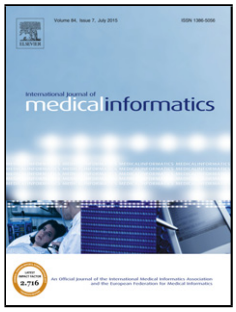


Title: A New Computational Intelligence Approach to Detect Autistic Features for Autism Screening

Authors: Fadi Thabtah, Firuz Kamalov, Khairan Rajab

PII: S1386-5056(18)30054-6
DOI: https://doi.org/10.1016/j.ijmedinf.2018.06.009
Reference: IJB 3717

To appear in: *International Journal of Medical Informatics*

Received date: 8-2-2018
Revised date: 28-5-2018
Accepted date: 12-6-2018

Please cite this article as: Thabtah F, Kamalov F, Rajab K, A New Computational Intelligence Approach to Detect Autistic Features for Autism Screening, *International Journal of Medical Informatics* (2018), https://doi.org/10.1016/j.ijmedinf.2018.06.009

A New Computational Intelligence Approach to Detect Autistic Features for Autism Screening

Fadi Thabtah
Manukau Institute of Technology
Auckland, New Zealand
Fadi.fayez@manukau.ac.nz

Firuz Kamalov
Canadian University of Dubai
Dubai, UAE
firuz@cud.ac.ae

Khairan Rajab
College of Computer Science and Information System, Najran University, Najran, Saudi Arabia
kdrajab@nu.edu.sa

Highlights

- New feature selection ranking method (VA) based on simplified likelihoods of observed and predicted values of variables is proposed
- Derive small yet effective autistic traits without hindering ASD screening performance
- VA maintained performance according to predictive accuracy, sensitivity and specificity rates
- In depth experimentations using 3 autism datasets and five known feature selection methods

Abstract

Autism Spectrum Disorder (ASD) is one of the fastest growing developmental disability diagnosis. General practitioners (GPs) and family physicians are typically the first point of contact for patients or family members concerned with ASD traits observed in themselves or their family member. Unfortunately, some families and adult patients are unaware of ASD traits that may be exhibited and as a result do not seek out necessary diagnostic services or contact their GP. Therefore, providing a quick, accessible, and simple tool utilizing items related to ASD to these families may increase the likelihood they will seek professional assessment and is vital to the early detection and treatment of ASD. This study aims at identifying fewer, albeit influential, features in common ASD screening methods in order to achieve efficient screening as demands on evaluating the items' influences on ASD within existing tools is urgent. To achieve this aim, a computational intelligence method called Variable Analysis (Va) is proposed that considers feature-to-class correlations and reduces feature-to-feature correlations. The results of the Va have been verified using two machine learning algorithms by deriving automated classification systems with respect to specificity, sensitivity, positive predictive values (PPVs), negative predictive values (NPVs), and predictive accuracy. Experimental results using cases and controls related to items in three common screening methods, along with features related to individuals, have been analysed and compared with results obtained from other common filtering methods. The results exhibited that Va was able to derive fewer numbers of features from adult, adolescent, and child screening methods yet maintained competitive predictive accuracy, sensitivity, and specificity rates.

Keywords: Accuracy, Autism Spectrum Disorder, Behaviour Science, Classifiers, Computational Intelligence, Data Mining, Feature Analysis, Machine Learning, Sensitivity, Specificity

1. Introduction

Autism Spectrum Disorder (ASD) refers to a neurodevelopmental disorder associated with limitations in social interactions, communication, and behaviour that is becoming increasingly common in many parts of the world (Ruzich, et al., 2015). The causes of ASD have been linked to genetic and neurological components, but are primarily diagnosed using non genetic variables related to behaviour such as social interaction, play and imagination, repetitive behaviours, and communication among others (Chakrabarti, et al., 2009; Geschwind, et al., 2001). Existing estimates reveal that about 1.5% of the world's population is on the spectrum, and it is believed that a huge number of individuals on the spectrum remain undetected (Brugha, et al., 2011; Fitzgerald, 2017). Therefore, high demands exist for faster diagnosis services corresponding with the growing awareness of ASD (Russell, et al., 2016).

Conventional diagnostic procedures for ASD require medical professionals to conduct a clinical assessment of the patient's developmental age based on a variety of domains (e.g., behaviour excesses, communication, self-care, social skills). This widely accepted approach is referred to as Clinical Judgment (CJ) (Wiggins, et al., 2014). The Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview (ADI) are two commonly used assessment tools that guide the diagnostic process (Lord, et al., 2000; Lord, et al., 1994). The latter is a structured interview comprised of 93 items, typically conducted by a licensed professional in cooperation with the patient's caregiver. ADOS offers four different modules for assessment, each applicable to a specific developmental age range, which may be applied across the spectrum from verbally fluent to non-verbal individuals. Patients are assessed based on their observed and/or reported performance with a set of developmentally appropriate activities associated with each module (Lord, et al., 2000).

In addition to CJ diagnostic methods, self-administered assessment methods (screening tools) have been developed, based mostly on existing CJ methods such as Child Behaviour Checklist (CBCL), Autism Behaviour Checklist (ABC), Autism Spectrum Quotient (AQ), Childhood Autism Rating Scale (CARS), and many others (Achenbach, 1991; Krug, et al., 2008; Baron-Cohen, 2001; Schopler, et al., 2009). These tools are typically conducted by caregivers, parents, teachers, or the patient (when feasible), and require responses to a large number of questions which makes many of them lengthy and inefficient (Thabtah, 2017a; Wall, et al., 2012b). Subsequently, reducing the number of variables in these screening methods may render the time consuming diagnostic process more efficient and empower busy medical clinics to conduct pre-diagnostic assessment more frequently. Further, many existing CJ and screening methods require the combination of tedious assessment methodologies that can take several visits (i.e., days) to complete alongside other common obstacles associated with the diagnostic processes that may affect the quality of health care patients receive. Therefore, identifying a small yet influential set of variables in the screening process is fundamental for reducing the tedious process and speed up referrals for patients.

Typically, ASD is pre-diagnosed based on observable and measurable behavioural indicators (e.g., social skills, engagement in age-appropriate play and leisure, behaviour excesses, communication skills, etc.) (Allison, et al., 2014). These indicators are often represented by items given in a questionnaire format for most screening methods (i.e. CBCL, ABC, AQ, CARS-2 and many others). However, in some cases the number of items could be large, making the experience of a pre-diagnosis test by users frustrating and inefficient with respect to time (significant for busy medical clinics). For instance, AQ, ABC, and CBCL contain 50, 57, and 100+ items respectively. This study seeks to delineate the significance of assessment components in order to ultimately reduce the number of components necessary for ASD classification without inhibiting the sensitivity, specificity, and accuracy of the process. Existing paradigms seem to subscribe to the idea that more questions translates into a more accurate classification. However, there is a need to re-examine features within ASD screening tools in order to generate smaller item sets while simultaneously maintaining the performance of the test.

Limited examinations of machine learning and computational intelligence perspectives on ASD have been previously conducted, in particular regarding the reduction of feature sets in screening tools. For example, Allison, et al., (2012), reduced the number of items used in the AQ screening method by Baron-Cohen (2001) to 10, rather than 50, items by using discriminant index (DI). In the same study, Baron-Cohen, et al.'s (1992) Quantitative Checklist for Autism in Toddlers (Q-CHAT) was also reduced from 25 items to 10 by using DI. The item's DI is calculated by taking the difference between the rate of individuals who have scored 1 on an item (showing a positive response to an autistic trait) from the rate of individuals who are with ASD. Allison, et al. (2012), computed the ID for each item in the AQ and Q-CHAT screening methods based on sample instances of adults, adolescents, and children, and then picked the top ten items in regard to IDs for each screening method. Nevertheless, the way items have been chosen were based on a DI, which is a simple rather than intelligent method. Other studies by Wall, et al. (2012a; 2012b), claimed that the number of items for ADOS-R (Module 1) in clinical environment can be minimised to 8 by using decision tree classifiers. However, a later study by Bone, et al. (2014), revealed serious pitfalls related to the clinical setup and both conceptual and implementation issues that have not been addressed by the authors, thus making their results questionable.

This study aims to build on previous attempts to reduce the number of items in AQ short versions to their minimum by discovering the least number of items that affect the process of ASD classification. The methodology used for achieving this aim is based on a newly proposed computational intelligence method being referred to as Variable Analysis (Va). Va intelligently computes the correlations between items in three AQ screening method versions based on normalised scores of the Information Gain (IG) and Chi-Square (CHI) methods (Liu & Setiono, 1995; Quinlan, 1986). By doing so, the scores are

stabilised and the discrepant behaviour of the results of these filtering methods is substantially reduced. Va assigns a weighted vector per item as a score, then ranks items based on their significance with the ASD class label. As such, a lesser number of items can be retained, thus improving the efficiency of the screening as well as pinpointing the most important items that contribute to autistic traits. By identifying a set of the least number of items, modern technologies such as mobile platforms can be used to allow more people to undergo screening for ASD, and will enhance accessibility throughout the healthcare community generally.

In the near future, this proposal can dramatically change the prospective of designing CJ diagnostic tools as well. By identifying the most influential items for ASD screening that should be included in the self-assessment tool, it will provide users with valuable concealed knowledge and guide the process of correct classification in a more efficient manner. The new method ensures that the remaining set of items are different from each other yet are highly correlated with the class of ASD. Va was tested on real datasets related to autism that have been collected recently using a mobile application called ASDTests (Thabtah, 2017b). In particular, three datasets (for children, adolescents, and adults) have been utilised in the experiments to show the merits and the issues with the proposed method. A machine learning approach has been employed in a verification step in order to derive classification systems based on the distinctive item sets of Va. The performance of Va was evaluated based on common metrics, including sensitivity, specificity, PPVs, NPVs, and predictive accuracy (see section 4 for further details).

