Thabtah *et al. Health Inf Sci Syst* (2019) 7:12
https://doi.org/10.1007/s13755-019-0073-5

Health Information Science and Systems

RESEARCH

A machine learning autism classification based on logistic regression analysis

Fadi Thabtah[1*], Neda Abdelhamid[2] and David Peebles[3]

Abstract

Autistic Spectrum Disorder (ASD) is a neurodevelopmental condition associated with significant healthcare costs; early diagnosis could substantially reduce these. The economic impact of autism reveals an urgent need for the development of easily implemented and effective screening methods. Therefore, time-efficient ASD screening is imperative to help health professionals and to inform individuals whether they should pursue formal clinical diagnosis. Presently, very limited autism datasets associated with screening are available and most of them are genetic in nature. We propose new machine learning framework related to autism screening of adults and adolescents that contain vital features and perform predictive analysis using logistic regression to reveal important information related to autism screening. We also perform an in-depth feature analysis on the two datasets using information gain (IG) and Chi square testing (CHI) to determine the influential features that can be utilized in screening for autism. Results obtained reveal that machine learning technology was able to generate classification systems that have acceptable performance in terms of sensitivity, specificity and accuracy among others.

Keywords: Autism spectrum disorder, Classification, Clinical decision making, Data mining, Feature analysis, Machine learning, Sensitivity, Specificity

Introduction

ASD is a life-long complex neurodevelopmental disorder characterised by impairments in the development of socio-communicative skills, cognitive abilities and by repetitive or restricted behaviours and interests [4]. The symptoms of autism are more visible and easy to identify in children aged two to three years. According to Towle and Patrick [32], one out of every 68 children has autism. Consequently, various screening methods have been developed globally by medical experts and psychiatrists seeking to identify autistic traits in their primitive stage so as to readily provide the necessary medications [3].

ASD can be formally diagnosed by specialised physicians within a medical unit using a diagnostic method such as the Autism Diagnostic Interview (ADI) [18]. The process of formally diagnosing ASD is time consuming [7, 27] as it requires time to be allowed for:

a. Training
b. Administration (asking a large number of questions)
c. Scoring and consensus coding

To expedite the referrals of individuals exhibiting autistic symptoms for further evaluation, self-administered screening methods have been developed primarily based on questionnaires, e.g. Autistic Quotient (AQ), Social Responsiveness Scale and Australian Scale for Asperger Syndrome (ASAS) [6, 10, 13]. Lessening the diagnostic time and minimising the number of items used during the diagnosis process is essential, especially now after the rapid development in the smart phone industry. This technology would enable individuals and their parents, caregivers and teachers to access screening tools using smart devices and to receive instant results and hence faster medical referral.

One possible way to improve efficiency and efficacy of existing ASD screening methods is to adopt intelligent solutions based around machine learning [1, 8, 20–22, 31]. This approach necessitates sufficient instances of cases and controls to construct autism detection systems

*Correspondence: Fadi.fayez@manukau.ac.nz
[1] Digital Technologies, Manukau Institute of Technology, Auckland, New Zealand
Full list of author information is available at the end of the article

that can be embedded within the screening method. However, historical data related to behavioural science applications, particularly autism, is rare posing a key challenge in improving ASD screening and reducing false positive and false negative rates [25]. Presently, few autism datasets associated with clinical diagnosis are available and are mostly genetic in nature, e.g., AGRE [14], National Database of Autism Research (NDAR) [15] and Boston Autism Consortium (AC) [12] but there is no behaviour data for screening of ASD.

To overcome the above challenges, we propose in this paper a machine learning framework with two datasets related to autism research that hold behavioural characteristics. The proposed datasets are based on the AQ-10 adult and AQ-10 adolescent screening methods respectively [3]. Each dataset consists of over 20 variables, ten of which are associated with the screenings plus the individual's features such as age, gender, ethnicity, etc. The datasets are anonymous and have been collected using a recently developed mobile application called ASDTests [26]. In this research, predictive and feature analyses have been conducted on the datasets to pinpoint the best influential features for autism screening of adults and adolescents. The feature analysis was performed using information gain (IG) and chi-square testing (CHI) methods [23, 19] in which a few effective features of autism have been detected ("Results analysis" section gives further details). Furthermore, a predictive analysis using a machine leaning algorithm called Logistic Regression was conducted. The purpose of the machine learning analysis was to obtain sensitivity, specificity and predictive accuracy on the results of the feature selection methods. By developing the new datasets and performing feature and predictive analyses the below distinctive advantages are gained:

(1) Valuable instances related to adult and adolescent cases and controls are now available for further analysis by researchers to improve ASD screening
(2) New features based on computational intelligence methods (IG and CHI) to indicate autistic traits are provided to autism researchers
(3) True performance of the screening with respect to different evaluation metrics are obtained using various different features subsets
(4) Parents, caregivers, special education teachers in schools, and medical clinics, among others, are aware of the most influential features in the ASD screening process.

This paper is structured such that "Screening methods and quick review" section discusses the autism screening methods used and the tool used to collect the datasets, while "The machine learning framework" section presents the proposed datasets along with features and characteristics; "Results analysis" section highlights the results analysis. Finally, conclusions are provided in "Conclusions" section.

