International Journal of Soft Computing and Engineering, 5*(6). pp. 7288–7291.

31. Pancer K., & Derkacz, A. (2015). *Consistency-based pre-processing for classification of data coming from evaluation sheets of subjects with ASDS.* Federated Conference on Computer Science and Information Systems, pp. 63–67. Lodz, Poland.

32. Pratap, A., & Kanimozhiselvi, C. S. (2012). *Application of naive Bayes dichotomizer supported with expected risk and discriminant functions in clinical decisions - Case study*. Fourth International Conference of Advanced Computing (ICoAC). Pp. 1-4. Chennai, India. IEEE.

33. Pratap, A, Kanimozhiselvi, C. S., Vijayakumar, R., & Pramod, K. V. (2014). Predictive assessment of autism using unsupervised machine learning models. *International Journal of Advance Intelligence Paradigms, 6*(2). Pp. 113–21.

34. Quinlan, J. (1986). Induction of decision trees. *Machine Learning, 1*(1), pp. 81-106.

35. Quinlan, J. (1994). C4.5: Programs for machine learning. Morgan Kaufmann Publishers.

36. Shopler, E., Reichler R., & DeVellis, R. (1980). Toward objective classification of childhood autism: Childhood autism rating scale (CARS). *Journal of Autism and Developmental Disorders, 10*. Pp. 91-103.

37. Thabtah, F. (2017a). *Autism spectrum disorder screening: Machine learning adaptation and DSM-5 fulfilment*. International Conference on Medical and Health Informatics. Taichin City, Taiwan.

38. Thabtah, F. (2017b). ASDTest: A mobile app for ASD Screening. www.asdtests.com

39. Thabtah, F. (2018a). *Machine learning in autistic spectrum disorder behaviour research: A review and ways forward.* Informatics for Health and Social Care. DOI: 10.1080/17538157.2017.1399132

40. Thabtah, F. (2018b). An accessible and efficient autism screening method for behavioural data and predictive analyses. Health informatics journal, pp. 1-23. 1460458218796636. https://doi.org/10.1177/1460458218796636

41. Thabtah, F., Kamalov, F., & Rajab, K. (2018). A new computational intelligence approach to detect autistic features for autism screening. *International Journal of Medical Informatics*, https://doi.org/10.1016/j.ijmedinf.2018.06.009

42. Thabtah, F., Peebles, D. (2019). A new machine learning model based on induction of rules for autism detection. Health Informatics Journal, 146045821882471. doi:10.1177/1460458218824711

43. Thabtah, F., Abdelhamid, N., Peebles, D., (2019). A machine learning autism classifcation based on logistic regression analysis. Health Information Science and Systems, 7(12). Pp 1-11.

44. Wall, D. P., Kosmiscki, J., Deluca, T. F., Harstad, L., & Fusaro, V. A. (2012). Use of machine learning to shorten observation-based screening and diagnosis of autism. *Translational Psychiatry, 2*(4), pp. 1-8.

45. Wiggins, L., Reynolds, A., Rice, C., Moody, E., Bernal, P., Blaskey, L., Rosenberg, S., Lee, L., Levy, S. (2014). Using standardized diagnostic instruments to classify children with autism in the study to explore early development. Journal of Autism and Developmental Disorders. 45(5), pp. 1271-1280.

46. Witten, I. & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques.* 2nd Ed, Elsevier.

47. Wolfers, T., Buitelaar, J. K., Beckmann, C. F., Franke, B., Marquand, A. F. (2015). From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neuroscience and Bio Behavioural Review, 57*, pp. 328–349.