The paper is structured as follows: section two reviews works related to the use of machine learning for reducing items in autism screening research. Section 3 is devoted to the description of the computational intelligence method. In section 4, the data collection process, data characteristics, experimental settings, and results analysis are conducted and explained in detail. Finally, the paper is concluded in section 5.

2. Literature Review

2.1 Related Works

Wall, et al. (2012a) utilised a number of data mining methods, in particular Alternating Decision Tree algorithm (ADTree), to reduce the number of items in the ADOS-Revised diagnostic method (Module 1). The aim of their study was to speed up the referral of the ASD diagnosis so that patients and their family members can quickly access healthcare services. To achieve this aim, the authors initially removed instances that are not clear ASD cases and then analysed the classification systems generated by the ADTree algorithm on an imbalanced dataset related to autism. The classification systems were produced using the Waikato Environment for Knowledge Analysis (WEKA) platform data analytics tool, which contain the implementation of ADTree algorithm (Hall, et al., 2011). After investigating the results of the ADTree algorithm the authors discovered that out of the 29 items of ADOS-Revised (Module 1) only 8 features appeared in the classification system and therefore concluded that the 29 items could thus be replaced with just 8 items. However, a clear shortcoming of this study is that when other data mining algorithms are used to process the same datasets other items appear besides the 8 items and some of the 8 items may disappear, therefore it is difficult to verify the results obtained. Another shortcoming of Wall, et al., (2012a) is the removal of the overlapping instances between the ASD and No ASD class labels before applying the ADTree algorithm during data processing since this action simplifies the ASD classification rather than providing a realistic solution.

Bone, et al. (2014), clarified pitfalls in Wall, et al., (2012a), and Wall, et al., (2012b), by employing data mining or machine learning algorithms to handle ASD classification. Bone, et al. (2014), argued that careful steps must be taken before using data processing methods, especially when the application under investigation is a clinical one that necessitates careful setup. A number of shortcomings were highlighted and explained in depth by the researchers. In particular, there was no time reduction gained when using the reduced 8 items of ADOS-Revised (Module 1) since the entire activities of the ADOS-Revised procedure must be taken by the clinician prior to applying ADTree on the dataset. More importantly, the ADOS-Revised procedure must be taken within a clinical setup and cannot be self-administered. Therefore, this study clearly pinpointed limitations of the results obtained earlier by Wall, et al. (2012a) and Wall, et al. (2012b).

An investigation on reducing the time of AQ and Q-CHAT method screening tests was conducted by Allison, et al. (2012). The authors utilised a dataset that consists of cases and controls of ASD collected through a web-based recruitment strategy. The correlations of the variables in the dataset have been computed using a DI that was defined as the proportion of positive instances for a variable, i.e. PI, in the training dataset from PI proportion, which is derived from the control training set. The aim was to determine the highest ranked items in the AQ and Q-CHAT methods based on the DI scores. Once the scores were sorted, the authors then evaluated the top 10 ranked items using different metrics including sensitivity and specificity among others. Results based on these metrics revealed that minimising AQ and Q-CHAT to just ten items is possible. Nevertheless, these 10 items can only be utilised to screen a population on a first level of ASD traits and not for clinical diagnosis.

Despite the results achieved by Allison, et al. (2012), other computational intelligence and machine learning techniques can validate the scores obtained by the DI measure used and thereby establish higher reliability of the outcome. This can also provide new potential opportunities by exploring automated classification and scoring of both control and cases. This study takes the output of the ten items in the AQ short versions of Adult, Adolescent, and Child datasets and evaluates their significance with the ASD class using the new methodology.

Duda, et al. (2016), and Kosmicki, et al., (2015), investigated the impact of different data mining techniques to speed the time needed by the Social Responsiveness Scale (SRS) and ADOS methods respectively. The dataset used in the Duda, et al. (2016), study was imbalanced, i.e. 2775 ASD and 150 ADHD instances. Containing 65 variables, the dataset was adopted from Simplex Simon Collection (SSC) version 15 (Fischbach & Lord, 2010). Before applying the data mining techniques, instances with more than four missing values were discarded and specific variables selected based on forward selection. Moreover, under sampling was applied to resolve the imbalanced target variable in the dataset by setting the ratio of ASD to ADHD instances to 1.5:1. After processing the training dataset, results pinpointed to higher predictive accuracy for classifiers produced by Logistic Regression when compared to decision tree classifiers such as Random Forest. Despite the reduction of the variables using feature selection by Duda, et al. (2016), there was no clear mechanism on how to distinguish cases of ADHD from those of ASD in an automated manner.

Kosmicki, et al. (2015), experimented with eight algorithms and applied the step wide backward feature selection method to reduce the number of items in ADOS modules (2, 3). Based on the data mining algorithms' results, there were 9 items claimed to be effective by the authors in ADOS module (2) and 12 items of module (3). These items' data, when mined, showed consistency in sensitivity and specificity when compared to the full item sets in these two ADOS modules.

Thabtah (2017c) investigated different pitfalls related to the use of machine learning in ASD research. The author pointed out that the majority of existing models related to machine learning have not considered replacing the static handcrafted rules and scoring procedures with automated classifiers during the process of ASD classification. Rather, existing studies have only considered employing machine learning on static datasets related to autism and measured the effectiveness of these techniques with respect to only sensitivity and specificity primarily. The authors also revealed conceptualization issues related to datasets and feature selection that have not been considered by the majority of existing machine learning studies.

2.1 Screening Methods Used in the Study

Initially, AQ is a self-administrated ASD screening tool developed by Baron-Cohen (2001) together with other behavioural scientists from the Autism Research Centre, University of Cambridge, in order to identify autistic traits and other neurodevelopmental symptoms in adults with an average level of intelligence. The AQ questionnaire consists of 50 different questions covering the areas of social skills, attention switching, imagination, communication, and attention to detail. The AQ test is available online, and each question has four possible rating responses ('Definitely Agree,' 'Slightly Agree,' 'Slightly Disagree,' and 'Definitely Disagree') depending on which final score is calculated. The final score can range from 0-50, and a higher score indicates an increased level of autistic symptoms. The AQ authors suggested that a cut-off score of 32 would optimise the validity of screening adults in a clinical setting. Later, in 2006 and 2008, two different versions of AQ were developed to cover adolescents and children

(Baron-Cohen, et al., 2006; Auyeung, et al., 2008; Allison, et al., 2012). AQ-Child is a parent administered questionnaire specially designed for children aged 4-11 years, whereas the AQ adolescent is designed for teenagers aged 12-15 years. All the versions of AQ contain 50 different items and take approximately 20-30 minutes to complete.

Baron-Cohen, et al. (2006) and Auyeung, et al. (2008), developed full AQ versions for adolescents and children respectively. To improve the usability of the tests, Allison, et al. (2012) presented a compressed version of the original AQ adult version called AQ-10-adult. Even though AQ-10 is shorter than the original version, it has comparable performance to the original AQ version. The questions of QA-10 have four possible responses like the original AQ version, and the scoring rule often considers one point per question. To be clear, a point is added if the answer is either 'Slightly Agree' or 'Definitely Agree' for questions 1, 7, 8, and 10. In addition, a point is added if the user's responses to questions 2, 3, 4, 5, 6, and 9 are either 'Slightly Disagree' or 'Definitely Disagree'. The overall score is then calculated using a handcrafted diagnosis rule, and anyone who scores above the threshold of six is considered to have autistic traits. Moreover, Allison, et al. (2012), proposed shorter versions for the full adolescent and child AQ tests. Score calculations for the adolescent and child short versions are different from the AQ adult short version. Details of the questions, and the scoring, of the shortened adolescent and child AQ versions can be found in Allison, et al. (2012).

When it comes to validity, the AQ-child has the highest sensitivity (95%) and specificity (95%) out of all the versions of AQ. However, overall sensitivity and specificity of AQ are reported as 77% and 74% respectively at a cut-off score of 32 (Auyeung, et al., 2008).

3. The Proposed Computational Intelligence Method

CHI-SQ and IG are two different methods normally used in classification research to evaluate the worthiness of variables in an input dataset (Liu & Setiono, 1995; Quinlan, 1986). Often, the input data, also called the training dataset, consists of many variables plus a target variable called the class. The aim of these methods is to reduce the dimensionality of the training dataset by producing a smaller number of variables so that efficiency or the quality of predictive models can be enhanced (Abdelhamid & Thabtah, 2014). The worthiness of the variable is typically measured statistically, using specific metrics by computing the correlation between each variable and the target variable (class). In the case of ASD screening, the variable will be an item/question in the screening method, and the class is whether an individual has ASD traits. For instance, CHI-SQ computes the correlation between variable-class (*v*,*l*) using their expected and observed probabilities in the training dataset (T) based on Equation (1),

$$CHI - SQ(v,l) = \frac{S \times (AD - BC)}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \quad (1)$$

where A is the frequency of the variable-class (*v*,*l*) in T, B is the frequency of the variable *v* without class *l* in T, C is the frequency of class *l* without variable *v* in T, D is the frequency of instances not having both (*v*,*l*) in T, and *S* is the size of T.