Screening methods and quick review

Autistic quotient (AQ)

The datasets proposed are based on a screening method developed by Baron-Cohen et al. [6] and called AQ. AQ was developed with the intention of detecting discernible features connected to Asperger Syndrome in adults with average intelligence levels. The AQ is a self-screening instrument with 50 items covering social aptitude, cognitive functioning, detail-orientation and social communications skills. Each item is measured on a four Likert-type ordinal scale ranging from Definitely Agree, Slightly Agree to Slightly Disagree and Definitely Disagree. A total instrument-based score results from an additive scaling procedure ranging from 0 to 50 with higher scores corresponding to higher possession of intellectual development deficits.

Baron-Cohen et al. [6] indicated that a cut-off score of 32 on the AQ is relevant and anyone receiving that score or higher is considered intellectually challenged. Auyeung [5] extended the AQ to be applied in various new settings such as on adolescents and children of various ages, backgrounds and contexts. For instance, two versions of AQ have emerged, one for children ranging in age between 4 and 11 years old, and one for adolescents ranging between 12 and 15 years of age. Most AQ variants require approximately 20–30 min to be completed and contain about 50 items. AQ-child enjoys higher validity and reliability psychometric properties compared with other versions. Auyeung [5] reported adequate sensitivity, as well as specificity metrics for AQ, at 77 and 74% respectively.

Allison et al. [3] created the AQ 10-adult and AQ 10-child, shortened versions of the original AQ, to facilitate the tool's clinical application across various settings. This attempt is said to increase the efficiency of the screening. Validation analyses yielded similar sensitivity and specificity measures for those shortened versions, similar to the original AQ. Each question on the shortened versions is worth one single point. Positive answers, Definitely Agree or Slightly Agree, receive a point in questions 1, 7, 8 and 10. If the respondent answered Definitely Disagree or Slightly Disagree, a point will be added to questions 2, 3, 4, 5, 6 and 9. A score of six or higher is considered to be clinically relevant and representing autism or intellectual development disorders.

Recent machine learning approaches to autism detection

Current ASD screening tools generally employ domain experts rules and scoring functions to classify cases and

controls. Psychiatric and behavioural science specialists have designed these rules, and the quality of outcomes and decisions depends substantially on the subjective contributions of these professionals and the interpretations of the specialised clinical staff conducting the assessments. Instead, the diagnosis of ASD might be empowered by automated decisions generated by intelligent algorithms such as machine learning. To date, there are no self-administered ASD diagnostic methods that have integrated machine learning models into the process, despite a few research attempts on doing so [8, 9, 11, 28, 29, 33, 34].

Wall et al. [33, 34] investigated the potential use of outcomes based on machine learning algorithms to assist clinicians conduct ADOS-R (Module 1) diagnosis method. The authors claimed based on the results obtained by using different machine learning techniques that ADOS-R (Module 1) items can be replaced with just eight items (common features found in the machine learning classifiers). Therefore, the efficiency of conducting ADOS-R (Module 1) can be significantly improved. However, a later research by Bone et al. [9] revealed serious pitfall in the methodology and implementation of the studies conducted by Wall et al. [33, 34]. To be exact, no saving time related to administration can be obtained simply because the researchers must use the full items in ADOS-R (Module 1) before applying the machine learning technique. Moreover, the whole experimental setup of Wall et al. [33, 34] was not conducted in clinical set up and without having a domain expert or a licensed clinician to verify the results obtained. More importantly, Bone et al. [9] and Thabtah [27, 30] showed that Wall et al. [33, 34] studies have not considered integrating machine learning within ADOS-R diagnosis methods rather the authors just applied in a conventional way a number of machine algorithm on static dataset related to autism. Thus, if the dataset characteristics change the results will indeed change and therefore such results cannot be generalised.

Duda et al. [11] investigated the applicability of six data mining algorithms to detect Attention Deficit Hyperactivity Disorder (ADHD) and ASD from a dataset with over 2900 instances. The aim of the study was to reduce the number of items required to come up with a diagnosis using the Social Responsiveness Scale (SRS) conventional diagnostic method. The authors claim that six items found by different machine learning techniques can be effective in detecting ADHD and ASD in the SRS method and therefore the SRS complete set of items can just be replaced with these six items. However, in this study, hard to predict instances have been discarded prior applying the machine learning techniques. In addition, there was no clear methodology how the discovered items can be utilized as a screening method and under which conditions.