On the other hand, IG calculates the worthiness of a variable, using the entropy of the class with and without the presence of a variable according to Equation 2,

$$G(l,v) = E(l) - E(l|v) \quad (2)$$

where l is the class, v is the variable, E () is the Entropy and $G()$ is Gained information. So, the IG for a training set $T = \{(v, v, \dots, v,\ l)\}_i$ where v_a is the value of the vth variable, and l is the corresponding class, can be calculated for vth by Equation 3:

$$G(l,v) = -\sum_{l \in L} p(l) \log p(l) + p(v) \sum_{l \in L} p(l|v) \log p(l|v) + p(\bar{v}) \sum_{l \in L} p(l|\bar{v}) \log p(l|\bar{v}) \quad (3)$$

One of the notable problems with statistical data reduction methods is the discrepancies in their variables' scores, especially when applied to different datasets. For example, when the CHI-SQ and IG methods are applied to the "arrhythmia" dataset that contains 280 variables as in Lichman (2013), they select 110 (37 variables respectively, using 15 and 0.15 minimum thresholds for CHI-SQ and IG). When

these thresholds are lowered, the results showed more abnormality with respect to variables' scores and ranks. This example, if limited, shows that filtering methods such as CHI-SQ and IG highly deviate in their results because of the different metrics used in computing the scores. Therefore, there is a need to unify the scores of variables in order to stabilize them. This can be achieved by utilising multiple scores per variable that can be initially normalised. By normalising the scores of methods, such as IG and CHI-SQ, similarities and deviations will be minimised and a unified global weight can then be assigned to each variable which may result in less redundant features. This is the basic idea of the Va method.

Va aims at unifying score volatility in IG and CHI-SQ by a) amalgamating their scores per variable, 2) Normalising their scores to one score per variable, and 3) utilising the new score as a new metric for ranking the variables. By accomplishing this idea, a new truer rank will be given to each variable which reduces variable-to-variable correlations and maintains good variable-to-class correlation. This metric could substantially reduce the number of variables selected, especially for medical and behaviour science applications such as autism classification. In ASD screening research it is vital to keep only the highest impact items during the screening in order to render screening time efficient while sustaining performance. This will be advantageous to the clinician, as well as other healthcare users such as patients, caregivers, parents, and teachers, who now would only have to answer a concise set of items during the pre-diagnosis process.

Va creates a new vector based on the CHI-SQ and IG scores and computes the magnitude of the vector as M_score. In Va, M-score is utilised as a robust measure to differentiate among the goodness of variables (features that are selected from the ones that get discarded). The mathematical formulations of Va are given in Equations 4-7. Initially, the scores derived by CHI-SQ and IG are different since each method employs a specific metric to evaluate the worthiness of variables, therefore these scores must be normalised in order to make them comparable. To accomplish this task, Va normalises the scores by defining $CHI\_SQ_{max}\ and\ IG_{max}$ to be the maximum score for a variable obtained by the CHI-SQ and IG methods on a training dataset, T (Equations 4-5). The normalised scores of the vth variable in CHI-SQ and IG results can then be defined.

$$\overline{CHI\_SQ_v} = \frac{CHI\_SQ_v}{CHI\_SQ_{max}} \qquad (4)$$

$$\overline{IG_v} = \frac{IG_v}{IG_{max}} \qquad (5)$$

Next, the *score vector* of feature v can be defined to be

$$sv_a = \begin{pmatrix} IG_v \\ CHI\_SQ_v \end{pmatrix} \qquad (6)$$

The new score vector is fundamental since it holds important information related to the scores of CHI-SQ and IG. Magnitude of the score vector, sv_v, can be utilised as a scalar measure of the vector as shown in Equation 7. This is since the magnitude of a vector can be computed by taking the square root of the sum of the squares of its coordinates. This new measure (sv_v) can easily be employed to compare between variables in which those with a larger sv_vcan assigned to a higher order and thus have an increased chance of being selected.

$$|sv_v| = \sqrt{(IG_v)^2 + (CHI_v)^2} \qquad (7)$$

The strategy for assessing variables proposed by Va differs from existing approaches that combine scores in feature selection (i.e. AND and OR) since the Va strategy offers a mathematical configuration for examining the space of scores. Magnitude of the score vector can be used to compare features to one another. Features with a greater value of $|sv_v|$ will be ranked higher. Unlike other ways of combining scores from different methods, such as AND / OR, this approach yields a true metric on the space of all pairs of scores. This allows for a mathematical structure for analysing the space of combined scores.

Variables in the proposed Va method with low scores are eliminated and only keep variables from high scores. To determine the line between influential and non-influential variables, any variable with a score below 50% of the largest score obtained on a dataset is discarded. Nevertheless, the user is allowed to determine the cut-off points (selection criterion) since application data varies based on the criteria of determining the acceptable score per variable. For instance, medical diagnostic applications, such as ASD screening, may require a small set of variables if compared with text categorisation or email classification and thus a cut-off score of 50% is ideal.

The cut-off values have been suggested based on extensive warming up experimental analysis on different datasets in which 5%, 10%, 25%, 50%, and 75% values have been tested. The results showed stability on the number of remaining variables when the cut-off is set to 50%. One possible way to identify potential cut-off points is to consider the balance of the number of variables remaining and the predictive accuracy obtained. In this context, if the cut-off is set too high, such as 75%, the expected number of variables detected by Va will be very limited and may cause a significant reduction in accuracy. On the other hand, if the Va cut-off is set to a smaller value, i.e. 10%, larger numbers of variables can be returned and may enhance the predictive power of the models but may over-fit the model. There should be a trade-off between the number of variables which remain and the expected predictive accuracy of the models.

Distinctive features of Va are its simplicity and ability to minimise variations of feature selection scores by existing methods, which may increase legitimacy in the final result by the user. In addition, Va proposes an amalgamated vector that gets assigned to each variable and results in fewer variables without drastically influencing the models. Minimising variable redundancy during the process of data processing causes a simultaneous reduction in the number of variables produced. This is advantageous for both decision makers as well as computing machines. The latter will use fewer resources during data processing, enhancing its efficiency. The former will have fewer variables to exploit, which improves their models' interpretability and subsequently their decision making. In the context of ASD, patients, their families, and medical staff will only have to to exploit a lesser number of autistic traits when seeking to understand the causes and effects of ASD, at least at the screening level.

4. Data and Experimental Results

4.1 Data Collection and Description

Three versions of the AQ screening method (AQ-10 Adult, AQ-10 Child, and AQ-10 Adolescent) and Q-CHAT-10 were implemented in a mobile application environment for the purpose of data collection for this study. The mobile app utilized is called ASDTests, and has been implemented and published in both Android and IOS operating systems in order to facilitate accessibility for users (Thabtah, 2017b). Figure 1A shows the landing page of the ASDTests app in which the user can select the appropriate age category before taking the screening test. The data collection tool enables different types of users to pre-diagnose ASD traits by selecting their test based on the appropriate age category. Each test consists of ten questions in a sequential order, and each is associated with an image in order to enable users to carefully select the appropriate answer. Users can use touch screens to navigate through the app, which can be run on smart phones (Android and IOS) as well as tablets. Figure 1B depicts one sample question from the toddler screening test.

Prior to completing the assessment, participants were required to consent to a disclaimer which explained the goal of the research, privacy policy, and use of the data. Participants were notified that their information would be kept anonymous and only shared for research purposes. During the data collection phase there was no direct access to participants and the ASDTests mobile application clearly stated that use of the data would be for research purposes only. The participants had to read this before submitting their answers. Since no name or sensitive information is involved, participant identities are anonymous (see variables in table 1A and their description in table 1B).

Once the user completes the test (all 10 questions – A1-A10), a screen emerges to review the answer. On this screen, the user can review their answers and amend any they wish. The screen serves as a quality assurance measure so that users can verify their answers before progressing to the data submission page. The key functionality of the ASDTests app is to collect relevant useful data about the case undergoing the screening. In particular, the features displayed in Table 1 are collected. Besides the main features that are related to the screening of ASD and the case under consideration, a target class variable has been created with a Boolean value to determine whether the individual undergoing the test has ASD or not. The class variable value is assigned automatically based on the final score

obtained by the individual taking the ASD test. For example, if the individual has selected an age category of 12-16 on the ASDTests app the scoring will be based on the AQ-Adolescent method. In this case, if the final score was between 6 and 10, the class value for this case will be assigned "Yes," otherwise it would be assigned "No." A class value with "Yes" relates that the case requires further assessment by an expert while a class value with "No" indicates that the individual has no autistic traits. The features shown in Table 1 are stored in a MYSQL database, and can be used for further data analysis in order to understand key features that may influence ASD diagnosis from a behavioural science perspective.

The values of the A1-A10 variables in each dataset have been recoded to either "0" or "1" depending on the actual answers given by the participants during the screening process (see table 1B for the actual questionnaire). In other words, during the AQ-10-Child method screening "1" was given for questions 1, 5, 7, and 10 if the participants answered any of them with "Definitely" or "Slightly Agree," and "0" otherwise. For the rest of the questions, "1" was utilised if the answer was "Definitely" or "Slightly Disagree," otherwise "0" was assigned. For the AQ-10-Adolescent method, "1" was allocated to questions 1, 5, 8 and 10 if the given answer was "Definitely" or "Slightly Agree" for each, whereas "1" was allocated to "Definitely" or "Slightly Disagree" responses on the remaining questions. Lastly, for the AQ-10-Adult method, "1" was given for "Definitely" or "Slightly Agree" answers for questions 1, 7, 8, and 10. For the rest of the questions in this method "1" was assigned when "Definitely" or "Slightly Disagree" was chosen for questions 2, 3, 4, 5, 6, or 9. This representation of "1" or "0" per feature in the screening method can ease the process of data processing by the machine learning algorithms utilised during the building of the classification systems.

Table 2 shows 20 sample data instances that have been collected based on the AQ-10 Child assessment. A total of 1,452 instances that belong to toddlers, children, adolescents, and adults were collected over a period of 4 months using the ASDTests app and based on Q-CHAT-10, AQ-10 Child, AQ-10 Adolescent, and the AQ-10 Adult screening methods respectively. After an initial investigation on the collected instances it was clear that the vast majority of the instances that belong to toddlers and infants have been associated with a "no ASD" class label, making such a group of data completely imbalanced. To be exact, 96% of the cases who took the test for the Q-CHAT-10 (toddlers) have not been associated with ASD, and therefore the toddlers instances are separated from other instances linked with children, adolescents, and adults as well as omitted from further analysis. This left 1,100 instances that belonged to three target audiences (children, adolescents, and adults). In addition, since the "scoring result" was used conventionally by the AQ-10 version to classify cases and controls, this variable has been discarded prior to data processing.

Figure 2A shows the instances of distribution with respect to age, and Figure 2B shows the class distribution per age category. It is clear from Figures 2A and 2B that there are more adult instances than adolescent and children, as well as more instances associated with the "No ASD" class label. A basic explanation for more none ASD cases is that the population normally contains a much higher quantity of individuals with no ASD traits than those with ASD markers. Moreover, Figure 2B reveals that child instances are somewhat balanced with respect to class labels when compared with adult and adolescent instances respectively.

The average age in years for children, adolescents, and adults in the three subsets based on the screening methods used are 6.3, 14.1, and 29.7 respectively. More male instances have occurred than female, as the number of instances for male and female in the three subsets (1100 total instances) are 625 and 475 respectively. The top ethnicities which participated in the data collection were Caucasian-European, Asian, Middle Eastern, South Asian, and African/African-American with 381, 185, 128 and 65 respectively. In the three subsets of data, there were 707 and 393 instances linked with the "ASD trait" and "No ASD trait" class labels. More of the tests for adults have expectedly been taken by the individuals themselves while many tests for the child category have been taken by parents, teachers, or caregivers. Among the gathered instances, there were 194 cases with family members diagnosed with ASD and 165 cases of individuals born with jaundice. Lastly, some values were missing in variables such as ethnicity and who_is_taken_the_test.

4.2 Experimental Settings

In this section, the performance analysis of the proposed autistic trait feature selection method is presented based on the three subsets of data (child, adolescent, and adult). These datasets have been collected using the ASDTests mobile app which implements four different screening methods (AQ-10 Adult, AQ-10 Child, AQ-10 Adolescent, and Q-CHAT-10). In particular, the autistic trait feature sets selected by the Va method is evaluated and their performance compared with sets chosen by four other common methods, namely IG, Correlation Attribute Evaluation, Correlation Features Set (CFS), and CHI-SQ (Quinlan, 1986; Witten & Frank, 2005; Hall, 1999; Liu & Setiono, 1995). The case of no feature selection is also considered. Reasons behind choosing these filtering methods are three fold:

a) They produce scores per feature and therefore are ranking based methods.
b) They have different mechanisms for computing the scores of the available features and proved their merits in many classification benchmarks.
c) They are all implemented within the WEKA environment.

The basis of the comparison is different evaluation metrics; including sensitivity, specificity, and predictive accuracy among others (see Equations 9-11). Two machine learning algorithms have been employed, named Repeated Incremental Pruning to Produce Error Reduction (RIPPER) and C4.5 (Decision Tree), in order to produce ASD classification systems from the different subsets of features chosen by Va, IG, Correlation, CFS, and CHI (Cohen, 1995; Quinlan, 1993). These classification systems will show the true performance (upsides and downsides) of the Va when contrasted with different filtering methods, particularly predictive accuracy, sensitivity, and specificity. The reason for employing two different predictive algorithms is to generalise the results obtained, especially goodness of the distinctive features. RIPPER is a rule-based classifier that normally employs excessive rule pruning to generate If-Then rules. Meanwhile, C4.5 is a decision tree algorithm that constructs classifiers using entropy in the format of trees. Both algorithms are well studied in the machine learning and data mining communities, and produce high quality results with respect to classification accuracy according to different experimental studies by Abdelhamid, et al. (2017), Thabtah and Kamalov (2017), and Mohamed, et al. (2014).

Va has been implemented in the Java programming language and embedded within WEKA version 3.9.1 (Hall, et al., 2011). WEKA is a machine learning tool that contains large collections of learning, visualization, filtering, and dimensionality reduction techniques. For fair comparison, all experiments of the considered data reduction methods have been conducted in WEKA. In addition, to build the classification systems from the distinctive feature sets derived by the filtering methods ten-fold cross validation has been adopted (Witten & Frank, 2005). Ten-fold cross validation is a testing method utilised in learning algorithms that ensures the training dataset is split into ten parts. The learning algorithm, in this case RIPPER or C4.5, will train on nine parts and generate a classifier. This classifier is then tested on the remaining parts to derive different evaluation metrics, primarily the error rate. The same procedure is repeated on the same dataset ten times, arbitrarily splitting the dataset into ten parts each time in order to produce an error rate. Lastly, all error rates generated are averaged to come up with one global error rate for the classifier. All experiments have been performed on a computing machine with 2.0 GHz processor and 8 RAM memory.

Since screening individuals in this article can be seen as a binary classification problem (ASD, No-ASD) evaluation metrics have been adopted that go along with the nature of the problem. Common metrics derived from the confusion matrix (Table 3), such as accuracy, specificity, and sensitivity, are used to evaluate the performance of the features selected by the filtering methods (Witten & Frank, 2005). Table 3 depicts the possible answer for a test instance decision during a screening process. Predictive accuracy (Equation 9) is one of the common measures employed in ASD research and the domain of classification. Using this measure, the number of test instances that have been correctly predicted from the total number of test instances are computed. Sensitivity (Equation 10) denotes the proportion of the test instances that are actually ASD (true positive rate) whereas specificity (Equation 11) is the proportion of the test instances who do not have ASD (true negative rate). Positive Predictive Value (PPV) and Negative Predictive Value (NPV) have also been included as shown in equations 12

and 13 respectively. NPV and PPV represent the percentages of negative and positive diagnostic test instances that are true negative and true positive outcomes respectively.

$$Error\_Rate = 1\text{-}\ Accuracy \quad (8)$$

$$Accuracy = \frac{|TP+TN|}{|TP+TN+FP+FN|} \quad (9)$$

$$Specificity = \frac{|TP|}{|TP+FN|} \quad (10)$$

$$Specificity = \frac{|TN|}{|TN+FP|} \quad (11)$$

$$PPV = \frac{|TP|}{|TP+FP|} \quad (12)$$

$$PPV = \frac{|TN|}{|TN+FN|} \quad (13)$$

4.3 Results Analysis

Multiple experiments have been conducted using feature selection and machine learning against the datasets below:

1) The subset of instances that have been derived using the AQ-10 Child screening method.
2) The subset of instances that have been derived using the AQ-10 Adolescent screening method.
3) The subset of instances that have been derived using the AQ-Adult10 screening method.

4.3.1 Feature Reduction Results

Figure 3 shows the number of features chosen by the filtering methods. It is clear from the derived figures that Va consistently selects a lesser number of features for all datasets considered in comparison with the other methods. Specifically, Va selected 8, 8, and 6 items from the AQ-Child, AQ-Adolescent, and AQ-Adult datasets respectively. This indicates that the proposed method not only seeks for feature-class correlations but also eliminates feature-feature correlations, leading to less redundant features in the final set. It is notable from the figures that the remaining number of features in Va from the larger datasets is smaller than those from the smaller datasets. In particular there were 6 features remaining after applying Va against the Adult dataset, which is an indication that Va works well in situations where instances are more represented among class labels. This may boost the performance of classifiers generated against the features selected by Va during assessment of large datasets when compared to those selected from the limited size datasets. To investigate this case the features selected by Va from the adult datasets are displayed in Table 4A.

Table 4A shows that there are six common features (items 3, 4, 5, 6, 9, and 10 in the AQ-10 Adult screening method) detected by Va from the adult dataset. When these items were checked in the AQ-10 Adult screening method, it was discovered that most of these items correspond to social interaction and social communication (Category A) in the DSM-5 manual. However, these items do not fully fulfil the new ASD diagnostic criteria under the Diagnostic and Statistical Manual of Mental Disorders DSM-5 (American Psychiatric Association, 2013). For instance, items 4-6 and item 10 are covering conditions under categories A and A.1 in DSM-5 (social communication and social interaction). This means that despite these items fulfilling multiple criteria in Category A, they still do not cover any condition in

Category B (Restricted and Reparative Behaviour) whiles the DSM-5 requires at least two criteria in category B to be met before an individual can be diagnosed with ASD. Therefore, these sets of items do not comprehensively cover the required minimum conditions for meeting an ASD classification. Nevertheless, screening for ASD does not necessarily require fully meeting the diagnostic conditions of ASD as its ultimate aim is merely to reveal potential autistic traits rather than diagnosing individuals since to do so necessitates the involvement of expert clinicians and clinical setup. Identifying the most influential features in AQ short versions to their minimum number is, therefore, a definite advantage.

Tables 4B and 4C show the reduced sets of items detected by the Va method for the AQ-10 Adolescent and AQ-10 Child assessments. The results show a reduction in the items chosen by Va from these two screening methods, detecting 8 items from the adolescent and child datasets respectively. An interesting finding based on the figures within these tables is that there are 5 items in common between the AQ-10 Adolescent and AQ-10 Child screening methods as detected by Va (highlighted in yellow within Tables 4B and 4C). In addition, item_4 (adolescent) can also be matched with item_4 (child), making the common items between AQ-Adolescent and AQ-Child significantly high (six out of eight). These results demonstrate that there are high levels of overlapping between the AQ-10 Child and AQ-10 Adolescent screening methods, at least at the autistic traits level. Less overlapping occurs between the AQ-10 Adult and AQ-10 Adolescent methods as detected by Va. Overall, the results clearly show that the top three items detected by Va from the AQ-10 adolescent and AQ10 -Child screening methods are related to communication and social traits. The top three items selected by Va from the AQ-10 Adult screening method, however, are related to behaviour and social traits.