The machine learning framework

Data collection

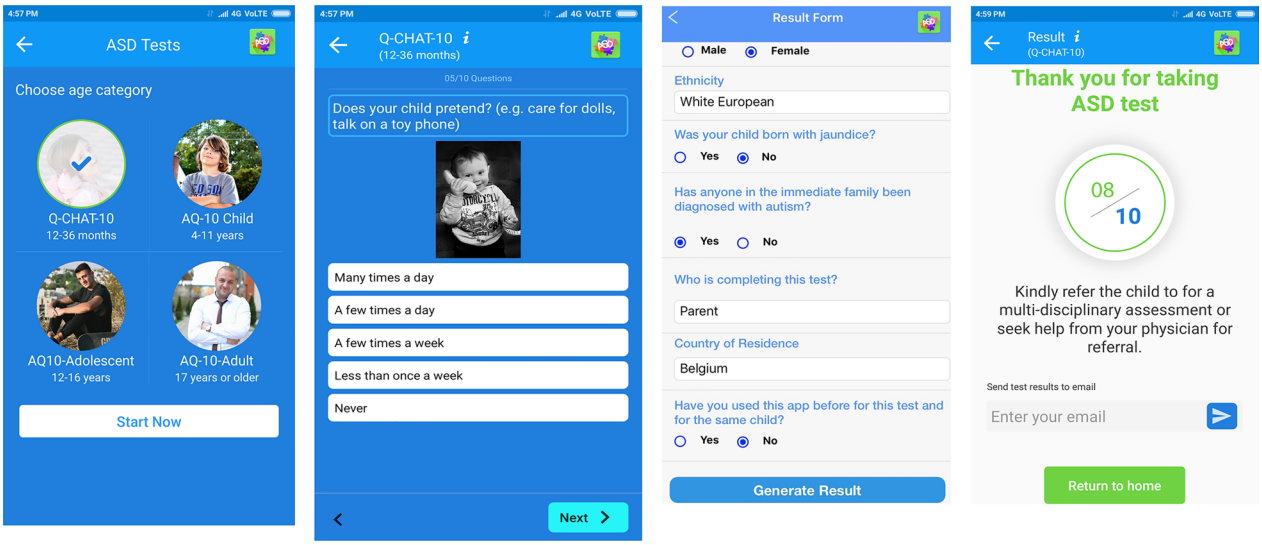
The instances (cases and controls) have been collected using a mobile application for autism screening called ASDTests [26]. This app was designed and implemented in 2017 and it contains four primary screening methods (Q-CHAT-10, AQ-10-child, AQ-10-adolescent, AQ-10-adult) [3, 24] to accommodate the target audience (toddlers, children, youths and adults) as displayed in Fig. 1. The ASDTests app is available online for free download in both IOS and Android versions.

Initially, the user from the initial screen (Fig. 1a) selects the screening type based on the age category. Each type of screening consists of ten sequential questions each of which is displayed on a separate screen and is associated with an image to enable users to carefully select the appropriate answer (Fig. 1b). 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 displays one sample question from the toddler test. When the user completes and reviews the questions then a submit screen appears (Fig. 1c). In the app information screen, a consent of data usage for research purpose besides other fields for data recording are provided to the users; participants can either choose to contribute or not. Once the user submits after undergoing the tests, a result screen appears to pinpoint the score computed and a textual interpretation of the score (Fig. 1d). For instance, if an adult has completed the screening and obtained a score less than six then the result will state "no autistic traits are found" otherwise "please consider seeing a medical specialist for further assessment". It should be noted that scores are calculated per screening type in an automated manner in the app and based on the handcrafted rules offered in each designated screening method. For further details on score calculations please refer to [3].

Before completing the screening, users were asked to consent to a disclaimer which explained the goal of the research, privacy policy, and use of the data. Users were informed that their data would be kept anonymous and only shared for research purposes. The users had to read this before submitting their answers.

The machine learning framework for ASD screening

Figure 2 shows the machine learning framework for the autism classification problem. In the framework, whenever a test case (an individual) undergoes the screening process, the machine learning method will assign the appropriate class label to the test case based on the recommended class given by the Logistic Regression model. Several different users can exploit the ASDTests



a Screening method screen (Thabtah, 2017b) **b** A sample question toddler (Thabtah, 2017b) **c** Data collection screen (Thabtah, 2017b) **d** Results screen (Thabtah, 2017b)

Fig. 1 **a** Screening method screen [26]. **b** A sample question toddler [26]. **c** Data collection screen [26]. **d** Results screen [26]

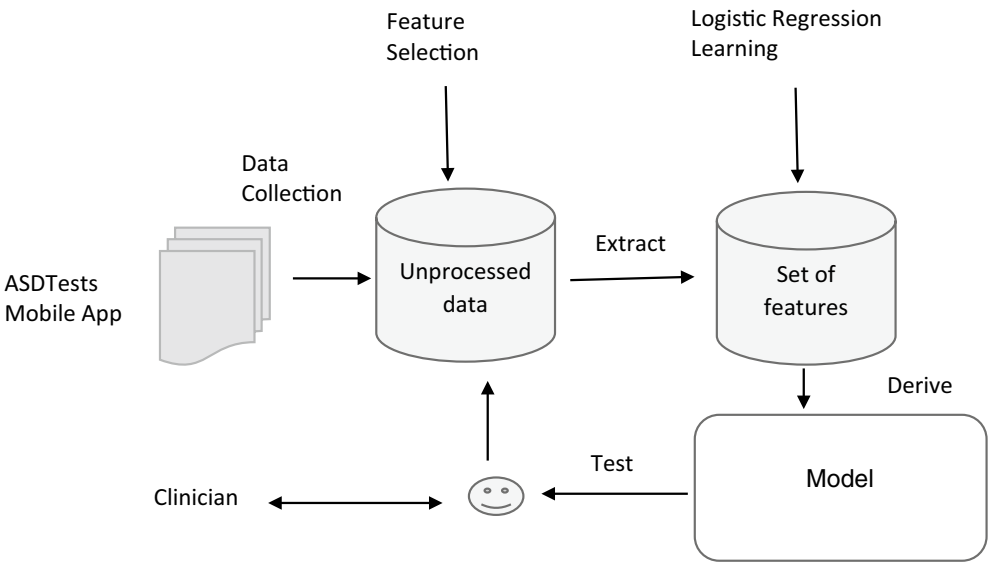


Fig. 2 The proposed machine learning framework for autism screening

app including clinicians, parents, care givers and medical staff.