Table 5 depicts the percentages of relative difference (Equation 12) of features chosen from the datasets between Va and the considered methods. The rates show that Va reduced the number of remaining features significantly when compared with the considered filtering methods. For example, Va minimised features in the Child dataset by 27.3%, 46.7%, 38.5%, and 20% when compared with results obtained by the CHI, IG, Correlation, and CFS filtering methods. The reduction is also clear in the adult dataset, where Va reduced the number of features by 53.8%, 62.5%, 60.0%, and 45.5% respectively when compared to results obtained by CHI, IG, Correlation, and CFS. The results of Table 5 clearly reveal that in the screening methods the Va method was also able to cut down the number of specific items as it takes into account multiple scores per feature and then normalises these scores into a new, single, global score that is then assigned to the feature. This gives the true rank per feature and reduces score discrepancies since each global score was computed by multiple scores which were generated by different filtering methods. In other words, Va reduces the deviations of the feature scores and ensures stability and the true weight per feature.

$$\frac{(\#\ \text{of features chosen by Method i} - \#\ \text{of features chosen by Va})}{\#\ \text{of features chosen by Method i}} \quad (12)$$

4.3.2 Accuracy, Sensitivity Specificity, PPVs, NPVs Results

To reveal the performance of Va when compared with other filtering methods dealing with the ASD classification problem, a number of experiments using two machine learning algorithms, RIPPER and C4.5, have been conducted on the sets of feature data obtained in Section 4.3.2. Answers to the questions “will Va maintain the performance despite reducing the number of features” and “what will be the differences in sensitivity, specificity, PPVs, NPVs and accuracy between Va and other known filtering methods assessing the different ASD datasets” were sought. To answer these questions, instances that belong to each feature set, obtained earlier using the RIPPER and C4.5 algorithms, were processed.

Typically, the classification of the features obtained by the Va method are performed using rule-based classifiers that produce “If Then” rules. These rules have been derived by RIPPER and decision tree algorithms. Using each training dataset (i.e. child, adolescent, adult) as well as each target class, RIPPER starts with an empty rule (If empty then ASD) and then appends features into the rule until it cannot grow any further. At this stage, RIPPER evaluates the rule against a pruning set in order to improve its predictive accuracy. Once the rule gets evaluated, RIPPER generates the rule and removes

its corresponding training instances and repeats the process on the remaining uncovered instances until no more data is left for class ASD. On the other hand, C4.5 uses entropy to build decision trees that in turn are converted into rules sets. In classifying a test instance, both RIPPER and C4.5 assign the class of the first rule that matches the test instance's items (features values) to the test instance. There was no involvement of the "scoring result" variable in the classification process of the machine learning algorithms. Only features that have been chosen by Va have been used in building the classifiers. This paper offers a new way of reducing the amount of features by offering limited influential variables for autism screening to the machine learning algorithms in order to enhance ASD screening performance.

Figures 4A and 4B show the predictive accuracies for the different subsets of data chosen by the filtering methods and using the C4.5 and RIPPER algorithms. In these figures the accuracies derived by C4.5 and RIPPER against CHI-features, IG-features, Correlation-features, Va-feature, and "no feature selection" (the 21 original features) were considered. The accuracy results revealed that Va scales well with IG, CHI, Correlation, and CFS, especially for the adolescent data based on the derived classifiers. Despite the slight drop in accuracy for Va derived features for the adult and child datasets, Va maintained an acceptable accuracy rate and, more importantly, was able to significantly reduce the number of features. Specifically, for the largest dataset (Adult) the C4.5 algorithm derived classifiers from Va feature sets with just under 2.8% less accuracy than those of IG, CHI, Correlation, and CFS. If the Va only selecting 6 features from the adult dataset is considered, this 2.8% drop in accuracy can be tolerated in light of the more than double in number of features remaining in these datasets by the other filtering methods considered (features that relate to both screening items and individuals characteristics).

Surprisingly, when the Va chosen feature set of the adolescent data is mined, the highest accuracy rate when using the RIPPER algorithm is derived in comparison with all remaining feature subsets. In fact, the accuracy rate of the classifier derived from the Va feature set was 10% and 6% higher than the original set of features (21) when the RIPPER and C4.5 algorithms were utilised for data processing on the adolescent dataset. One probable reason for this is that Va selected eight features from the AQ-Adolescent dataset and therefore more items covering autistic conditions, such as social, communication, and repeated behaviour, have been selected. One notable result from the adolescent dataset showed that the considered filtering methods' feature set, when processed by the machine learning algorithms, performed with less than the acceptable level of predictive accuracy. This can be attributed to the fact that only a limited number of instances for the adolescent instances are available and instances which are highly imbalanced in this dataset have many more "No ASD" cases than ASD. Nevertheless, Va showed good performance in regard to its accuracy rate in the presence of a limited and imbalanced dataset, which is a distinct advantage.

Figures 5A - 5B and 6A-6B display the sensitivity and specificity rates derived by the RIPPER and C4.5 classifiers from the different distinctive feature sets' data. Often in medical research, including autism, acceptable levels of sensitivity and specificity should be at least 80% (Towle, et al., 2016). The results of the sensitivity and specificity rates derived by the machine learning algorithms against the feature sets' data has shown an acceptable level, except for the adolescent subset, for the majority of the filtering methods aside from Va. For that particular subset the specificity rates, derived by the machine learning algorithms against the Va and IG feature sets, showed adequate rates thus proving that these two filtering methods perform well even when a limited number of instances are present. When processed, the Va feature set showed slightly lower sensitivity and specificity rates on the child and adult datasets but maintained adequate rates. In particular, the specificity rates derived from the Va features of the adult dataset were 2.8%, 1.9%, 1.4%, 3.6%, and 3.0% less than those of the "no feature selection," IG, CHI, Correlation, and CFS feature sets. Va has also only used 6 features while original data, IG, CHI, Correlation, and CFS are associated with 21, 14, 14, 14, and 11 features respectively.

On the other hand, for the adolescent dataset, Va was superior to "no feature selection," CHI, Correlation, and CFS, having achieved higher accuracy by 17.1%, 2.5%, 17.10%, and 4.9% respectively, thereby proving that Va can handle noisy data better than the other filtering methods. The sensitivity rates derived by RIPPER from the features of Va, CHI, IG, Correlation, CFS, and "no feature

selection" on the AQ-Adolescent dataset were 87.30%, 80.95%, 80.95%, 84.13%, and 80.00%. These rates clearly show a good level of sensitivity by the considered filtering methods and the superiority of Va on this particular dataset.

For the adult dataset, the sensitivity rates derived by the C4.5 and RIPPER algorithms from Va's features were 80.95% and 82.54% respectively, and were lower than the results obtained by the remaining filtering methods. For this data subset, out of 189 positive instances that were supposed to have "ASD" classification the C4.5 algorithm misclassified 36 instances to ASD (False negatives). In addition, 14 instances were incorrectly classified by C4.5 as having ASD, having been screened clinically as not on the spectrum (No ASD) (False Positives). These misclassifications have caused lower sensitivity rates for the features chosen by Va, at least for the adult dataset. It is believed that this can be attributed to not having enough selected features that cover the needed criteria of ASD. In addition, some of these "No ASD" instances have overlapping features with ASD cases while not fulfilling the entire criteria of ASD. These instances are the hardest to be predicted since they confuse the learning algorithm during the classification process, resulting in a slight increase in the false positive and false negative rates. Overall, Va has performed well in terms of sensitivity and specificity rates for the adolescent dataset and comparably adequate but below most filtering methods on the other two datasets (Adult, Child).

Figures 7A-7B and 8A-8B demonstrate the PPV and NPV rates derived by the RIPPER and C4.5 algorithms from the different datasets considered. These PPV and PNV results show acceptable levels, except for the adolescent dataset. For this dataset the classifiers extracted by the RIPPER algorithm on Va's selected subset were superior in terms of PPVs and NPVs in comparison to those extracted from the remaining features sets. For example, the NPVs produced by RIPPER from the Va features set are 0.64%, 7.76%, 6.57%, 17.07%, and 4.88% larger than those derived from "no feature selection," IG, CHI, Correlation, and CFS feature sets respectively. On the other hand, the classifiers derived from the Va data subset showed slightly lower PPV and NPV rates on the child and adult datasets but maintained adequate rates. The PPV and NPV results were consistent with the predictive accuracy results derived previously.

5. Conclusions

Self-administered ASD assessment tools, also known as screening tools, are typically conducted by a caregiver, medical staff, or the patient when feasible and require responses to a large number of items. In addition, the validity and accuracy of assessments based on these tools rely upon classification methods with antiquated technologies which should be a concern for users in the healthcare community. Possible ways to improve the classification accuracy and efficiency of the current screening tools is to adopt new intelligent methods based on machine learning and computational intelligence. The latter can be utilised to identify a concise set of items that can be implemented using new technologies such as mobile platforms, thus rendering the existing time consuming tools unnecessary. To process data based on the outcome of the computation intelligence in order to automate the classification process and enhance the predictive accuracy of the test can be conducted by utilising the former. This paper has proposed a new computational intelligence method called Va that significantly reduces the number of features needed for ASD screening methods while maintaining sensitivity, specificity, and predictive accuracy rates. Va has been implemented in Java within the WEKA machine learning tool. Three primary datasets related to AQ-Adult, AQ-Child, and AQ-Adolescent, in addition to multiple features related to individuals, have been collected over a four month period by using a mobile app called ASDTests. To measure the performance of the Va method, five different filtering methods have been contrasted with Va along with using two machine learning algorithms called RIPPER and C4.5. The role of the machine learning algorithms was to build classifiers from the feature sets' data derived by Va and the considered filtering methods. These feature sets have been derived from the adult, child and adolescent datasets. The results clearly demonstrated that Va was able to choose a fewer number of items from the three datasets than the other filtering methods considered. In particular, Va reduced the AQ-10 Adult, AQ-10 Adolescent, and AQ-10 Child to 6, 8, and 8 items while maintaining acceptable levels of specificity, sensitivity, and predictive accuracy. These item sets, when processed by machine

learning algorithms, derive highly competitive ASD classifiers with adequate levels of accuracy, PPVs, NPVs, specificity, and sensitivity rates. The concise sets of items and classifiers generated are of high interest to the different individuals interested in ASD screening. These results can also assist in early detection of ASD traits, thus facilitating access to necessary support systems for the physical, social, and educational well-being of the patient and family in addition to increasing the likelihood of improved outcomes in the patient.

On the behalf of the authors we declare that this paper has not been submitted before and is not under review in other journals

Dr Fadi Thabtah

On the behalf of the authors I declare that there is no conflict of interest.

Dr Fadi Thabtah

Summary Points

- Screening of autism is often time consuming
- Many items are involved during the screening process
- Little research on computational intelligence and machine learning toward autism screening
- Identifying few yet relevant items for autism screening is challenging
- New computational intelligence method to red flag influential items is proposed
- Methodology was based around intelligent technology called machine learning
- In depth experimentations to reveal performance of the proposed method, i.e. sensitivity, specificity and accuracy of the screening and using 5 other different feature selection methods
- Results showed significant small autistic traits can be utilized for screening and maintained performance

References

[1] Abdelhamid, N., Thabtah, F.,and Abdel-jaber, H. (2017). Phishing detection: A recent intelligent machine learning comparison based on models content and features. 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), pp. 72-77. 2017/7/22, Beijing, China.

[2] Abdelhamid, N., and Thabtah, F. (2014). Associative Classification Approaches: Review and Comparison. *Journal of Information and Knowledge Management (JIKM), 13*(3).

[3] Achenbach, T. (1991). *Integrative guide for the 1991 CBCL/4-18,YSR, and TRF profiles*. Burlington: University of Vermont, Department of Psychiatry.

[4] Allison, C., Auyeung, B., and 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), 202–12Y.

[5] American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

[6] Auyeung, B., Baron-Cohen, S., Wheelwright, S., and Allison, C. (2008). The Autism Spectrum Quotient: Children's Version (AQ-Child). *J Autism Dev Disord 38*: 1230–1240.

[7] Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., and 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*, 5–17.

[8] Baron-Cohen, S., Hoekstra, R., Knickmeyer, R., and Wheelwright, S. (2006). The Autism-Spectrum Quotient (AQ) – adolescent version. *J Autism Dev Disord 2006, 36*: 343 -50

[9] Bone, D., Bishop, S., Black, M., Goodwin, M., Lord, C., Narayanan, S. (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*, 927-937.

[10] Bone, D., Goodwin, M., Black, M., Lee, C., Audhkhasi, K., and Narayanan, S. (2014). Applying Machine Learning to Facilitate Autism Diagnostics: Pitfalls and Promises. *Journal of Autism and Developmental Disorders 45*(5), 1–16.

[11] Brugha, T., et al. (2011). Epidemiology of autism spectrum disorders in adults in the community in England. *Archives of General Psychiatry 68*(5): 459–466.

[12] Chakrabarti, B., Dudbridge, F., Kent, L., Wheelwright, S., Hill-Cawthorne, G., Allison, C., et al. (2009). Genes related to sex steroids, neural growth, and social emotional behavior are associated with autistic traits, empathy, and Asperger syndrome. *Autism Research 2*(3), 157–17.

[13] Cohen, W. (1995). Fast Effective Rule Induction. In In Proceedings of the Twelfth International Conference on Machine Learning. Tahoe City, California, 1995. Morgan Kaufmann.

[14] Duda, M., Ma, R., Haber, N., and Wall, D. (2016). Use of machine learning for behavioral distinction of autism and ADHD. *Translational Psychiatry 9*(6), 732.

[15] Fischbach, G., Lord, C. (). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron 68*, 192–195.

[16] Fitzgerald, M. (2017). The Clinical Gestalts of Autism: Over 40 years of Clinical Experience with Autism. In *Autism - Paradigms, Recent Research, and Clinical Applications* (pp. 1–13). InTech. http://doi.org/10.5772/65906

[17] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations 11*(1).

[18] Hall, M. (1999). Correlation-based Feature Selection for Machine Learning. Thesis, department of computer science, Waikaito University, New Zealand.

[19] Geschwind, D., Sowinski, J., Lord, C., Iversen, P., Shestack, J., Jones, P. et al. (2001). The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. *American Journal of Human Genetics 69*, 463–466.

[20] Kosmicki, J., Sochat, V., Duda, M., Wall, D. (2015). Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. *Translational Psychiatry 5*, 514.

[21] Krug, D., Arick, J., and Almond, P. (2008). ASIEP-3 (Autism screening instrument for educational planning).

[22] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

[23] Liu, H. and Setiono, R. (1995). Chi2: Feature Selection and Discretization of Numeric Attribute. Proceedings of the Seventh IEEE International Conference on Tools with Artificial Intelligence, November 5-8, 1995, pp. 388.

[24] Lord, C., Risi, S., Lambrecht, L., et al. (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 Disorders 30*, 205–223.

[25] Lord, C., Rutter, M., and 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 and Developmental Disorders 24*, 659–685.

[26] Mohammad, R., Thabtah, F., McCluskey, L. (2014). Intelligent rule-based phishing websites classification. *IET Information Security, 8*(3): 153-160.

[27] Quinlan, J. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

[28] Quinlan, J. (1986). Induction of Decision Trees. *Mach. Learn. 1*(1): 81-106.

[29] Russell, A., et al. (2016). The mental health of individuals referred for assessment of autism spectrum disorder in adulthood: A clinic report. *Autism*, *20*(5), 623–627.

[30] Ruzich, E., Allison, C., Smith, P., Watson, P., Auyeung, B., Ring, H., and Baron-Cohen, S. (2015). Measuring autistic traits in the general population: a systematic review of the Autism-Spectrum Quotient (AQ) in a nonclinical population sample of 6,900 typical adult males and females. *Molecular Autism 6*(2).

[31] Schopler, E., Reichler R., DeVellis, R. (1980). Toward objective classification of childhood autism: Childhood autism rating scale (CARS). *Journal of Autism and Developmental Disorders 10*:91–103.

[32] Thabtah, F. (2017a). Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment. Proceedings of the 1st International Conference on Medical and Health Informatics 2017, pp.1-6. Taichung City, Taiwan, ACM.

[33] Thabtah, F. (2017b). ASDTests. A mobile app for ASD screening. www.asdtests.com [accessed November 30th, 2017].

[34] Thabtah, F. (2017c). Machine Learning in Autistic Spectrum Disorder Behavioural Research: A Review. To Appear in *Informatics for Health and Social Care Journal*. December, 2017 (in press).

[35] Thabtah, F., and Kamalov, F. (2017). Phishing detection: a case analysis on classifiers with rules using machine learning. *Journal of Information & Knowledge Management, 16*.

[36] Towle, P., and Patrick, P. (2016). Autism spectrum disorder screening instruments for very young children: a systematic review. *Autism Res Treat 2016*:4624829.

[37] Wall, D., Kosmiscki, J., Deluca, T., Harstad, L., Fusaro, V. (2012a). Use Of Machine Learning To Shorten Observation-Based Screening And Diagnosis Of Autism. *Translational Psychiatry 2*.

[38] Wall, D., Dally, R.. Luyster, R., Jung, J., Deluca, T. (2012b). Use Of Artificial Intelligence To Shorten the behavioural diagnosis of autism. *PlOs One 2012; 7:e43855*.

[39] 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), 1271-1280.

[40] Witten, I. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*.