The results might suggest that the individual (toddler, child, adolescent or adult) undertake a more rigourous screening for autism. Every time a screening process occurs it gets added into a training dataset on the secured cloud where the screening method embedded in the app assigns a true class (ASD traits/No ASD traits) to the case in an automated manner. The raw dataset contains over 20 variables (including the class variable) of which ten are screening questions based on AQ short versions [3].

When the raw data are extracted, several pre-processing techniques were applied, including discretization of continuous variables (age), missing values replacement and transformation of the screening questions into binary representation (more details on the data transformation

are given in "The datasets and features" section). A feature selection process is employed to assess the variables in the training dataset using filter methods in order to determine redundant and useless features, so they can be discarded. In addition, the feature selection process will identify influential features that can be offered to the machine learning algorithm during the training phase. We have adopted Information Gain (IG) and Chi-Square Testing methods for the process of feature analysis [19, 23] ("Results analysis" section gives more details on the results related to feature selection).

Once the set of influential features are identified then a logistic regression algorithm is utilised to learn classification model for detecting autism traits [17]. This algorithm utilizes the Ridge estimator multinomial logistic regression to build classifiers. When a dataset contains c number of classes for m data cases with n variables, the parameter matrix can be computed as n*(c−1) matrix. The likelihood for class j with the exception of the last class is calculated as

$$\mathrm{Pj(Xi)} = \exp\big(\mathrm{XiBj}\big) / \big(\big(\mathrm{sum[j} = 1 \ldots (\mathrm{k}-1)]\exp\big(\mathrm{Xi} * \mathrm{Bj}\big)\big) + 1\big) \quad (1)$$

Originally logistic regression is used for data analysis in statistics in where it determines the relationships between one or more independent variables and a dependent variable. Typically, when this method is used for prediction problems, the input dataset contains two possible values for the dependent variable (target class). The aim is to model the relationships between the independent variables and the class using some logic function probabilities as described in Eq. (1). More details on how the classification is performed using multinomial logistic regression can be found in [17].

Whenever a test case (individual) undergoes a screening, the logistic regression model in our framework allocates the right class to the individual using the input variables values (independent variables values). This is instead of using the scoring function embedded in the screening method which was designed by a domain expert. The proposed framework replaces the scoring function with a more accurate model learnt from former cases and controls who had undergone screening and already have been classified. The relationships between these and the dependent variable (ASD traits/No ASD traits) can be discovered and exploited to make more accurate detection of ASD traits during the screening process. In addition, in the proposed machine learning framework, the validity of the test can be solely placed at the hand of the medical experts and clinicians to verify the decision when needed. Hence, this framework not only improves the accuracy of autism screening but in addition it indeed helps in speeding up the referral process for a formal autism diagnosis procedure. Consequently, individuals with autism and their family can have access to appropriate medical resources at earlier stage if we consider that waiting times for autism diagnosis is lengthy.

The datasets and features

Table 1 shows the primary features related to the screening method and individual features related to the users. A special feature called the target class variable has been created to determine whether the individual undergoing the test has ASD traits or not. The class value is assigned automatically by the ASDTests app based on the final score obtained from the individual taking the ASD screening. For example, if the individual has selected an age category of 12–16 (AQ-10-adolescent) when using the ASDTests app, the scoring will be based on the AQ-10-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" indicates 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 can be used for data analysis in order to understand key features that may influence ASD screening from a behavioural perspective. All bold features have been ignored during data processing (See "Results analysis" section for further details).

The A1–A10 variables (Table 1) have been transformed into either 0 or 1 depending on the true answers given by the users during the screening. In particular, for the AQ-10-Adolescent, 1 was given to questions 1, 5, 8 and 10 if the given answer was Slightly Agree or Definitely Agree for each, whereas 1 was given to Definitely or Slightly Disagree answers on the remaining questions. For the AQ-10-Adult method, 1 was given for Slightly Agree or Definitely Agree responses for questions 1, 7, 8, and 10. For the rest of the questions 1 was allocated when Definitely or Slightly Disagree was given for questions 2, 3, 4, 5, 6, or 9. The binary representation for the features in the dataset can ease the process of data mining by the learning algorithms

Table 2 shows sample data instances that have been collected based on the AQ-10-adolescent screening. For the adult and adolescent datasets, 1118 and 249 instances were collected respectively over a period of six months using the ASDTest app. After an initial investigation on the collected instances in the adult dataset, it was clear that the vast majority of the instances belonged to the "no ASD" class, making such a group of data imbalanced. To be exact, 68.3% of the adult individuals who underwent the screening were not associated with ASD traits;

Table 1 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 (0,1,2,3)	The type of screening methods chosen based on age category (0 = toddler, 1 = child, 2 = adolescent, 3 = adult)
Language	String	(English, Arabic, Farsi, Mandarin, Urdu, Swahili, French, Spanish, Portuguese, Turkish)
Why_are_you_taken_the_screening	String	Use input textbox
Question 1 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 2 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 3 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 4 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 5 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 6 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 7 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 8 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 9 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 10 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Screening score	Integer	The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner
Class	String	ASD traits or No ASD traits (automatically assigned by the ASDTests app).

this seems appropriate considering that more people will normally be without autism symptoms. However, in the adolescent dataset, the number of instances that are linked with ASD traits was 127 out 249. This means that the adolescent dataset is balanced with respect to class label. Looking further into the instances with ASD symptoms in the adult and adolescent datasets, it was a surprise to reveal that the majority of them were female, e.g. 69 in the adolescent and 185 in the adult. Among the 249 instances in the adolescent dataset, the majority have taken the screening by themselves (99) or by parents (103). For the adult dataset, the number of individuals who had taken the screening by themselves constitutes 82.28%. Figure 3a, b show the distribution of instances with respect to class labels for the adult and adolescent datasets respectively.