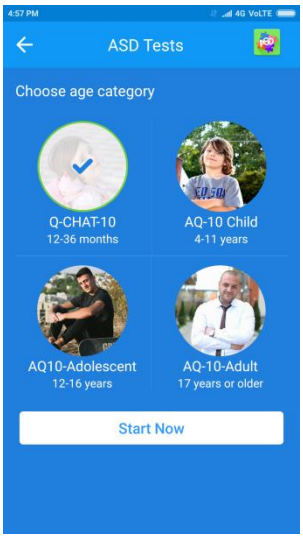


Fig 1A: Age selection screen

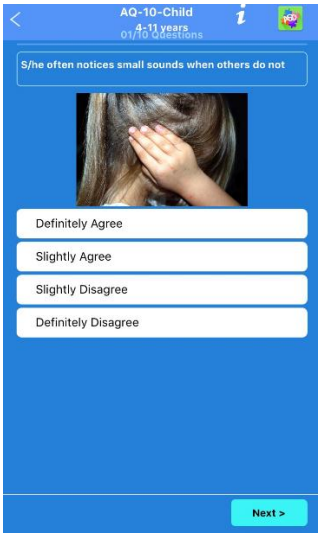


Fig 1B: A sample question: child

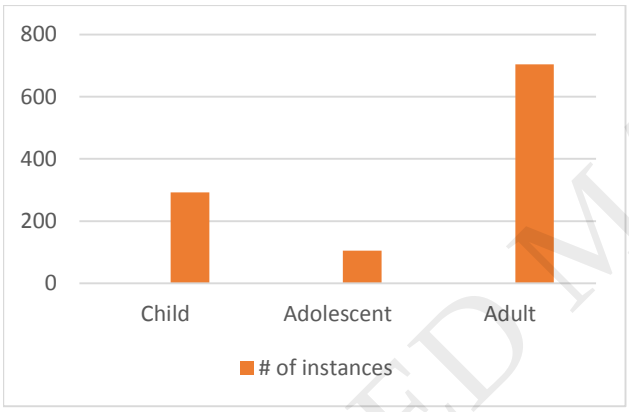


Fig. 2A the distribution of instances per age group (screening method target audience)

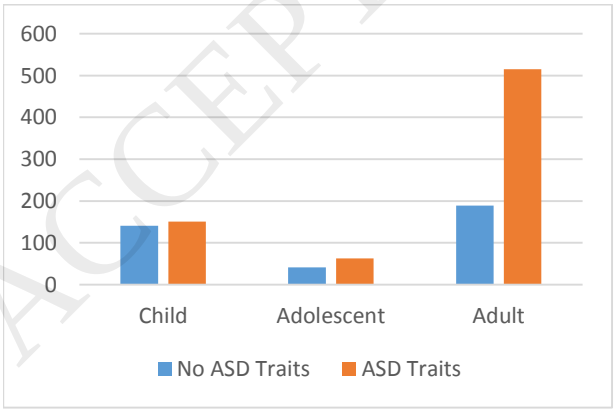


Fig. 2B the distribution of age instances per class label

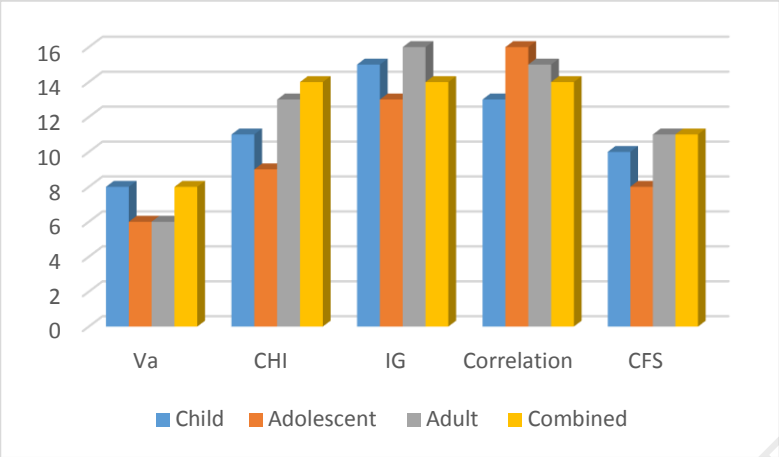
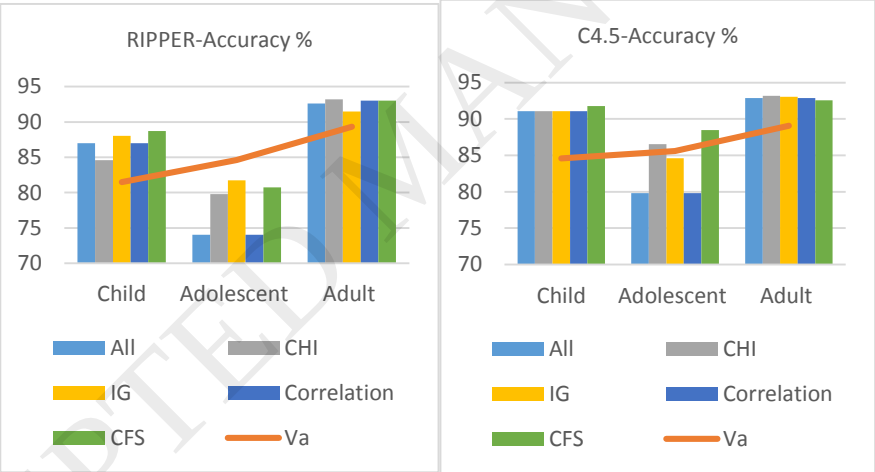
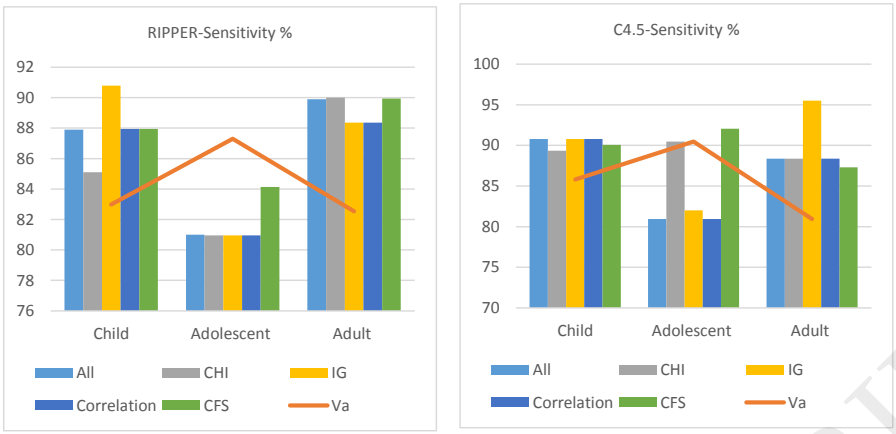


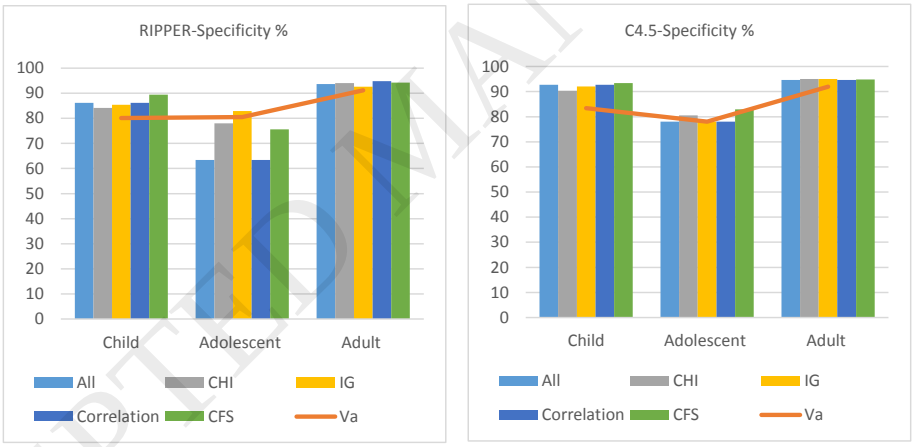
Fig. 3 number of variables selected from the ASD dataset based on the considered filtering methods



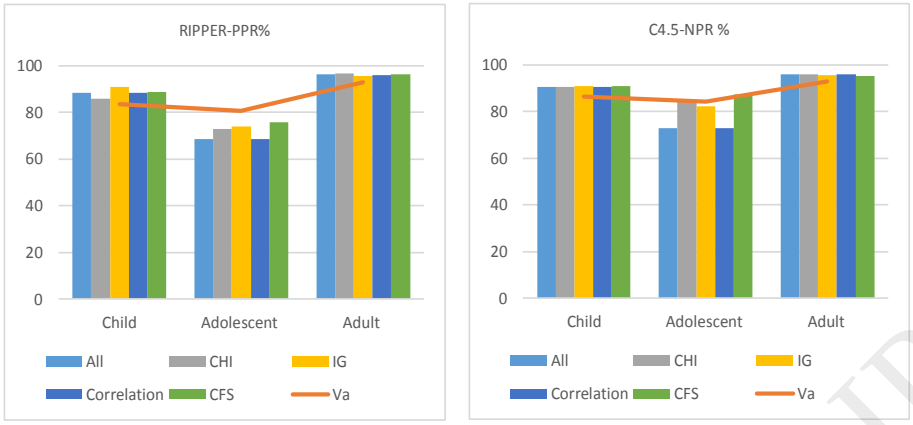
Figs. 4A & 4B Classification accuracies of RIPPER and C4.5 algorithms against the selected subsets of data of the considered methods



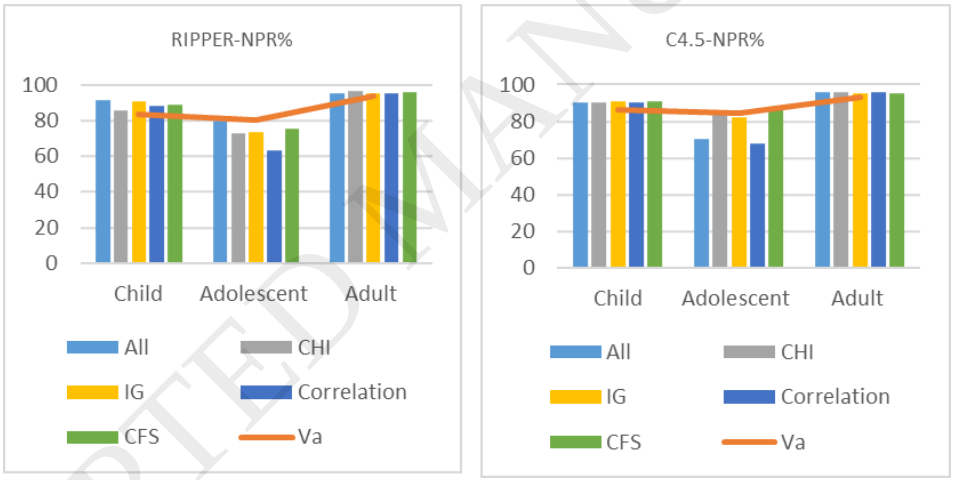
Figs 5A & 5B sensitivity rates of RIPPER and C4.5 algorithms against the selected subsets of data of the considered methods