Looking at other variables such as ethnicity, gender, family siblings with ASD, and country of residence, we discovered that the highest participated ethnicity was Caucasian followed by Middle Eastern then Asian for both the adult and adolescent datasets. Furthermore, the majority of the participants in both adults and adolescent screening tests resided in the United States and United Kingdom. There were 110 and 35 instances for the adults and adolescents respectively who had been born with jaundice. Among those 110 and 35 instances there were 48 and 20 who had actually been screened with ASD symptoms by the AQ-10-adult and AQ-10-adolescent methods respectively.

Moreover, there were 183 and 44 individuals who had family siblings diagnosed with ASD in the adult and adolescent datasets respectively. Among those 183 and 44 instances there were 88 and 22 who had been screened with ASD traits by AQ-10-adult and AQ-10-adolescent methods respectively. The average age in years for adolescents and adults in the datasets is 14.04 and 30.14 respectively, and the standard deviation for the age variable for the adolescent and adult datasets is 1.48 and 10.49 respectively. Finally, the adult and adolescent datasets contain 596 and 117 male participants and 522 and 131 female participants respectively.

Table 2 Sample 15 data instances collected for children using ASDTests app based on AQ-10 child screening method

Case No	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Age	Sex	Ethnicity	Jaundice	Family with ASD	Residence	Used_App_Before	Why taken the screening	Score	Screening Type	Language	User	Class
1	1	0	1	1	1	1	0	1	1	1	16	f	White	No	Yes	Andorra	No		8	Adolescent	English	Self	YES
2	1	1	1	0	0	0	0	0	0	0	15	m	Arab	No	No	UK	No		3	Adolescent	English	Self	NO
3	1	1	1	1	1	0	1	1	1	1	16	f	White	Yes	Yes	Estonia	No		9	Adolescent	English	Parent	YES
4	1	0	0	1	1	1	1	0	1	1	12	m	White	No	No	South Africa	No		7	Adolescent	English	Parent	YES
5	1	1	0	0	1	0	0	1	1	1	14	m	White	No	No	USA	No		6	Adolescent	English	Self	NO
6	0	1	0	1	1	1	0	0	1	1	14	m	White	No	No	USA	No	Check ASD	6	Adolescent	English	Health care professional	NO
7	1	0	0	0	1	0	0	1	0	1	14	m	White	No	No	USA	No		4	Adolescent	English	Health care professional	NO
8	0	0	0	1	1	0	0	1	0	0	15	f	Arab	No	No	Palestine	No		3	Adolescent	English	Parent	NO
9	1	1	1	1	1	0	1	0	0	1	12	m	White	Yes	Yes	UK	No		7	Adolescent	English	Parent	YES
10	0	0	0	1	0	0	1	1	0	0	15	m	Arab	No	No	Jordan	No		3	Adolescent	English	Parent	NO
11	1	0	0	0	0	0	0	0	0	1	16	f	Arab	No	No	New Zealand	No		2	Adolescent	English	Parent	NO
12	1	1	1	0	1	1	0	1	1	1	16	f	Native Indian	No	No	India	No	Check autism	8	Adolescent	English	Self	YES
13	1	0	1	1	1	1	1	1	1	1	13	m	White	No	No	UK	No		9	Adolescent	English	Parent	YES
14	0	0	1	1	0	1	1	1	1	1	12	f	Persian	No	No	Iran	No		7	Adolescent	English	Parent	YES
15	1	0	0	0	1	0	0	1	0	0	13	m	White	No	No	UK	No	Becaz	3	Adolescent	English	Parent	NO

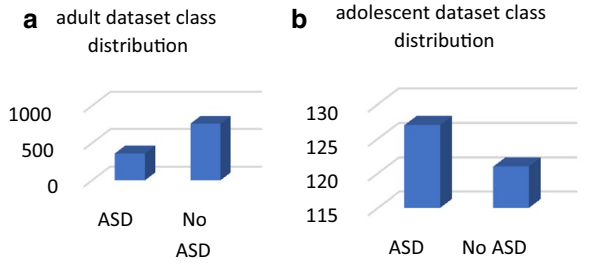


Fig. 3 **a**, **b** Class distribution for adult and adolescent datasets respectively

Results analysis

Settings and methods used

In this section, the feature analysis is presented based on the autism datasets (adolescent, adult) in order to assess which autistic traits have more influence on ASD screening. To achieve the aim, we apply CHI and IG feature selection methods and seek similarities and differences in the feature sets offered by these methods. The key is to determine a few yet effective features that can assess the different users and understand symptoms that red flag autism detection. Reasons behind choosing these feature selection methods are twofold:

(a) Different correlation metrics are employed for computing the scores of the available features
(b) Scores are ranked so influential features can be distinguished

All empirical runs on the autism dataset have been conducted on an open source Java platform named WEKA version 3.9.1 [16]. To build the classifiers we employed a logistic regression algorithm developed by Le Cessie and van Houwelingen [17] and embedded in WEKA. WEKA is a data analytics tool that holds large collections of pre-processing filters, supervised learning techniques, unsupervised learning techniques and visualization techniques, among others. In constructing the classifiers, the ten-fold cross validation method was used [35]. This testing method is usually employed in the training phase by the learning algorithms to avoid over-fitting. In using ten-fold cross validation, the input dataset is partitioned randomly into ten subsets and the algorithm will train on nine parts and derive a classifier. This classifier is then tested on the remaining parts to reveal its performance, i.e. error rate. The same process is repeated on the training 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 produce one global error rate for the classifier. All experiments have been performed on a computing machine with 2.0 GHz processor and 8 RAM of memory.

Data processing results analysis

Prior feature selection, we discarded language and screening-type variables since they contribute little to data processing. In addition, we discarded the final-score variable since it may over-fit the classifier by generating 100% accuracy. We also removed, "why_are_you_taking_the_screening?" since it has over 98% missing value in both datasets. All missing values within other variables have been treated as any other value. The variables that belong to AQ-10 (adolescents, adults) have been converted into 0 and 1. To be exact, any answer of a question in the AQ-10 methods with "Slightly Agree" or "Definitely Agree" during the screening process of the ASD-Test app will be converted into 1, and any answer with "Slightly Disagree" or "Definitely Disagree" will be converted into 0 (Table 2). All input typing errors recorded by the users during the data collection via the ASDTests app have been corrected using WEKA filters. The age has also been discretised using an entropy filter in WEKA prior to data processing. The total number of variables remaining in the adult and adolescent datasets was 20.

Table 3 shows the features along with their rank after applying IG and CHI filtering methods. It is obvious from the results that CHI and IG produce consistent results despite having different feature extraction procedures. The cutoff points that separate high correlated features from those that are low are 0.05 and 15 for IG and CHI respectively. Based on the cutoff points, features highlighted in red in Table 3 are ignored since they are associated with low scores. The results obtained by the IG and CHI filtering methods are clearly clustered into different groups of features (influential, semi-influential and low influential) as per highlighted colours in the table. There are slightly more influential features derived from the adult dataset by CHI and IG, possibly since the adult dataset contains more instances for both ASD and No ASD class labels.

For the adult dataset, the top three features that are in common based on the features sets of CHI and IG are A6, A5, A9 and A4. These are items within the AQ-10 screening methods (adult, adolescent) (see Tables 4, 5). These features are related to social and communication behaviours and do not fully accommodate ASD criteria based on the DSM-5 autism criteria [2]. Additional influential features were derived by CHI from the adults' dataset such as A3 and place_of_residence. However, we believe that place_of_residence has little impact on the classification of ASD traits and it was selected by both filtering methods since it has large numbers of possible values.

Table 3 Results and scores generated by CHI and IG methods for the adults' and adolescents' datasets

Adolescent Features		Adult Features	
IG features rank	CHI features rank	IG features rank	CHI features rank
0.2948146 A6	91.69033 A6	0.27396 A6	423.539 A6
0.2627309 Residence	82.74345 A3	0.27355 A5	408.3409 A9
0.2578447 A3	72.29336 Residence	0.26949 A9	367.1237 A5
0.2217318 A4	71.32022 A4	0.18492 Residence	247.0266 A4
0.1878661 A5	60.19897 A5	0.176 A4	241.11 Residence
0.123656 A9	39.96247 A9	0.15086 A3	219.0755 A3
0.0975631 A10	32.46908 A10	0.1326 A10	182.2764 A10
0.0909212 A7	30.53027 A7	0.10534 7 A7	160.4521 A7
0.0874036 Ethnicity	29.1736 Ethnicity	0.08544 Ethnicity	121.2511 Ethnicity
0.0812569 A2	27.38534 A2	0.07273 A1	104.7512 A2
0.0681774 A1	22.90584 A1	0.06908 A2	96.3257 A1
0.0613579 A8	20.80029 A8	0.05259 A8	75.7954 A8
0.0254088 User	8.50241 User	0.01591 Family_ASD	25.9467 Family_ASD
0.0177077 Age	6.00852 Age	0.01189 Age	17.97 Age
0.0044191 Used_App_Before	1.47752 Used_App_Before	0.00974 User	12.7184 User
0.0016761 Jaundice	0.57416 Jaundice	0.00466 Jaundice	7.5612 Jaundice
0.0006911 Sex	0.23757 Sex	0.00448 Used_App_Before	7.5544 Used_App_Before
0.0000911 Family_ASD	0.03133 Family_ASD	0.00339 Sex	5.2588 Sex

Table 4 The mapping between features and items in the adult screening method

Feature	Description based on AQ-10-Adult screening method
A6	I know how to tell if someone listening to me is getting bored
A5	I find it easy to 'read between the lines' when someone is talking to me
A9	I find it easy to work out what someone is thinking or feeling just by looking at their face
A4	If there is an interruption, I can switch back to what I was doing very quickly

Table 5 The mapping between features and items in the adolescent screening method

Feature	Description based on AQ-10-Adolescent screening method
A6	S/he is good at social chit-chat
A3	In a social group, s/he can easily keep track of several different people's conversations
A4	If there is an interruption, s/he can switch back to what s/he was doing very quickly
A5	S/he frequently finds that s/he doesn't know how to keep a conversation going
A9	S/he finds social situations easy

Therefore, we discarded this feature from taking any role in the screening.