Figs 6A & 6B Specificity rates of RIPPER and C4.5 algorithms against the selected subsets of data of the considered methods



Figs 7A & 7B PPV rates of RIPPER and C4.5 algorithms against the selected subsets of data of the considered methods



Figs 8A & 8B NPV rates of RIPPER and C4.5 algorithms against the selected subsets of data of the considered methods

Table 1A: Features collected and their descriptions

Feature	Type	Description
Age	Number	Toddlers (months), children, adolescent, and adults(year)
Gender	String	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean (yes or no)	Whether the case was born with jaundice
Family member with PDD	Boolean (yes or no)	Whether any immediate family member has a PDD
Who is completing the test	String	Parent, self, caregiver, medical staff, clinician ,etc.
Country of residence	String	List of countries in text format
Used the screening app before	Boolean (yes or no)	Whether the user has used a screening app
Screening Method Type	Integer (1,2,3)	The type of screening methods chosen based on age category (1=child, 2= adolescent, 3= adult)
A1: Question 1 Answer	Binary (0, 1)	See Table 1B for details
A2: Question 2 Answer	Binary (0, 1)	See Table 1B for details
A3: Question 3 Answer	Binary (0, 1)	See Table 1B for details
A4: Question 4 Answer	Binary (0, 1)	See Table 1B for details
A5: Question 5 Answer	Binary (0, 1)	See Table 1B for details
A6: Question 6 Answer	Binary (0, 1)	See Table 1B for details
A7: Question 7 Answer	Binary (0, 1)	See Table 1B for details
A8: Question 8 Answer	Binary (0, 1)	See Table 1B for details
A9: Question 9 Answer	Binary (0, 1)	See Table 1B for details
A10: Question 10 Answer	Binary (0, 1)	See Table 1B for details
Scoring Result	Integer	See Table 1B for details

Table 1B: Details of variables mapping to the screening methods

Variable	Corresponding AQ-10-Adult Features	Corresponding AQ-10-Adolescent Features	Corresponding AQ-10-Child Features
A1	I often notice small sounds when others do not	S/he notices patterns in things all the time	S/he often notices small sounds when others do not
A2	I usually concentrate more on the whole picture rather than the small details	S/he usually concentrates more on the whole picture rather than the small details	S/he usually concentrates more on the whole picture rather than the small details
A3	I find it easy to do more than one thing at once	In a social group, s/he can easily keep track of several different people's conversations	In a social group, s/he can easily keep track of several different people's conversations
A4	If there is an interruption, I can switch back to what I was doing very quickly	If there is an interruption, s/he can switch back to what s/he was doing very quickly	S/he finds it easy to go back and forth between different activities
A5	I find it easy to 'read between the lines' when someone is talking to me	S/he frequently finds that s/he doesn't know how to keep a conversation going	S/he doesn't know how to keep a conversation going with his/her peers
A6	I know how to tell if someone listening to me is getting bored	S/he is good at social chit-chat	S/he is good at social chit-chat
A7	When I'm reading a story I find it difficult to work out the characters' intentions	When s/he was younger, s/he used to enjoy playing games involving pretending with other children	When s/he is read a story, s/he finds it difficult to work out the character's intentions or feelings
A8	I like to collect information about categories of things (e.g. types of car, types of bird, types of train, types of plant, etc)	S/he finds it difficult to imagine what it would be like to be someone else	When s/he was in preschool, s/he used to enjoy playing pretending games with other children
A9	I find it easy to work out what someone is thinking or feeling just by looking at their face	S/he finds social situations easy	S/he finds it easy to work out what someone is thinking or feeling just by looking at their face
A10	I find it difficult to work out people's intentions	S/he finds it hard to make new friends	S/he finds it hard to make new friends

Table 2: Sample 20 data instances collected for Children using ASDTests app based on AQ-10 Child screening method

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Age	Sex	Ethnicity	Jundice	Family with autism	Contry_of_res	used_app_before	Scoring result	Who is teken the test	Class /ASD
1	1	0	0	1	1	0	1	0	0	6	m	Others	no	no	Jordan	no	5	Parent	NO
1	1	0	0	1	1	0	1	0	0	6	m	Middle Eastern	no	no	Jordan	no	5	Parent	NO
1	1	0	0	0	1	1	1	0	0	6	m	Middle Eastern	no	no	Jordan	yes	5	?	NO
0	1	0	0	1	1	0	0	0	1	5	f	Middle Eastern	yes	no	Palestine	no	4	?	NO
1	1	1	1	1	1	1	1	1	1	5	m	Others	yes	no	USA	no	10	Parent	YES
0	0	1	0	1	1	0	1	0	1	4	m	?	no	yes	Egypt	no	5	?	NO
1	0	1	1	1	1	0	1	0	1	5	m	White-European	no	no	USA	no	7	Parent	YES
1	1	1	1	1	1	1	1	0	0	5	f	White-European	no	no	New Zealand	no	8	Parent	YES
1	1	1	1	1	1	1	0	0	0	11	f	Middle Eastern	no	no	Bahrain	no	7	Parent	YES
0	0	1	1	1	0	1	1	0	0	11	f	?	no	yes	Austria	no	5	?	NO
1	0	0	0	1	1	1	1	1	1	10	m	White-European	yes	no	UK	no	7	Self	YES
0	1	0	0	1	0	0	0	0	1	5	f	?	no	no	Palestine	no	3	?	NO
0	1	1	1	1	1	1	1	1	1	4	m	White-European	yes	no	USA	no	9	Parent	YES
1	0	0	0	0	0	1	0	0	0	4	f	Black	no	no	USA	no	2	Parent	NO
1	1	1	1	1	1	1	1	1	1	6	m	White-European	no	no	UK	no	10	Parent	YES
1	1	1	1	1	1	1	1	1	1	8	m	White-European	no	no	New Zealand	no	10	Parent	YES
1	1	1	1	1	1	0	1	1	1	4	m	South Asian	no	no	India	no	9	Parent	YES
0	0	0	0	0	0	1	0	0	0	7	m	Others	no	no	USA	no	1	Parent	NO
1	0	1	1	1	0	1	1	1	1	11	m	White-European	no	yes	USA	no	8	Parent	YES
1	1	1	1	1	1	0	1	0	1	5	m	White-European	no	no	Australia	no	8	?	YES

Table 3: Confusion matrix for ASD screening problem

	Predicted Class Value	
Actual Class Value	**ASD**	**No-ASD**
ASD	True Positive (TP)	False Negative (FN)
No-ASD	False Positive (FP)	True Negative (TN)

Table 4a Features remained along with their scores on the AQ-Adult dataset after applying Va

Adult Dataset (AQ-Adult)		**Item description in the screening method**
Score	**Attribute**	**Description**
1.414	Item 9	I find it easy to work out what someone is thinking or feeling just by looking at their face
1.204	Item 6	I know how to tell if someone listening to me is getting bored
1.097	Item 5	I find it easy to 'read between the lines' when someone is talking to me
0.816	Item 4	If there is an interruption, I can switch back to what I was doing very quickly
0.708	Item 3	I find it easy to do more than one thing at once
0.559	Item 10	I find it difficult to work out people's intentions

Table 4b Features remained along with their scores on the AQ-Adolescent dataset after applying Va

Adult Dataset (AQ-Adult)		**Item description in the screening method**
Score	**Attribute**	**Description**
1.414	Item 5	S/he frequently finds that s/he doesn't know how to keep a conversation going
1.265	Item 4	If there is an interruption, I can switch back to what I was doing very quickly
1.174	Item 3	In a social group, s/h can easily keep track of several different people's conversations
1.174	Item 10	S/he finds it hard to make new friends
0.974	Item 6	S/he is good at social chit-chat
0.841	Item 8	S/he finds it difficult to imagine what it would be someone else
0.787	Item 9	S/he finds social situations easy
0.525	Item 7	When s/he was younger, s/he used to enjoy playing games involving pretending with other children

Table 4c Features remained along with their scores on the AQ-Child dataset after applying Va

Adult Dataset (AQ-Adult)		**Item description in the screening method**
Score	**Attribute**	**Description**
1.4142	Item 4	S/he finds it easy to go back and forth between different activities
1.0211	Item 9	I find it easy to work out what someone is thinking or feeling just by looking at their face
0.8586	Item 10	S/he finds it hard to make new friends
0.8269	Item 8	When s/he was in preschool, s/he used to enjoy playing games involving pretending with other children
0.7651	Item 6	S/he is good at social chit-chat
0.6899	Item 3	In a social group, s/h can easily keep track of several different people's conversations
0.6692	Item 1	She often notices small sounds when others do not
0.6339	Item 5	S/he doesn't know how to keep a conversation going with his/he peers

Table 5 relative reduction of the variables selected in % from the ASD datasets based on the considered filtering methods versus Va

Dataset	Screening Method	Va-CHI	Va-IG	Va-Correlation	Va-CFS
Child	AQ-Child-10	27.3%	46.7%	38.5%	20.0%
Adolescent	AQ-Adolescent-10	11.1%	38.5%	50.0%	0.0%
Adult	AQ-Adult-10	53.8%	62.5%	60.0%	45.5%