For the adolescent dataset, the top features related to autism that have been chosen were A6, A3, A4, A5 and A9. These features correspond to the items shown in Table 5 based on the AQ-10-adolescent screening method. It is clear that these features cover social and communication skills and hence partly fulfil the DSM-5 criteria for ASD diagnosis. It seems that the features chosen by CHI and IG related to adolescent are more focused on communication behaviours.

Table 6 depicts the sensitivity, accuracy and specificity rates derived by the Logistic Regression algorithm against subsets of the adult datasets chosen by IG and CHI filtering methods. The reported sensitivity, accuracy and specificity rates derived by the classifier from the complete number of features of the adult dataset excluding those highlighted in red in Table 1 are high. More interestingly, when the top 11 features selected by CHI filter are processed by the Logistic Regression algorithm the sensitivity, accuracy and specificity rates have been sustained without any drastic change in the performance. This is due to that all screening features of AQ-10-Adult are included in the final set offered by CHI. However, when we filtered out the features set of CHI to the top three features (A6, A9, A5, A4, A3, A10, A7) the rates of the evaluation metrics dropped by almost 5.5% upon processing these by the Logistic Regression algorithm. We investigated the top three common features of IG

Table 6 The accuracy, specificity and sensitivity rates against subsets of data for the adult dataset

Adult dataset	Accuracy	Sensitivity	Specificity
All-features	99.91	99.99	99.98
IG selected features			
(A6, A5, A9, A4, A3, A10)	90.51	90.50	93.02
(A6, A5, A9)	87.74	87.70	87.90
CHI selected features			
(A6, A9, A5, A4, A3, A10, A7, Ethnicity, A1, A2, A8)	99.91	99.99	99.99
(A6, A9, A5, A4, A3, A10, A7)	94.00	94.00	95.26

Table 7 The accuracy, specificity and sensitivity rates against subsets of data for the adolescent dataset

Adolescent dataset	Accuracy	Sensitivity	Specificity
All-features	97.58	97.60	95.86
IG selected features			
(A6, A5, A9, A4, A3, A10, A7)	92.33	92.30	90.02
(A6, A5, A9, A4, A3)	85.88	85.90	82.64
CHI selected features			
(A6, A3, A4, A5, A9, A10, A7)	92.33	92.30	90.02

and CHI on the adult dataset, i.e. (A6, A5, A9), Logistic Regression was able to produce classifiers with approximately 87% sensitivity, accuracy and specificity, which can be acceptable. These three features seem to be the most influential ones in adult screening.

For the adolescent dataset, the sensitivity, accuracy and specificity rates derived by the Logistic Regression against subsets of features chosen by IG and CHI are high yet lower than those derived by the same algorithm from the adult features subsets of the same filtering methods see Table 7). One reason of this could be that the adult dataset has more cases and controls which enabled the learning Regression against subsets of features chosen by IG and CHI are high yet lower than those derived by the same algorithm from the adolescent features subsets of the same filtering methods. One reason for this could be that the adult dataset has more cases and controls which enables the learning for the adolescent dataset, the sensitivity, accuracy and specificity rates derived by the Logistic algorithm to generate more accurate classifiers. The common features among IG and CHI features subsets are (A6, A9, A5, A4, A3). Logistic Regression generated classifiers with 85.88%, 85.9% and 82.64%. These rates are lower than the rates associated with the classifiers derived by Logistic Regression from the complete adolescent dataset.

Conclusions

The rapid growth in the number of ASD cases worldwide necessitates datasets related to behaviour traits. However, such datasets are rare making it difficult to perform thorough analyses to improve the performance of the screening. Presently, limited autism datasets associated with clinical diagnosis or screening are available and most of them are genetic in nature. Hence, we propose new machine learning framework with datasets related to the autism screening of adults and adolescents that have influential features and perform predictive analysis using Logistic Regression. In these datasets, we record ten behavioural features based on AQ (adults, adolescents) screening methods plus an individual's characteristics; these have proved to be effective in detecting the ASD cases from controls in behavioural science. We also perform an in-depth feature analysis on the two datasets using feature selection to determine the effective features that can be utilized in screening for autism. The feature analysis results reported that there are four effective features related to adult screening based on the AQ-10-Adult method (A4, A5, A6, A9) and five effective features related to adolescent screening based on the AQ-10-Adolescent method (A3, A4, A5, A6, A9). These chosen features are mainly concerned with communication and social behaviours. The Logistic Regression classifiers produced showed an acceptable level of sensitivity, accuracy and specificity rates based on the features

sets chosen by IG and CHI. Results also pinpointed that CHI and IG filter methods consistently derived common autistic features from both the adult and adolescent datasets respectively.

In conclusion, this research reported that machine learning technology specially function based ones (logistic regression) showed promising results in ASD screening at least for the adults and adolescents. In the near future, we intend to implement a new screening method using machine learning technology for toddlers and children.

Author details
[1] Digital Technologies, Manukau Institute of Technology, Auckland, New Zealand. [2] Information Technology, Auckland Institute of Studies, Auckland, New Zealand. [3] Department of Psychology, Centre for Applied Psychological Research, School of Human and Health Sciences, University of Huddersfield, Huddersfield, UK.

Received: 28 January 2019 Accepted: 22 May 2019
Published online: 01 June 2019

References

1. Abdelhamid N, Thabtah F, Abdel-jaber H. 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, 2017.
2. American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-5. Washington, D.C: American Psychiatric Association; 2013.
3. Allison C, Auyeung B, Baron-Cohen S. 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. J Am Acad Child Adolesc Psychiatr. 2012;51(2):202–17.
4. American Psychiatric Association (APA). *Diagnostic and statistical manual of mental disorders*. 5th ed. Arlington, VA: APA; 2013.
5. Auyeung BBC. The autism spectrum quotient: children's version (aq-child). J Autism Dev Disord. 2008;38(7):1230–40.
6. Baron-Cohen S, Wheelwright S, Skinner R, Martin J, Clubley E. The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. Journal of Autism Development Disorder. 2001;31:5–17.
7. Bishop D. *Definition, diagnosis & assessment in a history of autism by A. Feinstein*. Chichester: Wiley-Blackwell; 2010.
8. Bone D, Bishop S, Black M, Goodwin M, Lord C, Narayanan S. Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion. J Child Psychol Psychiatry. 2016;57:927–37.
9. Bone D, Goodwin M, Black M, Lee C, Audhkhasi K, Narayanan S. Applying machine learning to facilitate autism diagnostics: pitfalls and promises. J Autism Dev Disord. 2014;45(5):1–16.
10. Constantino J. *(SRS™) Social Responsiveness Scale*. WPS, 2005. https://www.wpspublish.com/store/p/2993/srs-social-responsiveness-scale. Accessed 9 Dec 2018.
11. Duda M, Ma R, Haber N, Wall DP. Use of machine learning for behavioral distinction of autism and ADHD. Transl Psychiatr. 2016;9(6):732.
12. Fischbach G, Lord C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron. 2010;68:192–5.
13. Garnett M, Attwood T. The Australian scale for Asperger syndrome. *Australian National Autism Conference*. Brisbane, Australia; 1995.
14. Geschwind D, et al. The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. Am J Hum Genet. 2001;69:463–6.
15. Hall D, Huerta MF, McAuliffe MJ, Farber GK. Sharing heterogeneous data: the national database for autism research. Neuroinformatics. 2012;10:331–9.
16. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I. The WEKA data mining software: an update. SIGKDD Explor. 2009;11(1):10–8.
17. Le Cessie S, van Houwelingen JC. Ridge estimators in logistic regression. Appl Stat. 1992;41(1):191–201.
18. Lord C, Rutter M, Le Couteur A. Autism diagnostic interview—revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. J Autism Dev Disord. 1994;24:659–85.
19. Liu H, Setiono R. 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.
20. Luo G. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. Health Inf Sci Syst. 2016;4(1):2.
21. Mohammad R, Thabtah F, McCluskey L. Intelligent rule-based phishing websites classification. IET Inf Secur. 2014;8(3):153–60.
22. Qabajeh I, Thabtah F, Chiclana F. Dynamic classification rules data mining method. J Manag Anal. 2015;2(3):233–53.
23. Quinlan J. Induction of decision trees. Mach Learn. 1986;1(1):81–106.
24. Robins D, Fein D, Barton M, Green J. The modified checklist for autism in toddlers: an initial study investigating the early detection of autism and pervasive developmental disorders. J Autism Dev Disord. 2001;31(2):131–44.
25. Thabtah F. Autism spectrum disorder screening: machine learning adaptation and DSM-5 fulfilment. Proceedings of the 1st International Conference on Medical and Health Informatics 2017, pp. 1–6. Taichung City, Taiwan, ACM; 2017.
26. Thabtah F. ASDTests. A mobile app for ASD screening. www.asdtests.com. Accessed November 30th, 2017.
27. Thabtah F. Machine learning in autistic spectrum disorder behavioral research: a review and ways forward. Inform Health Soc Care. 2018;43(2):1–20.
28. Thabtah F. An accessible and efficient autism screening method for behavioural data and predictive analyses. Health Inform J. 2018;19:1460458218796636. https://doi.org/10.1177/1460458218796636.
29. Thabtah. Detecting autistic traits using computational intelligence & machine learning techniques. Master of Research Thesis, School of Health, Department of Psychology, University of Huddersfield; 2019.
30. Thabtah F, Peebles D. A new machine learning model based on induction of rules for autism detection. Health Inform J. 2019. https://doi.org/10.1177/1460458218824711.
31. Thabtah F, Kamalov F, Rajab K. A new computational intelligence approach to detect autistic features for autism screening. Int J Med Inform. 2018;117:112–24.
32. Towle P, Patrick P. Autism spectrum disorder screening instruments for very young children: a systematic review. Autism Res Treat. 2016;2016:4624829.
33. Wall DP, Kosmiscki J, Deluca TF, Harstad L, Fusaro VA. Use of machine learning to shorten observation-based screening and diagnosis of autism. Transl Psychiatr. 2012;2(4):e100.
34. Wall DP, Dally R, Luyster R, Jung JY, Deluca TF. Use of artificial intelligence to shorten the behavioral diagnosis of autism. PLoS ONE. 2012;7(8):e43855.
35. Witten I, Frank E. *Data mining: practical machine learning tools and techniques*. Burlington: Morgan Kaufmann; 2005.